LNCS 8511

Masaaki Kurosu (Ed.)

Human-Computer Interaction

Advanced Interaction Modalities and Techniques

16th International Conference, HCI International 2014
Heraklion, Crete, Greece, June 22–27, 2014
Proceedings, Part II

2 Part II

HCI2014
INTERNATIONAL

Springer

Lecture Notes in Computer Science 8511

Masaaki Kurosu (Ed.)

Human-Computer Interaction

Advanced Interaction Modalities and Techniques

16th International Conference
HCI International 2014
Heraklion, Crete, Greece, June 22-27, 2014
Proceedings, Part II

Springer

Volume Editor

Masaaki Kurosu
The Open University of Japan
2-11 Wakaba, Mihama-ku, Chiba-shi
Chiba 261-8586, Japan
E-mail: masaakikurosu@spa.nifty.com

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-07229-6 e-ISBN 978-3-319-07230-2
DOI 10.1007/978-3-319-07230-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014938357

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

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 16th International Conference on Human–Computer Interaction, HCI International 2014, was held in Heraklion, Crete, Greece, during June 22–27, 2014, incorporating 14 conferences/thematic areas:

Thematic areas:

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

Affiliated conferences:

- 11th International Conference on Engineering Psychology and Cognitive Ergonomics
- 8th International Conference on Universal Access in Human–Computer Interaction
- 6th International Conference on Virtual, Augmented and Mixed Reality
- 6th International Conference on Cross-Cultural Design
- 6th International Conference on Social Computing and Social Media
- 8th International Conference on Augmented Cognition
- 5th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- Third International Conference on Design, User Experience and Usability
- Second International Conference on Distributed, Ambient and Pervasive Interactions
- Second International Conference on Human Aspects of Information Security, Privacy and Trust
- First International Conference on HCI in Business
- First International Conference on Learning and Collaboration Technologies

A total of 4,766 individuals from academia, research institutes, industry, and governmental agencies from 78 countries submitted contributions, and 1,476 papers and 225 posters were included in the proceedings. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers 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 (HCI), addressing the following major topics:

- Gesture-based interaction
- Gesture, gaze and activity recognition

- Speech, natural language and conversational interfaces
- Natural and Multimodal interfaces
- Human-robot interaction
- Emotions recognition

The remaining volumes of the HCI International 2014 proceedings are:

- Volume 1, LNCS 8510, Human–Computer Interaction: HCI Theories, Methods and Tools (Part I), edited by Masaaki Kurosu
- Volume 3, LNCS 8512, Human–Computer Interaction: Applications and Services (Part III), edited by Masaaki Kurosu
- Volume 4, LNCS 8513, Universal Access in Human–Computer Interaction: Design and Development Methods for Universal Access (Part I), edited by Constantine Stephanidis and Margherita Antona
- Volume 5, LNCS 8514, Universal Access in Human–Computer Interaction: Universal Access to Information and Knowledge (Part II), edited by Constantine Stephanidis and Margherita Antona
- Volume 6, LNCS 8515, Universal Access in Human–Computer Interaction: Aging and Assistive Environments (Part III), edited by Constantine Stephanidis and Margherita Antona
- Volume 7, LNCS 8516, Universal Access in Human–Computer Interaction: Design for All and Accessibility Practice (Part IV), edited by Constantine Stephanidis and Margherita Antona
- Volume 8, LNCS 8517, Design, User Experience, and Usability: Theories, Methods and Tools for Designing the User Experience (Part I), edited by Aaron Marcus
- Volume 9, LNCS 8518, Design, User Experience, and Usability: User Experience Design for Diverse Interaction Platforms and Environments (Part II), edited by Aaron Marcus
- Volume 10, LNCS 8519, Design, User Experience, and Usability: User Experience Design for Everyday Life Applications and Services (Part III), edited by Aaron Marcus
- Volume 11, LNCS 8520, Design, User Experience, and Usability: User Experience Design Practice (Part IV), edited by Aaron Marcus
- Volume 12, LNCS 8521, Human Interface and the Management of Information: Information and Knowledge Design and Evaluation (Part I), edited by Sakae Yamamoto
- Volume 13, LNCS 8522, Human Interface and the Management of Information: Information and Knowledge in Applications and Services (Part II), edited by Sakae Yamamoto
- Volume 14, LNCS 8523, Learning and Collaboration Technologies: Designing and Developing Novel Learning Experiences (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
- Volume 15, LNCS 8524, Learning and Collaboration Technologies: Technology-rich Environments for Learning and Collaboration (Part II), edited by Panayiotis Zaphiris and Andri Ioannou

- Volume 16, LNCS 8525, Virtual, Augmented and Mixed Reality: Designing and Developing Virtual and Augmented Environments (Part I), edited by Randall Shumaker and Stephanie Lackey
- Volume 17, LNCS 8526, Virtual, Augmented and Mixed Reality: Applications of Virtual and Augmented Reality (Part II), edited by Randall Shumaker and Stephanie Lackey
- Volume 18, LNCS 8527, HCI in Business, edited by Fiona Fui-Hoon Nah
- Volume 19, LNCS 8528, Cross-Cultural Design, edited by P.L. Patrick Rau
- Volume 20, LNCS 8529, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management, edited by Vincent G. Duffy
- Volume 21, LNCS 8530, Distributed, Ambient, and Pervasive Interactions, edited by Norbert Streitz and Panos Markopoulos
- Volume 22, LNCS 8531, Social Computing and Social Media, edited by Gabriele Meiselwitz
- Volume 23, LNAI 8532, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris
- Volume 24, LNCS 8533, Human Aspects of Information Security, Privacy and Trust, edited by Theo Tryfonas and Ioannis Askoxylakis
- Volume 25, LNAI 8534, Foundations of Augmented Cognition, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
- Volume 26, CCIS 434, HCI International 2014 Posters Proceedings (Part I), edited by Constantine Stephanidis
- Volume 27, CCIS 435, HCI International 2014 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 2014 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*, Dr. Abbas Moallem.

I would also like to thank for their contribution towards the smooth organization of the HCI International 2014 Conference the members of the Human–Computer Interaction Laboratory of ICS-FORTH, and in particular George Paparoulis, Maria Pitsoulaki, Maria Bouhli, and George Kapnas.

April 2014

Constantine Stephanidis
General Chair, HCI International 2014

Organization

Human–Computer Interaction

Program Chair: Masaaki Kurosu, Japan

Human Interface and the Management of Information

Program Chair: Sakae Yamamoto, Japan

Engineering Psychology and Cognitive Ergonomics

Program Chair: Don Harris, UK

Guy Andre Boy, USA
Shan Fu, P.R. China
Hung-Sying Jing, Taiwan
Wen-Chin Li, Taiwan
Mark Neerincx, The Netherlands
Jan Noyes, UK
Paul Salmon, Australia
Axel Schulte, Germany
Siraj Shaikh, UK
Sarah Sharples, UK
Anthony Smoker, UK
Neville Stanton, UK
Alex Stedmon, UK
Andrew Thatcher, South Africa

Universal Access in Human–Computer Interaction

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

Julio Abascal, Spain
Gisela Susanne Bahr, USA
João Barroso, Portugal
Margrit Betke, USA
Anthony Brooks, Denmark
Christian Bühler, Germany
Stefan Carmien, Spain
Hua Dong, P.R. China
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
Barbara Leporini, Italy
Eugene Loos, The Netherlands
Ana Isabel Paraguay, Brazil
Helen Petrie, UK
Michael Pieper, Germany
Enrico Pontelli, USA
Jaime Sanchez, Chile
Alberto Sanna, Italy
Anthony Savidis, Greece
Christian Stary, Austria
Hirotada Ueda, Japan
Gerhard Weber, Germany
Harald Weber, Germany

Virtual, Augmented and Mixed Reality

Program Chairs: Randall Shumaker, USA, and Stephanie Lackey, USA

Roland Blach, Germany
Sheryl Brahnam, USA
Juan Cendan, USA
Jessie Chen, USA
Panagiotis D. Kaklis, UK
Hirokazu Kato, Japan
Denis Laurendeau, Canada
Fotis Liarokapis, UK
Michael Macedonia, USA
Gordon Mair, UK

Jose San Martin, Spain
Tabitha Peck, USA
Christian Sandor, Australia
Christopher Stapleton, USA
Gregory Welch, USA

Cross-Cultural Design

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

Yee-Yin Choong, USA
Paul Fu, USA
Zhiyong Fu, P.R. China
Pin-Chao Liao, P.R. China
Dyi-Yih Michael Lin, Taiwan
Rungtai Lin, Taiwan
Ta-Ping (Robert) Lu, Taiwan
Liang Ma, P.R. China
Alexander Mädche, Germany
Sheau-Farn Max Liang, Taiwan
Katsuhiko Ogawa, Japan
Tom Plocher, USA
Huatong Sun, USA
Emil Tso, P.R. China
Hsiu-Ping Yueh, Taiwan
Liang (Leon) Zeng, USA
Jia Zhou, P.R. China

Online Communities and Social Media

Program Chair: Gabriele Meiselwitz, USA

Leonelo Almeida, Brazil
Chee Siang Ang, UK
Aneesha Bakharia, Australia
Ania Bobrowicz, UK
James Braman, USA
Farzin Deravi, UK
Carsten Kleiner, Germany
Niki Lambropoulos, Greece
Soo Ling Lim, UK
Anthony Norcio, USA
Portia Pusey, USA
Panote Siriaraya, UK
Stefan Stieglitz, Germany
Giovanni Vincenti, USA
Yuanqiong (Kathy) Wang, USA
June Wei, USA
Brian Wentz, USA

Augmented Cognition

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

Ahmed Abdelkhalek, USA
Robert Atkinson, USA
Monique Beaudoin, USA
John Blitch, USA
Alenka Brown, USA
Rosario Cannavò, Italy
Joseph Cohn, USA
Andrew J. Cowell, USA
Martha Crosby, USA
Wai-Tat Fu, USA

Rodolphe Gentili, USA
Frederick Gregory, USA
Michael W. Hail, USA
Monte Hancock, USA
Fei Hu, USA
Ion Juvina, USA
Joe Keebler, USA
Philip Mangos, USA
Rao Mannepalli, USA
David Martinez, USA
Yvonne R. Masakowski, USA
Santosh Mathan, USA
Ranjeev Mittu, USA
Keith Niall, USA
Tatana Olson, USA
Debra Patton, USA
June Pilcher, USA
Robinson Pino, USA
Tiffany Poeppelman, USA
Victoria Romero, USA
Amela Sadagic, USA
Anna Skinner, USA
Ann Speed, USA
Robert Sottilare, USA
Peter Walker, USA

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

Program Chair: Vincent G. Duffy, USA

Giuseppe Andreoni, Italy
Daniel Carruth, USA
Elsbeth De Korte, The Netherlands
Afzal A. Godil, USA
Ravindra Goonetilleke, Hong Kong
Noriaki Kuwahara, Japan
Kang Li, USA
Zhizhong Li, P.R. China
Tim Marler, USA
Jianwei Niu, P.R. China
Michelle Robertson, USA
Matthias Rötting, Germany
Mao-Jiun Wang, Taiwan
Xuguang Wang, France
James Yang, USA

Design, User Experience, and Usability

Program Chair: Aaron Marcus, USA

Sisira Adikari, Australia
Claire Ancient, USA
Arne Berger, Germany
Jamie Blustein, Canada
Ana Boa-Ventura, USA
Jan Brejcha, Czech Republic
Lorenzo Cantoni, Switzerland
Marc Fabri, UK
Luciane Maria Fadel, Brazil
Tricia Flanagan, Hong Kong
Jorge Frascara, Mexico
Federico Gobbo, Italy
Emilie Gould, USA
Rüdiger Heimgärtner, Germany
Brigitte Herrmann, Germany
Steffen Hess, Germany
Nouf Khashman, Canada
Fabiola Guillermina Noël, Mexico
Francisco Rebelo, Portugal
Kerem Rızvanoğlu, Turkey
Marcelo Soares, Brazil
Carla Spinillo, Brazil

Distributed, Ambient and Pervasive Interactions

Program Chairs: Norbert Streitz, Germany, and Panos Markopoulos, The Netherlands

Juan Carlos Augusto, UK
Jose Bravo, Spain
Adrian Cheok, UK
Boris de Ruyter, The Netherlands
Anind Dey, USA
Dimitris Grammenos, Greece
Nuno Guimaraes, Portugal
Achilles Kameas, Greece
Javed Vassilis Khan, The Netherlands
Shin'ichi Konomi, Japan
Carsten Magerkurth, Switzerland
Ingrid Mulder, The Netherlands
Anton Nijholt, The Netherlands
Fabio Paternó, Italy
Carsten Röcker, Germany
Teresa Romao, Portugal
Albert Ali Salah, Turkey
Manfred Tscheligi, Austria
Reiner Wichert, Germany
Woontack Woo, Korea
Xenophon Zabulis, Greece

Human Aspects of Information Security, Privacy and Trust

Program Chairs: Theo Tryfonas, UK, 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

HCI in Business

Program Chair: Fiona Fui-Hoon Nah, USA

Andreas Auinger, Austria
Michel Avital, Denmark
Traci Carte, USA
Hock Chuan Chan, Singapore
Constantinos Coursaris, USA
Soussan Djamasbi, USA
Brenda Eschenbrenner, USA
Nobuyuki Fukawa, USA
Khaled Hassanein, Canada
Milena Head, Canada
Susanna (Shuk Ying) Ho, Australia
Jack Zhenhui Jiang, Singapore
Jinwoo Kim, Korea
Zoonky Lee, Korea
Honglei Li, UK
Nicholas Lockwood, USA
Eleanor T. Loiacono, USA
Mei Lu, USA
Scott McCoy, USA
Brian Mennecke, USA
Robin Poston, USA
Lingyun Qiu, P.R. China
Rene Riedl, Austria
Matti Rossi, Finland
April Savoy, USA
Shu Schiller, USA
Hong Sheng, USA
Choon Ling Sia, Hong Kong
Chee-Wee Tan, Denmark
Chuan Hoo Tan, Hong Kong
Noam Tractinsky, Israel
Horst Treiblmaier, Austria
Virpi Tuunainen, Finland
Dezhi Wu, USA
I-Chin Wu, Taiwan

Learning and Collaboration Technologies

Program Chairs: Panayiotis Zaphiris, Cyprus, and Andri Ioannou, Cyprus

Ruthi Aladjem, Israel
Abdulaziz Aldaej, UK
John M. Carroll, USA
Maka Eradze, Estonia
Mikhail Fominykh, Norway
Denis Gillet, Switzerland
Mustafa Murat Inceoglu, Turkey
Pernilla Josefsson, Sweden
Marie Joubert, UK
Sauli Kiviranta, Finland
Tomaž Klobučar, Slovenia
Elena Kyza, Cyprus
Maarten de Laat, The Netherlands
David Lamas, Estonia
Edmund Laugasson, Estonia
Ana Loureiro, Portugal
Katherine Maillet, France
Nadia Pantidi, UK
Antigoni Parmaxi, Cyprus
Borzoo Pourabdollahian, Italy
Janet C. Read, UK
Christophe Reffay, France
Nicos Souleles, Cyprus
Ana Luísa Torres, Portugal
Stefan Trausan-Matu, Romania
Aimilia Tzanavari, Cyprus
Johnny Yuen, Hong Kong
Carmen Zahn, Switzerland

External Reviewers

Ilia Adami, Greece
Iosif Klironomos, Greece
Maria Korozi, Greece
Vassilis Kouroumalis, Greece
Asterios Leonidis, Greece
George Margetis, Greece
Stavroula Ntoa, Greece
Nikolaos Partarakis, Greece

HCI International 2015

The 15th International Conference on Human–Computer Interaction, HCI International 2015, will be held jointly with the affiliated conferences in Los Angeles, CA, USA, in the Westin Bonaventure Hotel, August 2–7, 2015. It will cover a broad spectrum of themes related to HCI, including theoretical issues, methods, tools, processes, and case studies in HCI design, as well as novel interaction techniques, interfaces, and applications. The proceedings will be published by Springer. More information will be available on the conference website: `http://www.hcii2015.org/`

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

Table of Contents – Part II

Gesture-Based Interaction

Gesture, Gaze and Activity Recognition

Speech, Natural Language and Conversational Interfaces

Natural and Multimodal Interfaces

Human-Robot Interaction

Emotions Recognition

Gesture-Based Interaction

RemoteHand: A Wireless Myoelectric Interface

Andreas Attenberger and Klaus Buchenrieder

Institut für Technische Informatik, Universität der Bundeswehr München,
Neubiberg, Germany
{andreas.attenberger,klaus.buchenrieder}@unibw.de

Abstract. While myoeletric signals (MES) have long been employed for actuating hand prostheses, their potential as novel input for the interaction with computer systems has received little attention up until now. In this contribution, we present *RemoteHand*, a system that fosters remote device control through the transmission of myoelectric data over WLAN. This allows to manipulate objects through the user's muscle activity regardless of their physical location. In our setup, a mechanical hand is controlled through electromyographic (EMG) sensors placed over the user's forearm muscles. This approach is compared to a conventional remote device control exercised by a tablet touchpad. The results of our user study show that wireless interaction through myoelectric signals is a valid approach. Study participants achieved interaction speeds equal to those of a standard input method. Users especially value myoelectric input with regard to novelty and stimulation.

Keywords: EMG, Myoelectric Signals, Prosthetic Hand, Remote Control, Wireless.

1 Introduction

Baseline work on employing myoelectric data for controlling upper limb prostheses dates back to 1948 [1]. Extensive research on MES processing has advanced since, mostly with a focus on prosthesis control [2]. In this contribution, we disclose a system for remote device interaction profiting from forearm muscle activity. While an actual prosthetic hand serves as the device to be actuated in immediate user-vicinity, our MES-based console opens new venues to interact with computer systems. This aspect of EMG signal acquisition has only received limited attention. Existing wireless EMG solutions are solely employed for gathering the signal data on a computer for further analysis or do not focus on MES exclusively. Our contribution presents a working, myoelectric control system with data transmitted through a WLAN connection, thus removing location constraints. With *RemoteHand* users are given the ability to remotely control a mechanical hand, making it possible to utilize their muscle activity for manipulating remote objects. Our user study shows, that wireless myoelectric sensing presents an invaluable amelioration in human computer interaction. On average, the participants achieved interaction speeds similar to or exceeding an established touch interface. *RemoteHand* also received high user-ratings regarding stimulating aspects and the novelty of the approach.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 3–11, 2014.

2 Related Work

In this contribution, we focus on surface EMG electrode systems rather than invasive subcutaneous, implanted sensors. Wireless EMG solutions applying surface electrodes are commercially available from companies like BTS[1] or DelSys[2]. However, such system solutions focus on the analysis of EMG data with respect to medical aspects. These products generally use a proprietary protocol for signal transmission, excluding the disclosure of the measured data for another purpose or subsequent processing with custom computer systems. Generally, such systems merely serve as preprocessing blocks and no control information is derived from the myoeletric signal.

While mainstream research in EMG control targets the advancement of prosthetic devices, we explicitly consider EMG sensor data as a novel means for human computer interaction. Augmenting interaction capabilites through myoelectric sensing was notably introduced by the artist Stelarc with an EMG-controlled third hand in 1980 [3]. Research on EMG input for human computer interaction has since only been deducted sparingly. Saponas et al. investigated the overall feasibility of myo-induced interaction solely focusing on gesture recognition [4]. They used eight sensors and only measured signals, not including an interaction component for the user. In a subsequent publication, the authors extend their approach to interactive systems and reveal a wireless EMG device prototype [5]. The proposition for a wearable EMG forearm band has recently resurfaced with the MYO band [3], which will additionally include accelerometers as integral components [6]. It was announced to be released to the market at the end of 2013, however shipping of final units is now planned for mid-2014[4]. Dubost and Tanaka employ EMG signals for interaction in musical performance [7]. Other systems include myoelectric sensing as an additional input method [8] or solely as a means to enhance interaction with an existing system [9] [10].

The *RemoteHand* prototype presented in this contribution enables wireless network transmission of EMG control information, so that the object or device to be interacted with can be physically distant from the user. Only standard hardware components are employed in our setup. Furthermore, in contrast to other approaches, our setup requires only two EMG sensors, reducing the amount of time spent on sensor positioning. Finally, as in traditional prosthesis control, we solely rely only on signal thresholding to derive control information without the need for a system training phase.

3 Prototype

A typical forearm EMG signal of a wrist flexion is shown in Figure 2. The average signal strength, as denoted by the RMS values, rises during muscle contraction

[1] `http://www.btsbioengineering.com/products/surface-emg/bts-freeemg/`

[2] `http://www.delsys.com/Products/Wireless.html`

[3] `https://www.thalmic.com/myo/`

[4] `https://www.thalmic.com/en/myo/faq/`

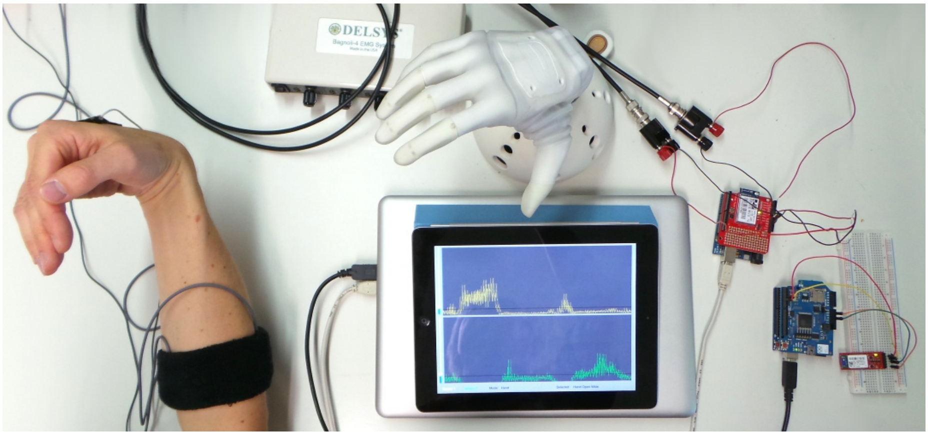


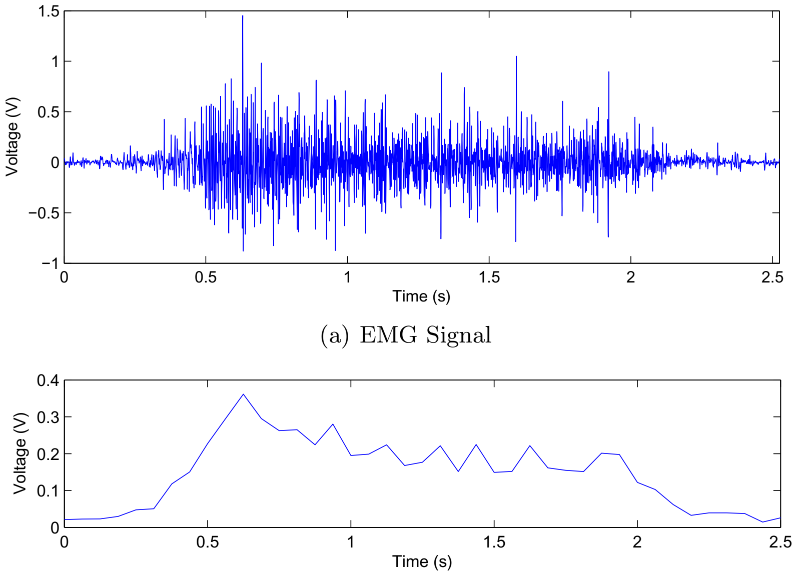
Fig. 1. The prototype setup with Arduino-Boards, EMG-Amplifier, i-Pad and the Michelangelo Hand$^{\text{TM}}$ by Otto Bock HealthCare

before returning to the background noise-level when the movement concludes. This sensor signal, picked up by EMG sensors, is then amplified, processed and sent over a wireless network connection for device control. The prototype is displayed in Figure 1, showing the Bagnoli EMG sensor and amplifier system manufactured by Delsys[5], an Arduino Uno board with a Sparkfun WiFly shield, an iPad with iOS 6, a second Arduino equipped with a Arduino WiFi and a Sparkfun Bluetooth Mate Gold shield and the Michelangelo Hand$^{\text{TM}}$ furnished by Otto Bock HealthCare. Not pictured is the WLAN access port that both the iPad and the WiFi-Shield connect to. The amplified myoelectric sensor data is connected to the analog inputs of the Arduino Uno with the mounted WiFly shield. The Arduino samples the analog inputs at about 10kHz and sends out RMS values to the iPad for each sensor with a window size of 32 values at a rate of 63Hz. This reduces the amount of data to be transferred as well as the processing power needed by the iOS app.

The RMS values are visualized in the app window shown in Figure 3. The lines shown in the realtime graph denote the adjustable thresholds, set to a level individual for each user, taking into account the background noise of the EMG signal [11]. As soon as the sensor signal exceeds the threshold, corresponding control commands are sent by the iPad app through the network to the second Arduino bearing the WiFi shield. The control command is in the format required by the Michelangelo Hand$^{\text{TM}}$ and transmitted through the Sparkfun Bluetooth shield.

Users can control *RemoteHand* with three hand gestures exhibiting different signal levels on the connected sensors: wrist flexion (high signal level on sensor 1 placed over the hand flexor muscles on the forearm), wrist extension (high activity on sensor 2 placed over the extensor muscles) and a fist or open palm

[5] `http://www.delsys.com/Products/Bagnoli_Desktop.html`



(a) EMG Signal

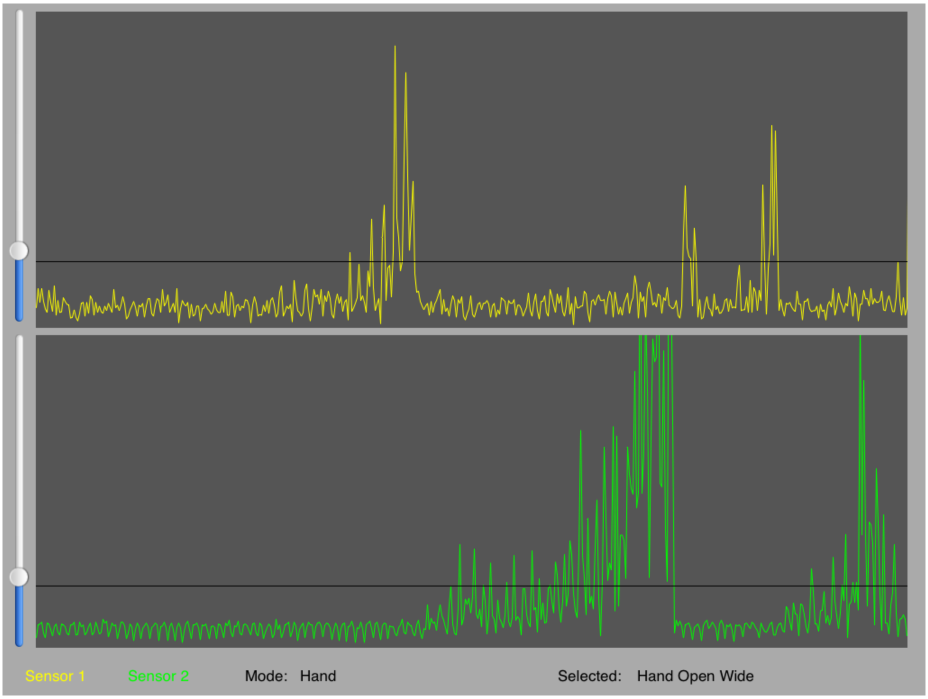
(b) RMS values calculated from the EMG signal

Fig. 2. The EMG signal and derived RMS values for the hand extensor muscles during wrist extension

gesture yielding a medium signal level on both sensors. Each of the first two gestures activates a different hand movement or grip, starting from the resting position. The hand remains in the selected grip or movement as long as the signal exceeds the threshold. The third gesture, with medium signal levels on both sensors, selects one of three modes, that can be accessed in the following order:

- Hand: In this mode, the hand is open as long as both sensor signals are below the threshold level. When exceeding the threshold, the hand closes either to a pinch (sensor 1) or a lateral grip (sensor 2).
- Pronation/Supination: When the threshold is exceeded for sensor 1, the hand is pronated, for sensor 2 the hand is supinated.
- Flexion/Extension: As soon as the threshold level for sensor 1 is reached, the hand is flexed at the wrist. In case of level saturation for sensor 2, the hand is extended at the wrist.

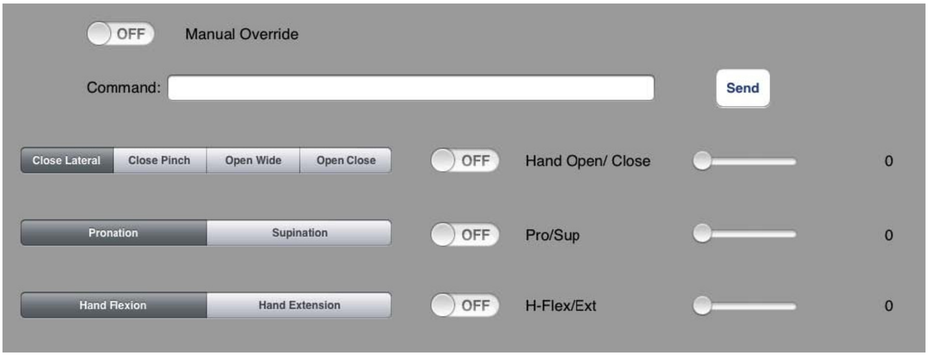
The iPad is also equipped with an app for touchpad interaction with the Michelangelo Hand$^{\text{TM}}$, serving as our baseline app for comparison with the myoelectric control. The corresponding control window is shown in Figure 4. The application allows the same movements as the myoelectric app. When a hand movement is activated and a movement speed is set with the slider, a corresponding control command for the hand is generated and can be transmitted



Mode: Hand

Selected: Hand Open Wide

Fig. 3. iOS app displaying the myoelectric sensor data sampled by the Arduino



OFF Manual Override

Command:

OFF Hand Open/ Close 0

OFF Pro/Sup 0

OFF H-Flex/Ext 0

Fig. 4. The touchpad interface for controlling the Michelangelo Hand

with the send button. The commands are again first received by the Arduino WiFi shield and then transmitted to the hand through the Bluetooth shield.

4 User Study

To determine how users perceive EMG data for remote device control, we asked a group of 10 able-bodied participants to perform a number of tasks with the following interaction options: a) using myoelectric data from the forearm, visualized on an iPad and b) solving the task by employing only the iPad's touch interface without myoelectric sensing. The participants aged 21-37 were recruited among students and research staff. Most of them studying or working in the field of computer science. None of the probands had previous experience with myoelectric device control. After a short introduction to the app and threshold adjustment, the following tasks were assigned amid the study:

- Task 1 (Hand Close & Open): Starting from the open position, the hand was to be closed and opened.
- Task 2 (Pronation & Supination): The hand was first to be rotated in one and then in the opposite direction. The order of the pronation and supination movements were not predetermined.

Each participant solved the tasks with both interaction options within a 30 minute time slot. The study was conducted with a 2x2 repeated-measures design. All sessions were recorded on video for further reference and future enhancement of the system. After completion of all tasks, the participants were asked to fill the abbreviated version of the standardized User Experience Questionnaire (UEQ)[6] for both interaction types (touchpad and myoelectric). One open ended item was added to the questionnaires prompting the participants to give further impressions, comments or suggestions with regard to the interaction method.

5 Discussion

The study revealed, that it was a major challenge to find an appropriate threshold when fitting the myoeletric control to the individual participant. Once the sensors were in place, participants were able to quickly solve both assigned tasks. As the second task required a mode change for the myoelectric control, it proved more challgenging and error-prone. Three participants needed a second attempt to carry out a positive change in mode. Two other participants experienced an unwanted mode change on the first task. Only one error occured during the use of the touchpad interface. The median task completion times with median absolute deviation (MAD) are displayed in Figure 5. One outlier was removed from task 1 with myoelectric control due to a value greater than three times the standard deviation and one user was not able to complete task 2 with the myoelectric control. Despite these errors, task completion times were similar for

[6] `http://www.ueq-online.org/`

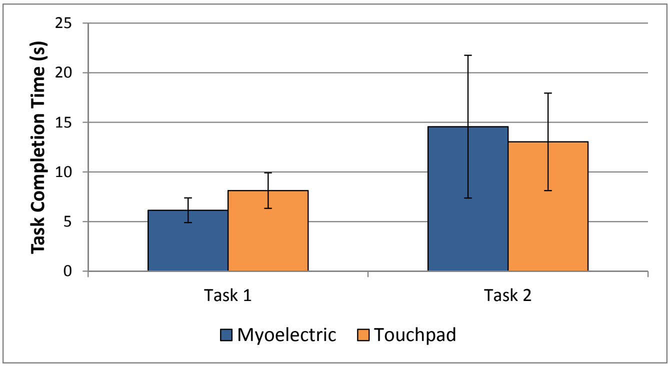


Fig. 5. Comparison of task completion times for both the myoelectric and the touchpad interface

both interaction types. It is to note that the differences are not statistically significant according to the applied paired samples *t*-test with $p > 0.48$ for each one of the tasks. However, this formalized test setup proved the observations we made during previous experiments with students and various probands.

The results of the UEQ are shown in Figure 6 with 95 percent confidence intervals. Both interaction types received positive feedback from the users with the myoelectric interation rating highly on stimulation and novelty. The lower score for perspicuity might stem from the current issue of having to adjust level-thresholds, which would probably be too difficult for a user to manage on their own. Novel users typically require help from more experienced end-users. This issue was also brought up by one participant in the corresponding questionnaire. One test-person found, that the system required a high amount of muscle activity for activation, which was likely caused by too strict threshold settings.

6 Conclusion and Future Work

In this paper, we presented *RemoteHand*, a wireless myoelectric interaction system for manipulating remote objects or as input device for assisted or guided steering applications. To validate the feasibility of the system, we conducted a user study comparing the wireless control of an artificial hand through a traditional touchpad and a myoelectric interface. While task completion times did not vary significantly between the two approaches, task completion for the myoelectric control included a period of error recovery for a number of users. By increasing the number of operation states, the mode change option could be

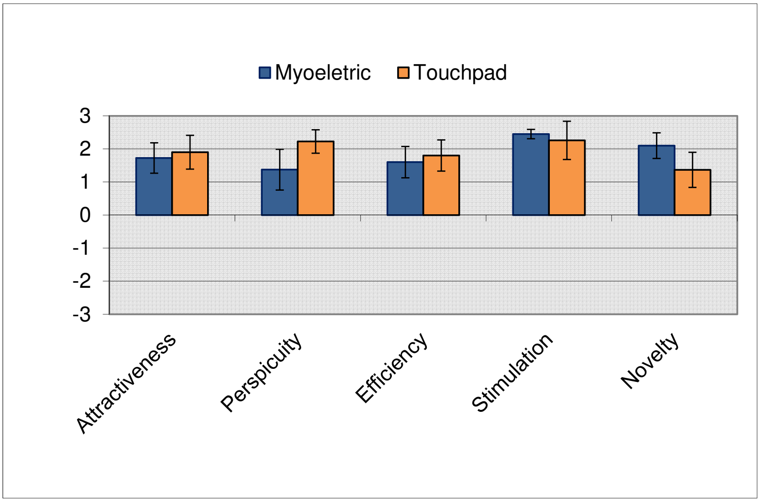


Fig. 6. Results of the UEQ for both interaction types

assigned to either the wrist flexion or the wrist extension gesture instead of gestures with a medium signal level on both sensors. This could increase the interaction speed as one of the participants in the study suggested. The results from the UEQ clearly show, that participants rate the system high with regard to the stimulating and novel aspects of this interaction type. These characteristics make myoelectric control also an interesting interface for electronic games. Due to the introduction of wireless transfer and on-par task completion times, other applications for controlling distant objects can now be investigated. By employing a low cost EMG system like the Olimex EKG/EMG shield for the Arduino[7], hardware cost can be further reduced, yielding a setup which can be applied to a wide range of scenarios. Implementing a method for automatic threshold setting might improve the perspicuity rating of the system. Furthermore, as the study has solely been conducted with able-bodied individuals, feasibility and potential applications for increasing the interaction possibilities exhibited by computer systems for actual prosthesis users cannot presently be estimated. As patients usually undergo training for MES control of their prostheses [2], they are however already familiar with myoelectric interaction and preliminary training of the interaction method is not necessary for this user group.

[7] `https://www.olimex.com/Products/Duino/Shields/SHIELD-EKG-EMG/`

Acknowledgements. We thank all participants for their participation in the study. Furthermore, we are grateful to Holger Fastenrath for his work toward implementing the Arduino and iPad software [12] and his help in conducting the study. We also thank Otto Bock HealthCare for supporting our research by supplying the Michelangelo Hand$^{\text{TM}}$.

References

1. Reiter, R.: Eine neue Elektrokunsthand. Grenzgebiete der Medizin 1(4), 133–135 (1948)
2. Muzumdar, A.: Powered Upper Limb Prostheses: Control, Implementation and Clinical Application. Springer (2004)
3. Stelarc: stelarc // Third Hand (2013), http://stelarc.org/?catID=20265
4. Saponas, T.S., Tan, D.S., Morris, D., Balakrishnan, R.: Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces. In: CHI 2008: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 515–524. ACM, New York (2008)
5. 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, pp. 167–176. ACM, New York (2009)
6. Techvibes, Y.: Combinator Backed Thalmic Labs Introduces MYO, the Motion Tracking Armband (2013), http://www.techvibes.com/blog/pre-order-thalmic-labs-myo-armband-2013-02-26
7. Dubost, G., Tanaka, A.: A wireless, network-based biosensor interface for music. In: Proceedings of International Computer Music Conference, ICMC (2002)
8. Zhang, X., Chen, X., Wang, W.H., Yang, J.H., Lantz, V., Wang, K.Q.: Hand gesture recognition and virtual game control based on 3D accelerometer and EMG sensors. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI 2009, pp. 401–406. ACM, New York (2009)
9. Benko, H., Saponas, T.S., Morris, D., Tan, D.: Enhancing input on and above the interactive surface with muscle sensing. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2009, pp. 93–100. ACM, New York (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, pp. 481–489. ACM, New York (2005)
11. Attenberger, A., Buchenrieder, K.: Wavelet-based detrending for emg noise removal. In: Rozenblit, J.W. (ed.) ECBS, pp. 196–202. IEEE (2013)
12. Fastenrath, H.: EMG Signalverarbeitung und Prothesensteuerung auf dem iPad. Master's thesis, Universität der Bundeswehr München (2013)

Early Prototyping of 3D-Gesture Interaction within the Presentation-Gesture-Dialog Design Space

Birgit Bomsdorf and Rainer Blum

Hochschule Fulda – University of Applied Sciences,
Marquardstr. 35, 36039 Fulda, Germany
{Birgit.Bomsdorf,Rainer.Blum}@informatik.hs-fulda.de

Abstract. Development of gesture interaction requires a combination of three design matters: presentation, gesture and dialog. In this contribution a first version of the tool ProGesture is introduced. The objective of its development is to cope with the resulting presentation-gesture-dialog design space in a flexible way. On the one hand, it aims at the early development phases, i.e. at rapid prototyping of 3D-gestures in combination with first UI sketches, such as mock-ups. On the other hand, it focuses on dialog and presentation modeling, and on testing based on executable models aiming at a smooth transition from informal UI sketches to formal models.

Keywords: 3D-Gesture Interaction, Early Prototyping, Model-Based Development.

1 Introduction

3D-gestures, such as touchless hand gestures and body movements are more and more used in human-computer interaction. Although gesture controlled user interfaces have been investigated for several years developing systematically intuitive and ergonomic 3D-gesture interactions is still challenging. Work in this field does not only aim at appropriate gestures taking into account the physiology of the human body and the users' goals, but also includes investigation of suitable UI widget types and presentation as a whole, as well as of the development process.

Nielsen et al. [1] contrast *technology-based* with *human-based approaches* for developing gestures. In the first one, gestures are implemented before being evaluated. Identified gestures are constrained by current technology resulting in solutions that may be undesired from the user perspective. Modifications, however, are costly if at all practicable. In human-based approaches users are asked to demonstrate gestures that should be implemented. This time, however, users may want gestures that are not realizable by up-to-date technology. *ProGesture,* the tool presented in this paper, supports rapid *Pro*totyping of 3D-*Gestur*e interactions allowing a combination of the two approaches to overcome the limitations. It is part of the authors' current work that addresses the issue of how to elicit and to evaluate gestures following a user-centered approach.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 12–23, 2014.

Model-based UI development involves a user-centered approach in engineering interactive systems. State-of-the-art works often favor the Cameleon framework[1]. It defines the modeling layers *Task & Concepts*, *Abstract User Interface (AUI)*, *Concrete UI (CUI)*, and *Final UI (FUI)*. The AUI is assumed to be independent of any modality of interaction while the CUI copes with modality. The idea is to systematically transform AUI into CUI. This is valuable once there is a common agreement on interaction objects and related interactions. In the area of 3D-interactions this commitment does not exist yet. Furthermore, finding gestures is mostly related to concrete presentations, e.g. in the form of first UI sketches, and not to abstract UI models. Nevertheless, tools reported in [2-4] have demonstrated the advantages of testing and prototyping, respectively, by means of executable models. ProGesture likewise enables to test elicited gestures based on dialog models, but moreover in the context of first UI sketches, as proposed in former work of the authors [5].

The structure of the paper is as follows: The next section takes a look at the design space spanned by the design dimensions presentation, dialog and gesture. Here, a coffee maker is taken as an example of developing gesture-based interactions according to a 1-, 2- or 3-axes design subspace. Then, related work is presented, that is followed by a short overview of the ProGesture tool. Next, usage scenarios are described revisiting the coffee maker examples introduced before. The paper concludes with a summarization of main results and an outlook.

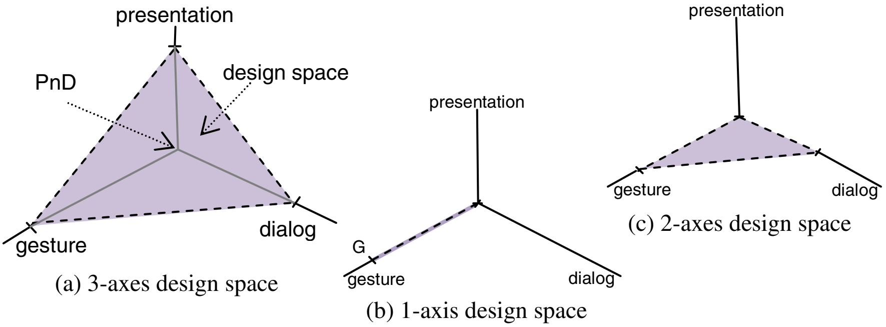
2 Gesture Interaction Design Space

The development space of gesture-based interactions is impacted by different aspects such as visualization techniques (e.g., 2D vs. 3D), sensor technique (e.g., Microsoft Kinect vs. LeapMotion) and gesture recognition algorithms (e.g., body movements vs. finger gestures). The sketches in Fig. 1 focus on presentation, gesture and dialog, whereby the axes represent the scope for development in each case. The more scope for design exists on each dimension the more space for finding a solution is available. If on the contrary the presentation, the gestures and the dialog are fixed the design space collapses to one point, to the "point of no design options" (PnD in Fig. 1a).

Example 1: Fig. 1b exemplifies a case in which the design space collapses to one dimension, here the gesture dimension. This is true if a gesture set is to be developed for an unchangeable application, e.g. for an existing coffee maker in a public area. The extent of the gesture design scope (distance of G and PnD in Fig.1b) is given by the applied sensor technology, the decision on gesture types such as body or only hand gestures, commitment to standard gestures[2] etc. Since presentation and dialog are fixed providing no design space they constrain identification of gestures and gesture set, respectively.

[1] `http://giove.isti.cnr.it/projects/cameleon/pdf/CAMELEON%20D1.RefFramework.pdf`

[2] Which does not exist at the moment of writing.



(a) 3-axes design space

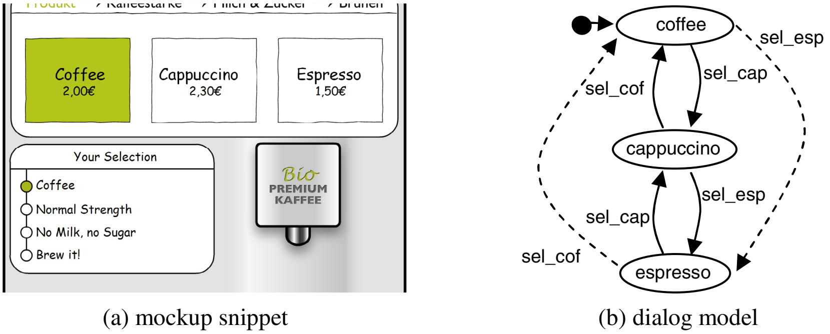
(b) 1-axis design space

(c) 2-axes design space

Fig. 1. 3D-gesture interaction design space

Example 2: If only the presentation is given developers deal with a 2-axes design space (Fig. 1c). The coffee maker mockup, for example, which is partially shown in Fig. 2, sets the presentation for selecting a coffee. Eliciting gestures for the task "select coffee type" may now result in different gestures affecting the dialog structure. The appearance and the alignment may suggest a slide gesture to put the virtual focus on the coffee of choice. The respective dialog is shown in Fig. 2b by means of a state transition diagram (please ignore the dotted arrows for this example).

The presentation, however, may also suggest a pointing gesture by which users can directly select a type of coffee. Deciding for this gesture results in a modified dialog structure (Fig. 2b, the dotted arrows now included). Users will expect no need to make a detour to cappuccino if the selection should be changed from coffee to espresso and vice versa.



(a) mockup snippet

(b) dialog model

Fig. 2. Given presentation mockup and related alternative dialog models

The presentation, more precisely the perceived affordance highly impacts the gestures users will use. The concept of affordances was introduced by Norman [6] and later on clarified as perceived affordances. It describes a desirable property of a UI or the objects of it which leads users to perform the correct actions to reach their goals.

If gestures, for example, are exhausting the user they have to be changed. Because of the demand for perceived affordances the presentation may have to be modified as well. All in all, designing gesture interaction requires a combination of three design matters: presentation, gesture and dialog.

Example 3: All of the three mentioned axes span the design space if a gesture controlled application is developed from scratch (Fig. 1a). Fig. 3 shows ideas of alternative designs for a coffee maker. In contrast to the example above the user is able to select coffee type, sugar and milk in a single dialog step. It is open which composition of the sugar and milk fields would be favored. Small changes to the layout could impact eliciting gestures from users, and possibly the dialog. The proposal on the left in Fig. 3 suggests to firstly select the coffee type, afterwards sugar, followed by milk and at last to confirm these choices. Even if no sugar is wanted the user has to pass over the sugar field (comparable to the dialog in Fig. 2b without the dotted arrows). This is not the case in the presentation on the right hand resulting in a modified dialog model.

In the case of the two presentations it is necessary to hold the gesture performing hand in front of the body while making all of the selections, from the first step up to the confirmation. Additional gestures may be allowed, e.g. a hold gesture to pause or a set gesture to fix a selection so that the user is able to relax the hand. Such a decision may also impact the presentation as well as the dialog.

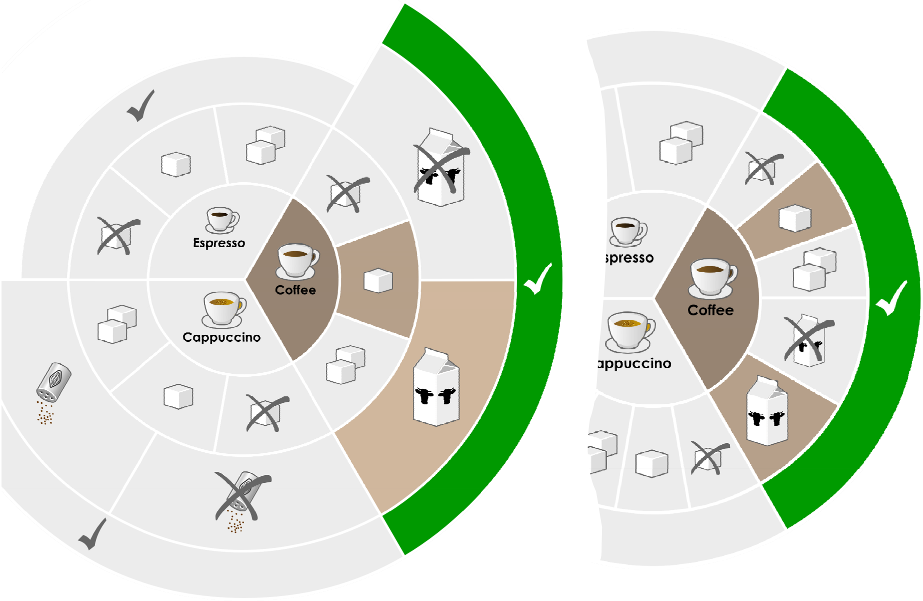


Fig. 3. Two versions of a design with only one dialog step

Developers, while prototyping 3D-gesture controlled UI, have to “move” within the presentation-gesture-dialog design space in a flexible way. They may, for example, fix the presentation as well as the dialog based on which they develop gestures in a subsequent design step. In the next moment, they may fix the gestures to redesign the presentation and dialog because of new insights. The development requires small iteration steps while following a user-centered design approach taking into account the mutual dependencies of presentation, gesture and dialog.

3 Related Work

Small iteration steps demand rapid prototyping starting as early as possible within the development process. One approach of early testing is based on executable models. In respective works [2], [3], [4], [5] user actions are simulated by activating buttons, while animation of the model diagrams visualizes system reactions in place of UI representations. It allows developers to evaluate a design while concentrating on the AUI independently of concrete gestures. Evaluating real gesture performance and sequencing is of vital importance once gestures are identified. ProGesture enables both, indicating gesture actions by means of buttons and, similarly to [7], by executing real gestures to control the executable model. In [7] gestures (in contrast to ProGesture only poses, no movements) and dialog models are specified separately. The gesture recognition is subsequently linked to the executable dialog model in an explicit modeling step. In ProGesture first steps towards an integrated tool are implemented aiming at specifying gestures during a test run and using it in its next step as user action. All in all, the tool presented in [7] and ProGesture support designers to cope with the gesture-dialog design space.

Another approach of early testing is revolutionary (also called throw-away) prototyping, e.g. mockups. Respective tools are more and more extended to include touch gestures, e.g. pidoco[3] and proto.io[4]. The focus is on presentation and interactions, i.e. on the presentation-gesture design space. The dialog is specified implicitly by means of connected, interactive areas by which the user can move within and between presentation units. ProGesture, in contrast, enables to test real-time gestures based on executable dialog models in the context of presentations to realize the presentation-gesture-dialog design space.

Explicit, formal dialog models are open to verification of specific properties. The same holds true for formal descriptions of 3D-gestures. Different gesture formalisms exist for the specification of poses and body movements, e.g. [8] and [9] that facilitate integration of gesture specification into the model-based approach. Furthermore, they enable animation of the gestures [10] supporting evaluation from the perspective of ergonomics. Current animation tools, however, do not incorporate the dialog or the presentation. Also the work presented in [7] necessitates video recording of users performing a gesture to analyze the movements afterwards. ProGesture implements the by-demonstration concept. Hereby, gestures are recorded and are immediately available for evaluation purposes – by real gestures in action or by animations.

Furthermore, in ProGesture recorded gestures can be assigned to a dialog and presentation, respectively. The tool presented in [11] follows a similar approach but considers touch gestures. There are, to the authors' knowledge, only two tools able to specify

[3] `http://www.pro-tact.de/`

[4] `http://proto.io/`

3D-gestures by demonstration. With the Kinetic Space Tool[5] a gesture is performed only one time and can be used afterwards by different persons. The tool can be linked to another application by a given communication protocol. With this mechanism, for example, it would possible to utilize the gestures in combination with executable models. An integration that enables to switch directly between, e.g. gesture specification and dialog modeling, is not possible. In addition, gesture recognition failed too often in our tests. Omek GAT[6] supports the by-demonstration concept as well. Specifying gestures demands several repetitions, around 30 performances are recommended. This is cumbersome particularly when it comes to rapid prototyping in small iterations.

4 Overview of the ProGesture Tool

The aim of the work presented here is to support developers of 3D-gesture-based UI by allowing them to "move" within the given design space in a most flexible way. This first version of ProGesture is intended as a proof-of-concept tool. In order to identify relevant requirements, practitioners interested in gesture interaction were involved from the beginning. It became apparent that it is not possible to raise a complete and thorough set of requirements at the moment due to the novelty of the topic and related uncertainties concerning good practice and design methods. Therefore, the tool's current version serves also as an experimental demonstrator to collect more precise and detailed requirements on how to develop ProGesture further as well as to investigate methodic aspects of the user-centered design process.

The tool is basically structured into three modules that are presented below: (1) gesture editor, (2) dialog editor and (3) model simulator. The gesture editor and the dialog editor are related to the gesture and to the dialog dimension, respectively. The presentation dimension, by contrast, is covered by creating UI sketches or elaborated UI outside of ProGesture and linking them within it to dialog models and thus to gesture specifications.

The model simulator allows testing and analyzing the dialog model at its own, in conjunction with gestures as well as in conjunction with gestures and presentation (involving all three axes then). A further module (4), which is currently under development, enables to link gestures directly to presentations and to perform evaluations, supporting iterative development within the presentation-gesture design space, i.e. without taking into account a dialog model.

From the functional and software architectural points of view the four modules are loosely coupled but work together closely. Replacement of each of the modules by a similar software package is possible, guaranteeing to be able to integrate future developments, also from third parties.

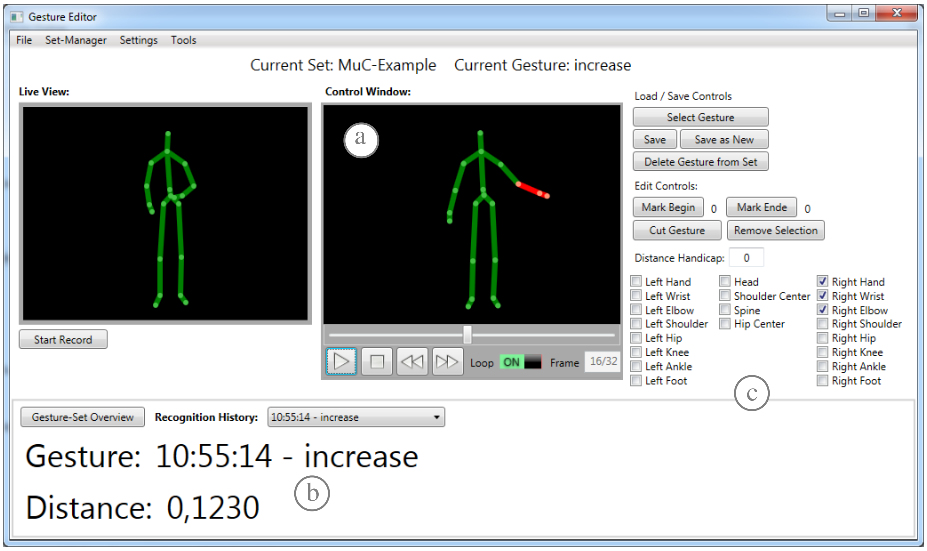
[5] Kinetic Space, Training and Recognizing 3D Gestures, `https://code.google.com/p/kineticspace`

[6] Gesture Authoring Tool, `http://www.omekinteractive.com/products/beckon-usability-framework`

4.1 Gesture Editor

The gesture editor provides functionalities for recording and editing gestures and organizing them in gesture sets. A gesture is specified “by-demonstration”, i.e. a user demonstrates the body movement in front of a sensor. Hence, gestures are not described explicitly using a specific human-readable notation, but as frame sequences comparable to video clips. The system currently employs the Microsoft Kinect sensor and works with the skeleton data delivered by the Kinect SDK. Consequently, each frame contains the separate positions of all the skeleton joints that are of relevance for the respective gesture. This frame sequence is accompanied by specific parameters for its interpretation, e.g. the tolerable deviation from the “ideal” movement.

Integral part of the gesture editor is a gesture recognizer implemented by our team. It is based upon the Dynamic Time Warping (DTW) algorithm [12] that is able to eliminate temporal variations when comparing two gesture sequences. The gesture recognizer is realized as an independent library and can therefore be employed to use recorded gestures in other applications. Actually, it is used by all ProGesture modules.



Current Set: MuC-Example Current Gesture: increase

Loop ON Frame 16/32

Mark Begin 0 Mark Ende 0

Distance Handicap: 0

☐ Left Hand
☐ Left Wrist
☐ Left Elbow
☐ Left Shoulder
☐ Left Hip
☐ Left Knee
☐ Left Ankle
☐ Left Foot

☐ Head
☐ Shoulder Center
☐ Spine
☐ Hip Center

☑ Right Hand
☑ Right Wrist
☑ Right Elbow
☐ Right Shoulder
☐ Right Hip
☐ Right Knee
☐ Right Ankle
☐ Right Foot

Recognition History: 10:55:14 - increase

Gesture: 10:55:14 - increase

Distance: 0,1230

Fig. 4. Screenshot of the gesture editor’s main window

The gestures editor provides first features for editing and testing of recorded gestures supporting iterative development within the gesture design space. When the user repeats a movement the editor reports the recognized gestures (Fig. 4b). Furthermore, recorded gestures can be played in a so-called gesture player (Fig. 4a) in order to analyze the skeleton movements and cut the frame sequence as needed. Additionally, the skeleton joints that are of relevance for the respective gesture can be selected here, at that time or earlier before starting the recording (Fig. 4c). They are highlighted in the gesture player. For example, for a typical wipe gesture of one hand the positions and movements of head, legs and feet and even of the other arm are not of relevance and should be excluded from the gesture recognition process.

4.2 Dialog Editor

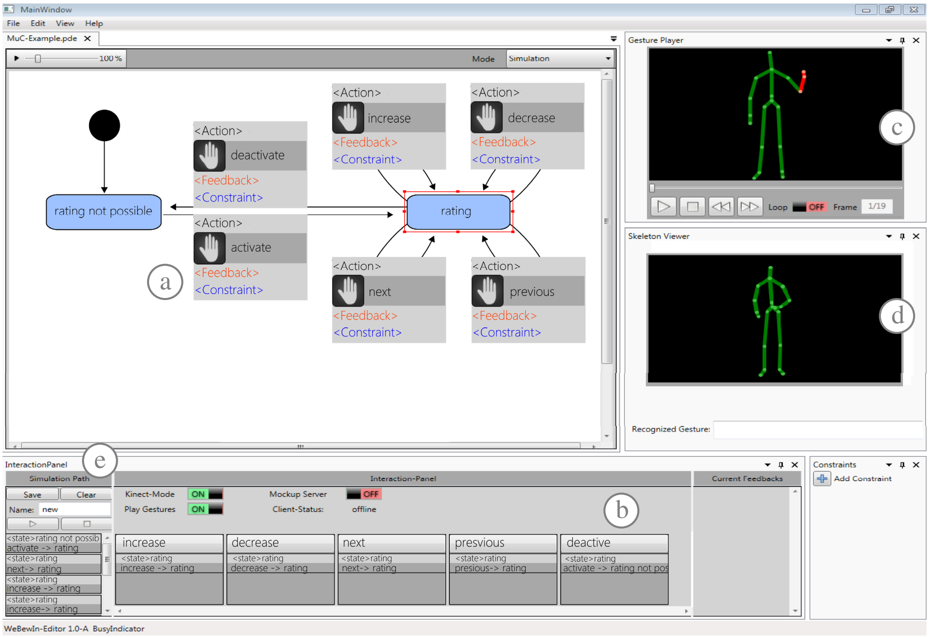
The dialog editor supports the specification of dialog models by means of state charts. Recorded gestures are assigned to state transitions, together with additional information relevant for the dialog sequence, i.e. constraints and feedback such as highlighting of a chosen option. The gestures to be assigned to transitions can be selected from a gesture set. In the case a gesture is not recorded yet and to be specified later on it can be added to the dialog model by a representative name. In doing so, the dialogue model is already executable. Therefore, gestures may be specified previous to, during or after the dialog modeling allowing the developer to arbitrarily shift the design focus within the gesture-dialog space.

4.3 Model Simulator

The model simulator is based on an executable model that is derived from the dialog model. It allows the simulation of the dialog to analyze dialog paths, gesture sequences and system reactions in different scenarios and in different situations. Dialogs can be tested in conjunction with or without the assigned presentation. Additionally, model execution can be triggered either by mouse clicks or by real gestures.

Thus, the simulator provides various options for the test and evaluation of 3D-gesture applications, based on combinations of its different features, such as:

- Gesture actions can be simulated with mouse clicks on the graphical representation of the gesture within the diagram (Fig. 5a). All outgoing transitions of an active state are listed in a separate window panel (Fig. 5b). Within this panel gesture actions can be triggered by mouse clicks on buttons labeled with gestures names (e.g. “increase”, “next” etc. in Fig. 5).
- The gesture player introduced above is also integrated into the simulator (Fig. 5c). Developers can use it to recall gestures as well as to analyze and discuss the defined body movements in the context of complete interaction respectively gesture sequences.
- The dialog model can alternatively be traversed by executing the gestures in reality, e.g. for evaluating physical effort. Feedback is given with a live skeleton view (Fig. 5d) while the gesture recognizer works in the background.
- The presentation can be connected, currently via simple network communication. This enables interaction with the presentation, e.g. based on real gestures, to evaluate the whole user experience, also taking into account the system feedback.
- Switching to the gesture editor, existing gestures may be modified or replaced by a new version and used instantly inside the model simulation.
- A history function stores model simulation sequences, i.e. states traversed and constraints changed (Fig. 5e). For subsequent analysis they can be played back together with the involved gestures.



Mode Simulation

Kinect-Mode ON

Play Gestures ON

Mockup Server OFF

Client-Status: offline

Name: new

Fig. 5. Dialog editor in model simulation mode

5 Example Scenarios of Using the ProGesture Tool

The previous section outlines prototyping features of the simulator, and also of the editor and the fourth module that was shortly mentioned. This section provides insights into the usage of ProGesture by means of three scenarios revisiting the coffee maker examples (cf. Sect. 2).

5.1 Scenario 1: Developing Gesture Interaction for an Existing Application

In scenario 1 a gesture set is to be developed for an existing application, e.g. for the coffee maker of example 1 where we want to augment an existing maker in a public area with a gesture recognition module – causing low technical effort and modifications of the UI are not possible for some reasons. By consequence, the presentation and dialog are fixed resulting in a design space with only one dimension, which is here the gesture axis.

Let us assume that in this scenario a developer's first step is to analyze the given UI with respect to the presentation's units and transitions between them as well as to the system feedback. As a result the corresponding dialog is documented with the ProGesture dialog editor, as far as required for gesture test purposes. In addition, the model is linked with the presentation units, the latter in the form of a simple mockup including photos of the real coffee maker's presentation. In this scenario it is assumed that the existing UI is composed of a text display listing the available options with hardware buttons close-by to make the selection.

Now, the developer specifies gestures with the ProGesture gesture editor, taking into account the existing coffee machine UI. As in this scenario, he has to replace hardware buttons by gestures, he decides for semaphoric gestures. A semaphoric gesture is precisely designed to designate one specific symbol within a given alphabet, e.g. like in this scenario, “chose option 1”, “… option 2” and “… option 3”. Here, this gesture type is used to express and confirm the coffee type selection in one step.

After assigning these gestures to the dialog model, the setting can be tested interactively based on the ProGesture executable model – using the real gestures. Problems may be identified, e.g. in the sequencing of the gestures and prospective solutions can be devised. Spontaneously, gestures can be modified or replaced by newly recorded ones using the gesture editor, and then used instantly within a follow-up test.

5.2 Scenario 2: Developing Gesture Interaction for a Given Presentation

In scenario 2, which refers to example 2 (cf. Sec. 2), only the presentation is given: The coffee maker mockup, of which a selected part is shown in Fig. 2, constraints the presentation for choosing a coffee type. Here, developers deal with a 2-axes design space, requiring a design solution concerning gestures and the corresponding dialog.

Let us assume, the developer again starts with analyzing the given UI, though this time represented by sketches without included concrete dialog behavior. Then, with the ProGesture gesture editor a set of gestures is compiled, whereby the developer can choose gestures from existing sets or record new ones. The mockup, as pointed out for example 2, may suggest a wipe gesture to make a selection. In this scenario, as a start, the designer decides for this option. The corresponding dialog model (cf. Fig. 2b) can be specified with the dialog editor either previous to the gesture specification, in parallel or afterwards. Then, the presentation is connected to the dialog model and an interactive test of the “look and feel” of the chosen 3D-gesture interaction is conducted instantly.

Possibly, as already mentioned in Sec. 2, it would be worth comparing an alternative: The mockup may also suggest a pointing gesture by which users select a type of coffee directly. This implies other gestures and a modification of the dialog structure – both of which is effectively done with the gesture and dialog editor. Switching between both design versions to compare them is easy and fast.

5.3 Scenario 3: Developing from Scratch

The last scenario exemplifies the development from scratch of a 3D-gesture controlled application (cf. example 3 in Sec. 2). As no design presets or limitations are given, the design space spans all of the three axes. This time our developer starts to investigate gestures and corresponding presentation clues, e.g. widgets, based on the user tasks involved when using the respective coffee maker.

The developer, with a specific gesture concept in mind, generates one or several presentation mockups using an arbitrary tool. Fig. 3 shows two versions of possible designs. Here, the central idea is, in contrast to the example in scenario 2, to let the user select coffee type, sugar and milk in a single dialog step. Since the two drafts differ in the sequence of selecting of sugar and milk, two slightly different dialog models have to be specified. This scenario does not imply a specific order of dealing with the three design dimensions – on the contrary, the designer can switch his focus arbitrarily.

At last, having connected the dialog model to the mockups, both versions are ready to be tested. As argued for example 3, the tests may unfold, that additional gestures may be needed to improve the usability. The resulting modifications for the dialog model and the gesture set as well as the subsequent evaluation of the new design can be realized quickly within ProGesture.

6 Summary and Outlook

The gesture recognizer used in ProGesture can also be utilized by a target application. In such a case only those user-elicited gestures are added to a gesture set that are at the same time technically realizable – resulting in a combined *technology-human-based approach* (cf. [1]). ProGesture additionally supports evaluation of gestures in the contexts of first UI drafts up to a final UI. Wizard of Oz experiments are often applied for rapid prototyping of gesture-based interactions (e.g. in [13], [14]). ProGesture can be used within such experiments but additionally supports a model-based development approach. It enables to test real as well as simulated 3D-gesture interactions based on executable dialog models.

ProGesture exists as a proof-of-concept tool that allows to gain first experiences and to get first feedback from industrial partners of the project (within which the tool was developed), as well as from an involved ergonomics expert. One of the main results so far is the importance of including mockups in eliciting and testing gestures, particularly if the development should result in an innovative design solution. All in all, it should be possible to shift the focus of development in a flexible way within the presentation-gesture-dialog design space. Developers may start with a prototype of a UI and then develop gestures (design space is given by the gesture dimension). While testing they may encounter problems with single gestures. In ProGesture a person just has to demonstrate alternative gestures and shortly afterwards use them to interact with the UI draft. In the case these gestures do not match the presentation (perceived affordances) and/or dialog behavior, the developer can alter the UI design and assign the gestures to it (acting in the presentation-dialog design space).

The concept of executable models was implemented to support the transformation from early models and prototypes to final code of target applications. Currently, the dialog editor and the model simulator are fully integrated into a coherent tool. This tool, the gesture editor and the additional, fourth module, that covers the presentation-gesture design space, are coupled by shared components. Subsequent work will aim at the integration of all of them. In ongoing work we investigate the development process and further applications of ProGesture. One objective is to refine the requirement specification of tool support for early prototyping of 3D-gesture interaction and to realize additional requirements in a follow-up version of ProGesture.

References

1. Nielsen, M., Störring, M., Moeslund, T.B., Granum, E.: A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. In: Gesture-Based Communication in Human-Computer Interaction - 5th International Gesture Workshop, pp. 409–420 (2004)
2. Biere, M., Bomsdorf, B., Szwillus, G.: The Visual Task Model Builder. In: Vanderdonckt, J., Puerta, A.R. (eds.) Computer-Aided Design of User Interfaces II, Proc. of the 3rd International Conference of CADUI, pp. 245–256. Kluwer (1999)
3. Mori, G., Paternò, F., Santoro, C.: CTTE: support for developing and analyzing task models for interactive system design. IEEE Trans. Softw. Eng. 28(8), 797–813 (2002)
4. Reichart, D., Forbrig, P., Dittmar, A.: Task Models as Basis for Requirements Engineering and Software Execution. In: Proceedings of the 3rd Annual Conference on Task Models and Diagrams TAMODIA 2004, pp. 33–42. ACM Press, New York (2004)
5. Bomsdorf, B., Szwillus, G.: Early prototyping based on executable task models. In: Conference Companion on Human Factors in Computing Systems, pp. 254–255 (1996)
6. Norman, D.A.: The Psychology of Everyday Things. Basic Books, New York (1998)
7. Feuerstack, S., Anjo, M.D.S., Pizzolato, E.B.: Model-based design and generation of a gesture-based user interface navigation control. In: Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems & the 5th Latin American Conference on Human-Computer Interaction (IHC+CLIHC 2011), pp. 227–231 (2011)
8. Loke, L., Larssen, A.T., Robertson, T.: Labanotation for design of movement-based interaction. In: Pisan, Y. (ed.) Proceedings of the Second Australasian Conference on Interactive Entertainment, pp. 113–120. Creativity & Cognition Studios Press, Sydney (2005)
9. Vilhjálmsson, H.H., et al.: The Behavior Markup Language: Recent Developments and Challenges. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 99–111. Springer, Heidelberg (2007)
10. Wilke, L., Calvert, T., Ryman, R., Fox, I.: From dance notation to human animation, The LabanDancer Project, Motion Capture and Retrieval. Computer Animation and Virtual Worlds 16(3-4), 201–211 (2005)
11. Lu, H.: Gesture Coder: A Tool for Programming Multi-Touch Gestures by Demonstration. In: CHI 2012: ACM Conference on Human Factors in Computing Systems, pp. 2875–2884 (2012)
12. Reyes, M., Dominguez, G., Escalera, S.: Feature weighting in dynamic time-warping for gesture recognition in depth data. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1182–1188 (2011)
13. Höysniemi, J., Hämäläinen, P., Turkk, L.: Wizard of Oz prototyping of computer vision based action games for children. In: Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community, New York, USA, pp. 27–34 (2004)
14. Rupprecht, D., Blum, R., Bomsdorf, B.: Towards a Gesture Set for a Virtual Try-On. In: Proceedings of IADIS International Conference Interfaces and Human Computer Interaction. IADIS Press, IADIS MCCSIS (2013)

The Study of the Full Cycle of Gesture Interaction, The Continuum between 2D and 3D

Mohamed-Ikbel Boulabiar[1], Gilles Coppin[1], and Franck Poirier[2]

[1] Lab-STICC, Telecom Bretagne, France
{mohamed.boulabiar, gilles.coppin}@telecom-bretagne.eu
[2] Lab-STICC, University of Bretagne-Sud, France
franck.poirier@univ-ubs.fr

Abstract. The goal of HCI researchers is to make interaction with computer interfaces simpler, efficient and more natural. In a context of object manipulation, we think that reaching this goal requires the ability to predict and recognize how humans grasp then manipulate objects. This is based on studies explaining human vision, reach, grasp taxonomies and manipulations. In this paper, we study the full cycle of gesture interaction using different points of view, then attempt to organize them using Norman's theory of Human Action, we link the psychology of object sensing to HCI goals and propose a simplification of gestures classes into four principal families. Our simplification of gestures classes still allow the expression of more detailed subclasses differentiated by the gesture properties.

Keywords: Gesture, 3D, Interaction, Hand, Grasping.

1 Introduction

From the birth, human beings tend to discover their environment using all their senses. They discover the existence of things by sight, then tend to touch, grasp and manipulate objects before using these recognized items or tools to accomplish other tasks [35]. Our study focuses on the whole cycle of an interaction [27,26] starting from observation, through grasping and focusing in manipulation and specifically direct maniplation of objects. We attempt to link many research areas for the quest of natural and direct interaction as expressed by Beaudouin-Lafon [2].

This research around natural interaction in 2D and 3D space is a high priority because interfaces and visualization techniques have evolved from command line to graphical user interface (GUI) with multidimensional graphical elements, but the interaction are still almost the same. The presence of mid-cost 3D stereoscopic displays, using special glasses, provides an immersion of 3D objects below and above the rendering surface. Some researchers worked on controlling such rendering with 2D multi-touch input [34] but this is still very limited and a real 3D spatial manipulation is more appropriate for direct interaction. In order to reflect real world manipulations with objects, many frameworks like [11]

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 24–35, 2014.

or [15] have abstracted the interaction between the hand and objects to those emulated by physics simulation engines. Even if these ways of interaction with objects provide a sense of reality, they still provide neither information about the psychology of grasping and manipulation, nor a prediction of the upcoming position of the hand and fingers. Hence comes the importance of psychological research to fill the gap between the goals of HCI among naturalness, efficiency, discoverability and the current state of studies.

One of the main foci of Human-Computer Interaction researchers is to make interaction easier and more intuitive allowing a larger spectrum of people to better use systems. This is called making interaction natural in the meaning of reducing the need to remember complex operations. Actions should be easily discoverable and the system may be learned through exploration. From the other side, there have been many studies on human dexterity and hand grasping taxonomies. Napier in [24] has proposed the classification of human grasp into two main categories: power grasp and precise grasp. According to this simple classification, we think that we can link studies on grasp and the gesture generation cycle with the objective of HCI research and start a new conquest for a more natural interaction.

Our contributions are : 1. Attempt to link between multi-fields of research around gesture, 2. Proposition of a new simplified taxonomy of gestures, 3. Rising the problematic of gestures prediction with virtual objects.

This paper proposes a representation of previous work on neuropsychology, grasp and gesture from different points of view and by using Norman model of human interaction to organize the separate areas of research, then explaining the limits of stopping the studies at grasping and not continuing to manipulation. In the next chapter we discuss manipulation of objects and we propose a simplified taxonomy.

2 Overview of the Related Work Around the Hand and Gesture

2.1 The Evolution of the Hand with Tool Use

When the first primates have left the trees, they started freeing their hands for new uses. As Napier [23] refers to, the use of hands in hominids evolved from self feeding. In contrast to most animals which use their both hands, hominids and some great apes are able to use one hand to grasp objects thanks to the thumb. It was thought that humans are the only creatures with the ability to create tools instead of just using them, they were even called “Homo Faber”. Even if it cited in Napier’s book, it was discovered recently that apes can use and even create tools [32]. This evolution of the ability to grasp, then the use and creation of tools in primates, diverged just in a small fraction that made the anatomy of human hand with more dexterity only for precise gestures manipulations like playing musical instruments [19].

2.2 Gestures and the Speech in the Brain

The recent man and the great apes like the chimpanzees can use hand gestures for communication [31]. David McNeill has emphasized on the relation between the gesture and the speech. In his book [20] he made an analysis of gestures relatively to speech. The taxonomy he proposed categorize gesticulation, speech-linked gestures, emblems and pantomime. The taxonomy of gestures vis-a-vis speech has its background as their are both proceeded by the same neural system in the brain [40].

2.3 Motor and Vision Brain Pathways

Manipulating an object requires detecting its presence mainly using the eyes, then planning an action. An experiment run by Aglioti et al. [1] demonstrated that the visual illusions have not impacted the motor action. This experiment has proven that the perception and the action are two separate paths that do not interfer with each others [5]. Another work has suggested that even for a hand action to occur, there are two pathways, one for moving the fingers and another for transporting the hand itself [29]. More detailed studies about hand movement and the physiology of grasping are made by Nowak et al. [28] studying subjects with disorders in the somatosensory system, or with the Parkinson's diseases among many others.

2.4 Proprioception, Exteroception and Manipulation Area

The proprioception is the individual perception of himself, and the self sense of the parts of the body. [38]. This knowledge is acquired during the first years of the individual existence and decline by aging. The proprioceptive kinesthetic sensory system intervene in controlling and correcting limb movements during a reach movement.

The importance of proprioception in a 3D gesture maniplation has been studied by Mine et al. [21]. The lack of feedback makes an interaction very difficult and should be compensated by the proprioceptive capacity. Memorizing real world positions by the exteroception, allow an efficient acces and reach to the objects. In reaching study, it is easier to remember a position relative to ones own hand more accurately than a position fixed in a virtual space.

2.5 Troubles in gestures Choices and Manipulation: Gestemes and Kinemes

The planning of manipulation starts with knowledge acquired in proprioception, which is the self sense of the parts of the body. It's a learning process which starts from the childhood and keeps enhancing through the age. In the grasping process, proprioception is primordial to orient the palm and reshape the hand [17].

The classification of Signoret and North cited in [33] considers the two distinct functional levels in gestures after a study on people having apraxia problems, and who fail in chosing either the right gesture for a specified object, or in performing the gesture through the right list of kinemes. These levels explained more in Bertranne thesis [4], are the goal of the gesture and the motor realization. Gesture production needs in the same time the choice of the right gesteme, then the sequence of kinemes performed with the hand. During the experiments, patients were asked either to just mimic a just performed gesture to evaluate the ideomotor apraxia, or to perform a named gesture to evaluate the ideational apraxia.

2.6 Grasp, Hand Shape and Fingers Configuration

The human dexterousity in grasping is efficient but very difficult to mimic. Researchers who want to build robots capable of grasping and moving objects focus on human grasping creating taxonomies which would be possible to mimic on robots [6]. The need for a robot capable of performing a multitude of tasks, Cutkosky [8] and Feix et al. [10] started by studying human grasp selection. These robots needs to grasp objects in order to manipulate them which include moving or rotating. Even if the goal is for robots, the main element in the process is the human, his/her hand and its dexterous manipulation [36] .

2.7 The Search for the Best Gesture Interaction in HCI

Many Human-Computer Interaction researchers have been working on tracking the best gestures to be used for a system. In their research, like the one conducted Wobbrock et al. and Gustafson et al. [39,13], they asked subjects during an experiment to perform what they thought would be the best gesture to achieve a specific task. During so, we may argue that the methods used in these studies are more statistical analyses than a proper reverse engineering of the human behaviors. Other studies focuses on finding newer methods of interaction or new gestures like Nacenta et al. work [22], or by the introduction of different sensors for the human manipulation [14].

2.8 Simulating a Rich Interaction

Physics simulation libraries are used to add a feeling of real interaction with objects like in the work by Frohlich et al. [11]. Given the properties of objects, Newtonian forces are computed to move, distort or bounce an object according on its type and how a user is touching it. Systems using these libraries are not able to gather semantics of gestures. They also can not predict user intention before the user hand reaches the object.

3 Organization of the Related Work

3.1 Organizing through a Real Scenario

The multidisciplinary nature of the related work makes studying the subject of gesture interaction difficult. Many researchers tried to take the problem from their point of view. Even with such diverse work, we can start to find what are the common things between them all and then glue parts together. All the research previously referred is around the human and objects. We think that in a real scenario, in which the human is manipulating objects, would put all the reseach pieces in place. Our practical case of study out of this work is with a human manipulating 2D and/or 3D stereoscopic objects which are on or above a table space as in figure 1 and the detailed interaction figure 2.

In this context, the user detects objects through vision with a possible 3D optical illusion (stereoscopy), reaches the object inside his interaction space, grasps the object depending on its form and the intended manipulation, then manipulates it through gestures. In the same time the operator uses 2D and 3D gestures, and this is why a single recognizer for the two cases is needed.

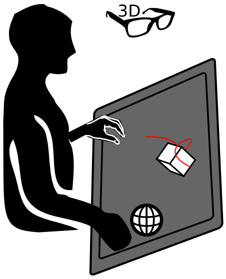


Fig. 1. The manipulation of the gestures on and above the table, using 2D and 3D virtual objects and captured via 3D sensors

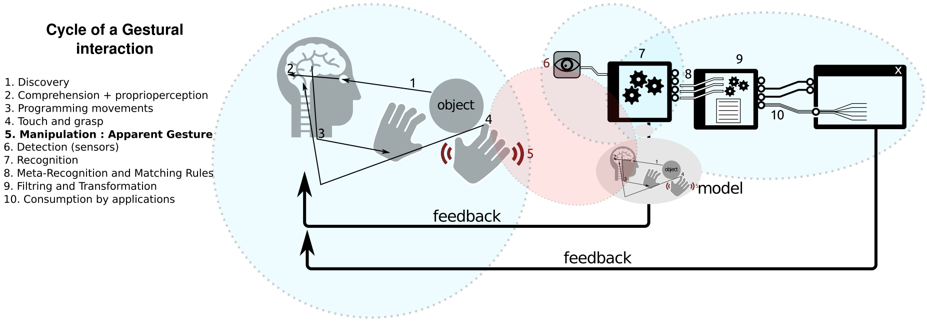

Cycle of a Gestural interaction
1. Discovery
2. Comprehension + proprioperception
3. Programming movements
4. Touch and grasp
5. Manipulation : Apparent Gesture
6. Detection (sensors)
7. Recognition
8. Meta-Recognition and Matching Rules
9. Filtring and Transformation
10. Consumption by applications
object
1
2
3
4
5
6
7
8
9
10
X
model
feedback
feedback

Fig. 2. The cycle of a gestural interaction, including the human, the hand, the object, the sensors, the handling daemons, and applications

3.2 Norman's Theory of Human Action as an Organizing Method

From a HCI point of view, and as the human is the generator of gestures, we propose the use of Norman's theory of Human Action Cycle [25] to position the areas of research together on the cycle. Norman describes seven stages-of-action in his model, during an experiment like ours, all these stages are fullfiled as it is included in his description of an interaction. The seven stages can be divided in 3 parts: Goal formation, Execution, and Evaluation. The perception can be in the Goal formation part. The Execution having translation of goals into a set of tasks can be attached to gestemes, the sequencing of tasks into action sequence can be attached to kinemes and reaching, the execution can be through the grasping and manipulation. The evalution part may include the feedback in a non-exhaustive way, and which is not in the scope of this paper.

3.3 The Urge to Link with the HCI Discipline

The goal of connecting areas together means understanding the problem deeper. This understanding can be transformed later into a model able to receive a partial amount of the hand information and then can extrapolate the missing data. We believe that a model based on the shared knowledge from different fields can satisfies the naturality question [12].

3.4 The Naturality from EMG and fMRI

The electromyogram (EMG) is the electrical signal detectable by electrodes on the skin of the muscle. And the functional magnetic resonance imaging (fMRI) is a neuroimaging technique that measures brain activity from the changes in blood flow coupled with neuronal activation. The brain activation during an operation allows the detection of what makes an operation stressful. Ehrsson et al. [9] studied forces and brain activation during a power and precision grip manipulation. The results show that even if the power grip forces are higher, the left-sided brain activity is the principal activation and it is low, while in precision grip with small forces generated, both sides and more regions are activated. We think that the more an operation is natural, the less zones are activated in the brain and the less stress it generates.

Other studies confirm the gesture and speech being proceed by the same neural system, should they be manipulated and detected using similar methods?

3.5 Comparison with a Legacy Device: The Mouse

Gesture technologies have always tried to compare to the mouse. Mainly claiming that they are able to ditch the mouse from its current dominent position in computing. These claims have been demonstrated to be wrong. The mouse is currently more efficient for the usual tasks [3]. The comparison was trying to beat the mouse.

The specificity of the mouse is that manipulating the mouse don't differ a lot from the hand rest position. While gesture systems tend to trigger fatigue and disconfort due to the gorilla arm effect. The rest position of the hand means less activation in the brain and less stress.

3.6 Enaction and Vicarious Learning

There are two ways of organizing knowledge in the brain, either enactive knowledge through action and motor skills like manipulating objects. Or learning by observation. Do we manipulate objects gestures by doing ourselves or by watching others do?

4 Prediction of Gestures

4.1 Affordance of Object Grasping

In a grasp, we can gather rich information in relation to the object properties, the setting, the relationship, the goal, and the anatomy of the user [37]. According to the previous studies on grasping, the way we grasp an object may be predicted according to the opposition plans, and in a virtual environment with virtual objects, we can simplify the model even more by ignoring a part of the grasp properties like the relationship and the goal.

Opposition plans [16,17] gives us hints about where fingers would be positioned for an object. In a virtual environment, we put the focus on pushing the user into a pad opposition as there is no object tactile feedback. This affordance of object grasping prepares the way into manipulation and thus gesture manipulation affordance.

4.2 Affordance of Gesture Manipulation

Manipulating an object in a virtual environment means moving the hands and stroking in the 3D space. The affordance of a manipulative gesture for an object is related to the affordance of grasping on the same object. The prediction of gestures is not always possible. But for many cases, the hand posture before reaching the object is relevent for what the user is willing to do. The form of the object, the hand posture before grasp and the position from where the hand through the limb reaches the object is a telling factor of the user gesture intention. These three, in some cases, remove many alternative possibilities of what is remaining to recognize as in figure 3. A user reaching a pawn object in a chess game from the upper side means he is willing to move it to another place. The scaling of the object is impossible, and the form of the pawn which has a base eliminates the possibility of rotation.

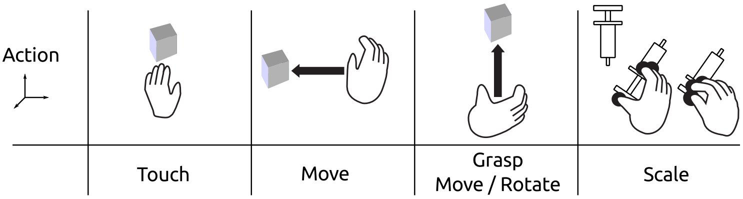


Fig. 3. Cases of gestures prediction

5 Simplifying of Gesture Manipulation Recognition

5.1 Role of the Hand and the Object

In our manipulation study, and in order to simplify the process, we suppose that the hand is the human tool used to act on a object whose existence is required. The notion of naturality of the interaction can not be processed without an object receiving the hand touch, grasp and movements.

5.2 Definition of Multitouch Gestures in 4 Classes

Multi-touch gestures are generally expressed in a wide form of possibilities, each one has a separate name which add a lot of complexity in definition and in use [41]. A burst nomination of multitouch gestures does not facilitate neither their user nor building a recognizer engine for them. With the hypothesis of having an object as a direct receiver of the gesture, we can organize the multi-touch gestures as shown by the figure 4 into four principal families[1]. Then, in order to detect sub-categories, we compute properties like the number of fingers, speed and time. The factorization of gestures into a small set of categories is previously cited by a work of Reisman et al. [30] as Rotate-Scale-Translate interaction.

5.3 Definition of Spatial Gestures in 4 Classes

We have questioned spatial gestures classes after handling multi-touch ones. We wanted to know whether it is possible to do the same classification into basic families as in figure 5. The work by Bullock et al. [7] made a specific classification for human maniplation, but using only rotation and translation. In his work, squeezing the syringue is considered as a translation task. We take the object into consideration, and if it is deformable, then a new category is required and in this case is the scaling gesture.

Meanwhile, not all gestures that exist can be expressed using one of the classes if the hypothesis of the receiver object is not satisfied. Non natural gestures that cannot be guessed unless being teached in advance like the Cyclostar by Malacria et al. [18] for dragging and zooming cannot be expressed by our system.

[1] Using multi-touch figures from `http://www.lukew.com/ff/entry.asp?1073`

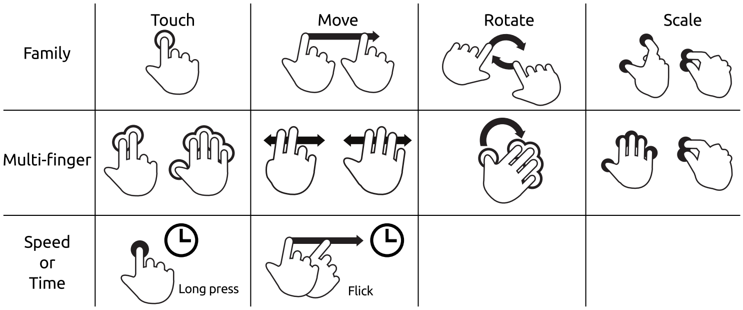


Fig. 4. Families of multi-touch gestures and specific gestures depending on number of fingers or the speed

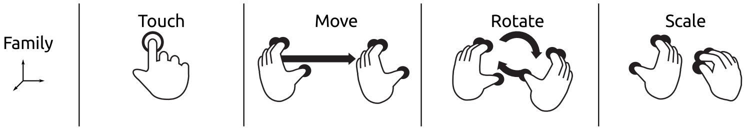


Fig. 5. Families of spatial gestures

5.4 Detecting Specific Gestures and the Points of View

Four classes of gesture is not enough to detect all gestures. We need more detailed characteristics allowing fine differentiation. From the characteristics we can list the number of fingers, the speed of the gesture, the delay in operation, the size of the interaction zone, the central point.

How a gesture recognition is handled differ from one system to another. Many systems wait for the gesture to start, record its stroke until the gesture finishes, then starts the recognition. The recognition results are only available at the end. Other more advanced systems, as the ones implemented in smartphones, don't wait until a gesture finishes, they start recognition when a fixed threshold is reached and each chunk of small movement beyond the threshold is decided from the four basic classes. Sometimes many decisions are fired at the same time, letting the final decision to meta recognition tools. In the case of Linux, recognition of long strokes as standalone gesture needs a meta system capable of matching predefined rules with the chunks feed.

6 Conclusion and Future Work

In this paper, we have explored most of the areas of research around gestures, in an attempt to extend the view of the problem to a wider public. We have shown that the recognition of gestures needs first a recognition of objects properties, human grasping, understanding of how we see and decide. The context and scenario of a gesture manipulation experience and the Norman's theory of human action helped us organize these blocks of research and raise the question why some legacy devices still beat new sensors and input devices. We have presented why the prediction of user gestures from hand posture during the grasping can improve current gesture recognizers. We have presented a taxonomy for classifying gestures either in 2D or in 3D in a few main classes and then detect sub-categories using a set of properties. Future work will target a deeper study to strengthen the links between the presented research areas and the developement of a real prototype to experiment the theoretical notions presented.

References

1. Aglioti, S., DeSouza, J.F., Goodale, M.A.: Size-contrast illusions deceive the eye but not the hand. Current Biology: CB 5(6), 679–685 (1995)
2. Beaudouin-Lafon, M.: Instrumental interaction: An interaction model for designing post-wimp user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 446–453. ACM (2000)
3. Bérard, F., Ip, J., Benovoy, M., El-Shimy, D.: Did Minority Report get it wrong? Superiority of the mouse over 3D input devices in a 3D placement task. IFIP (2009)
4. Bertranne, D.: Praxies idéomotrices corporelles: Création d'un test d'imitation de postures asymboliques (2007)
5. Bruno, N., Bernardis, P.: Dissociating perception and action in Kanizsa's compression illusion. Psychonomic Bulletin & Review 9(4), 723–730 (2002)
6. Bullock, I., Ma, R., Dollar, A.: A Hand-Centric Classification of Human and Robot Dexterous Manipulation. Ieeexplore.ieee.org, section III, 1–16 (2012)
7. Bullock, I.M., Dollar, A.M.: Classifying human manipulation behavior. In: 2011 IEEE International Conference on Rehabilitation Robotics (ICORR), pp. 1–6. IEEE (2011)
8. Cutkosky, M.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Transactions on Robotics and Automation 5(3), 269–279 (1989)
9. Ehrsson, H.H., Fagergren, A., Jonsson, T., Westling, G., Johansson, R.S., Forssberg, H.: Cortical activity in precision-versus power-grip tasks: An fmri study. Journal of Neurophysiology 83(1), 528–536 (2000)
10. Feix, T., Pawlik, R., Schmiedmayer, H.-B., Romero, J., Kragic, D.: A comprehensive grasp taxonomy. In: Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation, pp. 2–3 (2009)
11. Frohlich, B., Tramberend, H., Beers, A., Agrawala, M., Baraff, D.: Physically-based manipulation on the responsive workbench. In: Proceedings of the IEEE Virtual Reality, pp. 5–11. IEEE (2000)
12. Gamberini, L., Spagnolli, A., Prontu, L., Furlan, S., Martino, F., Solaz, B.R., Alcañiz, M., Lozano, J.A.: How natural is a natural interface? An evaluation procedure based on action breakdowns. Personal and Ubiquitous Computing (October 2011)
13. Gustafson, S., Bierwirth, D., Baudisch, P.: Imaginary interfaces: Spatial interaction with empty hands and without visual feedback. In: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, pp. 3–12. ACM (2010)
14. Harrison, C., Schwarz, J., Hudson, S.E.: Tapsense: Enhancing finger interaction on touch surfaces. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 627–636. ACM (2011)

15. Hilliges, O., Izadi, S., Wilson, A., Hodges, S., Garcia-Mendoza, A.,, B.: Interactions in the air: Adding further depth to interactive tabletops. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 139–148. ACM (2009)
16. Iberall, T., Bingham, G., Arbib, M.: Opposition space as a structuring concept for the analysis of skilled hand movements. Experimental Brain Research Series (1986)
17. MacKenzie, C.L.C., Iberall, T.: The grasping hand. Elsevier (1994)
18. Malacria, S., Lecolinet, E., Guiard, Y.: Clutch-free panning and integrated pan-zoom control on touch-sensitive surfaces: The cyclostar approach. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 2615–2624. ACM, New York (2010)
19. Marzke, M.W., Wullstein, K.L.: Chimpanzee and human grips: A new classification with a focus on evolutionary morphology. International Journal of Primatology 17(1), 117–139 (1996)
20. McNeill, D.: Gesture and thought. University of Chicago Press (2008)
21. Mine, M.R., Brooks, J. F.P., Sequin, C.H.: Moving objects in space: Exploiting proprioception in virtual-environment interaction. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, pp. 19–26. ACM Press/Addison-Wesley Publishing Co., New York (1997)
22. Nacenta, M.A., Kamber, Y., Qiang, Y., Kristensson, P.O.: Memorability of predesigned and user-defined gesture sets. In: CHI, pp. 1099–1108 (2013)
23. Napier, J., Tuttle, R.: Hands. Natural science. Princeton University Press (1993)
24. Napier, J.J.: The prehensile movements of the human hand. Surger 38(4), 902–913 (1956)
25. Norman, D.: The design of everyday things (2002)
26. Norman, D.: Natural user interfaces are not natural. Interactions, 6–10 (2010)
27. Norman, D.A., Draper, S.W.: User Centered System Design; New Perspectives on Human-Computer Interaction. L. Erlbaum Associates Inc., Hillsdale (1986)
28. Nowak, D., Hermsdörfer, J.: Sensorimotor Control of Grasping: Physiology and Pathophysiology. Cambridge University Press (2009)
29. Paulignan, Y., MacKenzie, C., Marteniuk, R., Jeannerod, M.: Selective perturbation of visual input during prehension movements. Experimental Brain Research 83(3), 502–512 (1991)
30. Reisman, J.L., Davidson, P.L., Han, J.Y.: A screen-space formulation for 2D and 3D direct manipulation. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, UIST 2009, p. 69 (2009)
31. Roberts, A.I., Vick, S.-J., Roberts, S.G.B., Menzel, C.R.: Chimpanzees modify intentional gestures to coordinate a search for hidden food. Nature Communications 5 (2014)
32. Roffman, I., Savage-Rumbaugh, S., Rubert-Pugh, E., Ronen, A., Nevo, E.: Stone tool production and utilization by bonobo-chimpanzees (pan paniscus). Proceedings of the National Academy of Sciences 109(36), 14500–14503 (2012)
33. Sève-Ferrieu, N.: Neuropsychologie corporelle, visuelle et gestuelle: Du trouble à la rééducation. Elsevier Masson (2005)
34. Valkov, D.: Interscopic Multi-Touch Environments. dfki.de, 339–342 (2010)
35. Victor, B.: A Brief Rant on the Future of Interaction Design (2011), `http://worrydream.com/ABriefRantOnTheFutureOfInteractionDesign/`
36. Vinayavekhin, P.: Dexterous manipulation planning from human demonstration. PhD thesis, University of Tokyo (2009)

37. Wimmer, R.: Grasp sensing for human-computer interaction. In: Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction, pp. 221–228. ACM (2011)
38. Wing, A., Haggard, P.: Hand and brain: The neurophysiology and psychology of hand movements
39. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009, pp. 1083–1092. ACM, New York (2009)
40. Xu, J., Gannon, P.J., Emmorey, K., Smith, J.F., Braun, A.R.: Symbolic gestures and spoken language are processed by a common neural system. Proceedings of the National Academy of Sciences 106(49), 20664–20669 (2009)
41. Yee, W.: Potential limitations of multi-touch gesture vocabulary: Differentiation, adoption, fatigue. In: Jacko, J.A. (ed.) HCI International 2009, Part II. LNCS, vol. 5611, pp. 291–300. Springer, Heidelberg (2009)

iPanel: A Computer-Vision Based Solution for Interactive Keyboard and Mouse

H. Chathushka Dilhan Hettipathirana[1] and Pragathi Weerakoon[2]

[1] Department of Computing
Informatics Institute of Technology, Sri Lanka.
Collaboration with University of Westminster
h.hettipathirana@my.westminster.ac.lk
[2] Informatics Institute of Technology, Sri Lanka
Pragathi.w@iit.ac.lk

Abstract. This paper represents an implementation of a computer vision based interface; iPanel which employs an arbitrary panel and tip pointers as a spontaneous, wireless and mobility device. Also the proposed system can accurately identify the tip movements of the panel and simulate the relevant events on the target environment. By detecting the key pressing, mouse clicking and dragging actions, the system can fulfill many tasks. Therefore, it enables users to use their fingers naturally to interact with any application as well as with any mobility enabled devices.

Keywords: Computer vision, Human computer interaction, gesture recognition, optical character recognition, wearable computing.

1 Introduction

Human computer interaction is the technique of studding the relations between people and computer or computer mediated information. Thus it involves the design, development and evaluation of models, systems and applications from a human-centered perspective. Since its inception in the 1980s, HCI has been primarily concerned with designing more usable computer systems, attractive conventional computing devices, be it the computer desktop, the Web, or the mobile phone. It evaluates the existing designs and shows how to improve them. And, it attempts to apply its methods to design more user friendly systems from the start. Human-computer interaction comprise many sub domains such as gesture reconstruction, event detection, video tracking, object recognition, learning, indexing, motion estimation, and image restoration. Each sub domain is a unique concept of computer vision and it attempt to address a particular area of HCI, where the computers are pre-programmed to solve tasks or the interactions (e.g. touch screens, tablet PCs).

It has been identified and observed that many researches are adopting gesture reconstruction and ended up with implementing excellent results (e.g. Microsoft is researching on how user can interact with computers or computational devices in more efficient and user friendly manner [7]. Thus gesture recognition, sensor based

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 36–47, 2014.

interactions and augmented realities are becoming more and more popular. This is the main reason that gesture recognition is more important and required to simplify the user interaction with the computers of computational devices. As an example, several people are discussing in a meeting room using a large display. They may need to draw, to suggest their ideas. However, it is unrealistic to facilitate every user a keyboard and a mouse. Even more, in a large lecture room, the lecturer sometimes needs to write down something on a small whiteboard. However, the audience far from him or remote audience may not be able to see clearly what he writes on the board. Therefore, need for a vision based system is necessary to analyze and understand what the lecturer writes and retrieve on a remote display, while avoiding bandwidth constraints. Furthermore, most of the smart mobile provide a QWERTY keyboard with tiny keys. It is really difficult to type with those keys. Yet, providing a large screen would lead to unnecessary problems such as size of the phone. In order to address the above a vision based solution has been suggested.

2 Human Computer Interaction

Human-Computer Interaction (HCI) is a technology researching on people, computer and the communications between them. Designing interactive computer systems to be effective, efficient, easy and enjoyable to use is important, so that people and society may realize the benefits of computation based devices such as mouse or keyboard. According to [2] use of these devices are recognized way of interaction with interfaces based on Window, Icon, Menu and Pointer (WIMP) paradigms which have succeeded for decades. Eventually software interfaces have got improved and interactive, lot of effort and code has been put behind the development of interactive software. Nonetheless, the use of traditional computational devices such as keyboard and mouse do not provide an expected way of interaction.

Most of the innovative interfaces such as Microsoft Surface [11] tend to support multi user interaction and are recognized to be augmented reality based products. Due to that reason there has been lot of concern on development on alternative and natural interaction methods to support interaction with such interfaces, while supporting for the existing conventional computing devices. Thus human-computer interaction design is human centered approach where human is given more priority. Also the previous work done by [2] note that "The human, the user, is, after all, the one whom computer systems are designed to assist. The requirements of the users should therefore be our first priority".

3 Computer Vision and Gesture Recognition

In its most general meaning, a gesture is any physical configuration of the body, whether the person is aware of it or not, whether performed with the entire body or just the facial muscles, whether static in nature or involving a movement. In the computer vision literature, gesture usually refers to a continuous, dynamic motion, whereas a posture is a static configuration.

Computer vision based gesture recognition is a sub domain of HCI and it comprises of a wide range of shapes, motions and texture based variations. And also it includes different gesture recognition methods such as applying Fourier transform ([6]), wavelet transform ([4]) or Principal Component Analysis (PCA) ([16]) on images, Edge orientation histogram, temporal templates ([17]) and oriented rectangular patches ([8]). Thus, it is very important to study on these gesture recognition method differences and select good features to define simple and natural gestures which will be easily adoptable to be used for human computer interaction. Recognition of gestures includes object detection, motion analysis, extraction of features, and machine learning. Besides real time recognition has been a stimulating task in all the time. Efficient recognition of positions can be adopted for an effectively simulation of keyboard events. For example, posture classification refers to the estimation of finger configurations, that is, the ability to distinguish a fist from a flat palm and so on. As described by [9] describes different kinds of gestures from what has become known as Kendon's Gesture Continuum. However practical limitations due to varying luminous conditions and complex backgrounds can exist. Thus, finger tracking and use of non-geometric features such as color and outline are also important for a reliable and strong recognition. There is an extensive body of related computer vision research which could fill many books. Here, author has summarized the major works that could fit the bill for real-time user interface operation through hand gesture recognition in a fairly unconstrained environment. To get an independent overview, the reader is referred to a paper by [5] a survey on "computer vision for interactive computer graphics" and an evaluation of the state of the art by [15]. Three common tasks for computer vision processing are; (1) The detection of the presence of an object in an image. (2) The spatial tracking of a once-acquired object over time. (3) Recognition of one of many object types

3.1 Preprocessing

Preprocessing is the progression of color space conversion, edge detection, morphological operations, noise removal and thresholding. Therefore, it is implemented in almost every vision based algorithms as an entry point to be suitable for the image processing. According to [14] color space conversion, noise removal, edge detection and outlines extraction has to be carried out during the preprocessing stage.

3.2 Detection

As [3] showed in early neuron scientific experiments the human visual system has the amazing ability to detect hands in almost any configuration and situation, and possibly a single "hand neuron" is responsible for recording and signaling such an event. The computer vision researches have not quite yet achieved this goal. However, it is vital that a hand is supposed to function as an input mechanism to the computer is strongly and consistently perceived in front of arbitrary background, for the reason that all further stages and functionalities depends on it. Object detection of artificial objects, such as colored sticks as in [18] can achieve very high detection rates regardless of low false positive rates. According to [20] face detection has attracted a great amount of interest and many methods relying on shape, texture, and/or temporal information have been thoroughly investigated over the years. Author has carried out some researches on finding hands in grey-level images based on their appearance and texture. As assert by [19] “Combining with skin color segmentation, view independent posture recognition can be used to detect hands. Since skin color segmentation has already limited the searching range, hand detection can be very efficient”. [12] demonstrated that, lately improved classifiers have succeeded compelling results for view and posture independent hand detection. However, most of the hand detection methods resort to less object-specific approaches and as an alternative employ color information (see, for example [21]), sometimes in combination with location priors (for example [10]), motion flow or background differencing (for example [13]).

3.3 Tracking

Background subtraction is very important for motion analysis and object tracking because of it’s a basic function that enables to build statistical model of background. And used for segmenting moving objects for the background. If the detection method is flexible and fast enough to operate at image acquisition frame rate, it can be used for tracking as well. However, tracking hands is extremely difficult since they can move very fast and their appearance can change enormously within a few frames. As [1] asserts that some of the most effective head trackers, for example, use a fixed oval shape model which is fast and appropriate for the inelastic head structure. Similarly, more or flexible hand models work well for a few select hand configurations and relatively static lighting conditions. Since tracking with an inflexible appearance model is not possible for hands in general, most approaches alternative to shape-free color information or background differencing as in the mentioned works by [10], and [13]. Other methods incorporate for example, texture and color information and can then recognize and track a small number of fixed shapes regardless of arbitrary backgrounds (for example, [22]). As per the research work, a particle filtering method is optimized for speed mean shift, and dynamic weights determine the blend of color with motion data. That explains, the faster the object moves, the more weight is given to the motion data, and slower object movements result in the color cue being weighted higher. Some of their performance is surely due to simple, however usually effective dynamical model (of the object velocity), which could add to the suggested solution as well. Object breakdown based on visual flow (for example, normalized graph cuts as proposed by [23]) can produce good results for tracking objects that display a limited amount of twists during global motions and, thus have a fairly unchanging flow ([24])

3.4 Skin Color

Skin color detection is widely used to detect hand configurations and thus it is very important in gesture recognition. Skin color classification is preferred for fast processing and due to its effective response to non-rigid objects such as hands.

Previously absorbed results shows it can be achieved only by skin color properties, for example, by [25] who used it in combination with a neural network to estimate gaze direction. [26] Demonstrate interface quality hand gesture recognition only with color segmentation means. Their method uses an HSV like color space, which is possibly beneficial to skin color identification.

The appearance of skin color differs mostly in intensity while the chrominance remains fairly consistent. Thus, and according to [27], color spaces that separate intensity from chrominance are suitable to skin segmentation when simple threshold-based segmentation is used. However, their results are based on a few images only, while a paper from [28] examined a huge number of images and found an excellent classification performance with a histogram-based method in RGB color space. It appears that very simple threshold methods or other linear filters accomplish better results in HSV space, while more complex methods, particularly learning-based, nonlinear models excel in any color space. [28] Also state that Gaussian mixture models are lower to histogram based approaches. This is true as long as a large enough training set is available. Otherwise, Gaussians can fill in for inadequate training data and achieve better classification results. [29] Showed that object tracking based on color information is possible with a method called CamShift which is based on the mean shift algorithm. These methods dynamically slide a “color window” along the color probability distribution to dynamically parameterize thresholding segmentation. A certain amount of lighting changes can be allocated with. Patches or drops of uniform color have also been used, especially in fairly controlled scenes. According to [30] achieve excellent segmentation with dynamic adaptation of the skin color model based on the observed image.

3.5 Contours Extraction

Contour processing is performed on images typically after performing edge detection or thresholding. Contour extraction is used after canny edge detection algorithm, to detect an inserted object using color cluster feature. In theory, the contour or outline of an object reveals a lot about its shape and orientation. If perfect segmentation is possible, comparison based on curve matching is a feasible approach to object classification. For example as [31] assert that, based on polar coordinates above can be done. One can benefit even more from curve descriptors that are invariant to scale differences and rigid transformations such as those by [32] and Shape Context descriptors. For less-than-perfect conditions however, more powerful 2D methods must be used. Those usually set on finding enough local clues in the image to place a shape model close to where the most likely placed of this shape can be found in the image data. Iterative methods frequently try to minimize an energy defined as images which are not aligned properly (far from an edge). For a hand in top view, these modes could theoretically be the movements of each finger. Statistical models of an object's 3D shape, often called “point clouds,” can also be built (as did, for example, [33]), but they shall not be further measured since their speed performance might drop. According to [34] took a popular approach and had their recognition method learn from extracted hand images instead of from actual photographs. During testing, edge data between the observation and the learned database are compared and 3D hand configurations can be estimated from 2D grey-level edges. According to their paper, matching takes less than a second for an approximate result, but too long for interactive frame rates. [35] assert that; detect hands uniquely in postures regardless of messy backgrounds. The distance between two curves or contours is the mean of the distances between each point on one curve and its closest point on the other curve.

4 Implementation

4.1 Hand Gesture Recognition

Hand gestures can be recognized with various means and varying fidelity. They are not in one particular identification technique, but with various sensing mechanisms. Hand detection for user interfaces must favor reliability over expressiveness: false positives are less tolerable than false negatives. Since detecting hands in arbitrary configurations is a largely unsolved problem in computer vision, the detector for iPanel allows reliable and fast detection of the hand in one particular posture from a particular view direction. Starting the interaction from this initiation pose is particularly important for a hand gesture interface that serves as the sole input modality as it functions as a switch to turn on the interface: without this and instead with an always-on interface, any gesture might inadvertently be interpreted as a command. The output of the detection stage amounts to the extent of the detected hand area in image coordinates. This software system is capable of detecting the human hand in monocular video, tracking its location over time, and recognizing a set of finger configurations (postures). It operates in real time on commodity hardware and its output can thus function as a user interface.

The software system that realizes the vision-based hand gesture recognition and allows for its utilization as a user interface consists of a number of software components that will be described in the following. iPanel main component pronounced “skindetector" is a library and the core gesture recognition module that implements all of the computer vision methods for detection, tracking, and recognition of hand gestures. This module receives the direct video feed from a camera and generated the analyzed gesture results to the main application. This application called WinTalk class library, which handles pipeline initialization and implements convenience functions. In addition to these runtime components, there is also an offline module that implements ANN (Artificial Neural Networking) training for the detection and recognition components.

The core module is a combination of recently developed methods with novel algorithms to achieve real-time performance and robustness. A careful orchestration and automatic parameterization is largely responsible for the high speed performance while multi-modal image cue integration guarantees robustness. Yet, initially an hard-coded values has been used to identify the hand gestures of the author in order to make sure that, development phase has not been misguided with different values. There are three stages: the first stage detects the presence of the hand in series of posture (It is required to have the vision interface always active since hand gestures which are used as mouse movements may be used as commands). Yet, identifying a series of postures could cause different errors in the application due to the misleading of generated events. However, the issue has been addressed by a different mechanism, and to be discussed in the following paragraphs. After this gesture based activation, the second stage serves as an initialization to the third stage, the main tracking- and posture recognition stage. This multi-stage approach makes it possible to take advantage of less general situations at each stage. Exploiting spatial and other constraints that limit the dimensionality and/or extent of the search space achieves better quality and faster processing speed. Author uses this at a number of places: the generic skin color model is adapted to the specifics of the observed user for posture recognition is positioned with fast model free tracking. However, staged systems are more prone to error propagation and failures at each stage. To avoid these pitfalls, every stage makes conservative estimations and uses multiple image cues (grey-level texture and local color information) to increase confidence in the results. "SkinDetector" assists as a library for gesture recognition that can be built into any windows application that demands a hand gesture user interface. However, it does not handle any user display-specific operations such as image acquisition or display. Thus, it requires some programming before it can be used. The final output of the vision system indicates for every frame the 2D location of the hand with the number of fingers if is tracked, or that it has not been detected yet. If the dominant hand's posture is recognized, it is described with a string identifier as a classification into a set of predefined, recognizable hand configurations.

4.2 Hand Detection

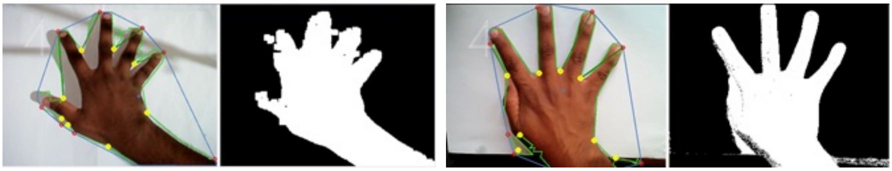


Fig. 1. (1) Actual output of the custom YCrCb based skin detection algorithm. (2) Output - Hsv based skin detection algorithm.

YCrCb / Custom YCrCb Based Skin Detection. Y′ is the luminance component and Cr and Cb are the red-difference and blue-difference chroma components. Y′ (with prime) is distinguished from Y which is luminance, meaning that light intensity is nonlinearly encoded based on gamma corrected RGB primaries. Yet, identifying the skin using YCrCb algorithm is very challenging because of the difficulty of identifying the correct Ycc color range.

Hsv Based Skin Detection. In the HSV color system, the colors of maximum saturation are not necessarily pure. The HSV, an alternate representation of a given RGB color space, and the saturated colors in HSV are in fact the colors bordering the corresponding RGB triangle in the chromaticity diagram. For this reason, the HSV color system has be identified as device dependent, meaning that it is not an absolute colorimetric space, but relative to the gamut of the RGB color space it describes. The third coordinate in HSV has the value or brightness; black has zero brightness. Starting from the hues disk one can imagine the HSV space as a collection of hues circles with varying color value, one on top of the other and of the same size or of sizes diminishing with value. The Fig.1 (2) shows, the identified hand postures using the Hsv skin detection algorithm.

Both algorithms custom YCrCb and Hsv based skin detection can used to identify the user skin of hand gestures. However, most of the researches shows, that the source frame is converted to both Ycc and Hsv color spaces and observed that Hsv color space provides better segmentation in practice over Ycc. Furthermore, it has been tested with different skin colors during various times of the day. The reason was Hsv provides clear separation of luminance and chrominance. Yet, it is more vital to train the algorithms through an ANN approach to recognize any type of skin in order to ensure that the every skin component has been identified in the image stream.

4.3 Recognition

In order to place flock features, initially context hulls are recognized through the identified counters, a centered point on top of the detected skin and a clock wise rotation. The Flock of Features follows small grey-level image artifacts. A weak global constraint on the features' locations is enforced, keeping the features tightly together. Features that are not likely to still be on an area of the hand appearance are relocated to close proximity of the remaining features and on an area with high skin color probability. This technique integrates grey-level texture and dimensionless color cues, resulting in more robustness towards tracking disturbances cause by background artifacts. From the feature locations a small area is determined that is scanned for the key postures that recognition is attempted for. Once prefect detection has been performed events are bind with the fingers.

4.4 Execution

In order to perform key pressing events author has been developed on his own algorithm though out series of researches. The identified method was to capture three different postures of figure movements and analyze them to meet the requirements of performing key press actions. In the algorithm each of the fingertip positions are stored on a collection along with a finger number. When a key press is performed through the gestures it is noticeable that particular fingertip's positions are change through Y axis of the screen while X axis on constant (But X axis could be slightly changed based on the movement). For example, if the initial X and Y position of a fingertip is X = 150, Y = 200; in the event of performing key press tip positions are changed to X = 155, Y = 265 and then again it changes to a position X = 152, Y = 225. [37] Discussed a similar algorithm in there research related to wearable multi-touch interaction. According to their solution three dimensional (X, Y and Z) fingertip detection has been used. However, author insists to use his own algorithm in order to minimize the complexity and to improve effective mechanism to identify the event of key press.

4.5 Finding Rectangles and Identify Characters

Finding rectangles involves finding axis aligned rectangles in binary image. These rectangles will help in separating the area of the image that contains text from the rest of the image. Even though, this is a huge process to carry out Tesseract API, provide all the required processing and identifying algorithms. Therefore, no custom algorithms have been developed in order to analyze images or to identify. Furthermore, to identify language dependent characters Tesseract API required initializing with relevant tesseract data and language dictionaries. Following Fig.4 (1) illustrates characters identified by the module.

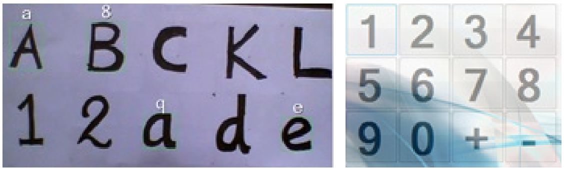


Fig. 2. (1) Characters identified by the OCR module. (2) User designed key arrangement and module generated virtual keyboard.

5 Evaluation

5.1 Expert Evaluation

An expert evaluation was conducted on the research to measure the validity and appropriateness of the approaches, methodologies, and models used by the author. The expert evaluation process has been started at the requirement gathering stage in order to understand and evaluate the user requirements. Then a thorough analysis and evaluation has been conducted in the design phase to avoid expensive mistakes, since the design can be altered prior to any major recourses commitment. Therefore, sample design and prototypes have been provided though it is difficult to get an accurate assessment of the experience of interaction. [4] assert that four different approaches that could adopt to expert analysis: (1) cognitive walkthrough, (2) heuristic evaluation, (3) the use of models and (4) use of previous work. The author uses three approaches in order evaluate system properly.

Cognitive walkthrough approach has been adopted and it evaluation is the code walkthrough to check certain characteristics (for example, that coding style is adhered to proper coding standards). The general idea behind the heuristic evaluation is that several evaluators independently critique the system to identify potential usability problems and to understand the severity of the problems. The final approach used to evaluate the iPanel system is "use of previous work". Expert knowledge on previous experience also has been involved in order to provide the feedback for the system.

Furthermore, their comment regarding research holds a significant impact for the future enhancements on the application. A questioner has been used to gather expert reviews on the iPanel. With the intension of understanding the accuracy of the output, and the suitability of implemented algorithms, along with identifying the independency on user hand gestures; 71% of the participation has been awarded “Excellent” for the accuracy of the hand gesture recognition algorithm, while 28% stated “Good”. Also it is notable to discuss all the experts have been agreed that, Hsv based skin color detection along with structural analysis can provide an effective person independent hand and finger detection.

6 Conclusion

Author has been developed the iPanel, a computer vision system for recognition of hand gestures in real-time and perform key strokes in order to allow real time interaction with a virtual keyboard. Novel and improved vision methods had to be devised to meet the strict demands of user interfaces. Tailoring system and applications for hand motions within a comfort zone that we have established improves user satisfaction and helps optimizing the vision methods. Multiple applications demonstrated such as windows, web, and mobile showed iPanel in action and indicated that it adds to the options of interaction with non-traditional computing environments.

References

1. Birchfield, S.: Elliptical head tracking using intensity gradients and color histograms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-25, pp. 232–237. IEEE, Stanford University (1998)
2. Dionisio, C.R.P., Cesar Jr., R.M.: A project for hand gesture recognition. In: Symposium on Computer Graphics and Image Processing, p. 345. IEEE, Sao Paulo University (2000)
3. Desimone, R., Albright, T.D., Gross, C.G., Bruce, C.: Stimulus-Selective Properties in Inferior Temporal Neurons in the Macaque. Journal of Neuroscience 4(8), 2051–2062 (1984)
4. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: Human-Computer Interaction, 3rd edn. Peaeson, New Delhi (2003)
5. Freeman, W.T., Anderson, D.B., Beardsley, P.A., Dodge, C.N., Roth, M., Weissman, C.D., Yerazunis, W.S.: Computer Vision for Interactive Computer Graphics. In: Computer Graphics and Applications. IEEE Computer Graphics and Applications, pp. 42–53 (May-June 1998)
6. Harding, P.R.G., Ellis, T.: Recognition Hand Gesture Using Fourier Descriptors. In: IEEE Pattern Recognition, August 23-26, vol. 3, pp. 286–289. Buckinghamshire Chilterns Univ. Coll., UK (2004)
7. Harper, R., Rodden, T., Rogers, Y., Sellen, A.: Being Human - human-computer interaction in the year 2020. Microsoft Research, Cambridge (2008)
8. Ikizler, N., Duygulu, P.: Human Action Recognition Using Distribution of Oriented Rectangular Patches. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 271–284. Springer, Heidelberg (2007)
9. Kendon, A.: Cross-cultural perspectives in nonverbal communication. In: How Gestures can Become Like Words, pp. 131–141 (1988)

10. Kurata, T., Okuma, T., Kourogi, M., Sakaue, K.: The Hand Mouse: GMM Hand-color Classification and Mean Shift Tracking. In: Second Intl. Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (July 2001)
11. Microsoft Cooperation: Microsoft Surface. Microsoft (2012), `http://www.microsoft.com/surface/en/us/default.aspx` (accessed October 08, 2012)
12. Ong, E.J., Bowden, R.: A Boosted Classifier Tree for Hand Shape Detection. In: Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition, pp. 889–894. IEEE (2004)
13. Segen, J., Kumar, S.: GestureVR: Vision-Based 3D Hand Interface for Spatial Interaction. In: The Sixth ACM Intl. Multimedia Conference. ACM (September 1998)
14. Shen, W., Wu, L.: A method of billiard objects detection based on Snooker game video. In: Future Computer and Communication (ICFCC), Beijing, China, May 21-24, vol. 2, pp. 251–255. IEEE (2010)
15. Turk, M.: Computer Vision in the Interface. ACM Communications 47(1), 60–67 (2004)
16. Vafadar, M., Behrad, A.: Human Hand Gesture Recognition using image processing for Human-Computer Interaction. In: IEEE Information and Knowledge Technology (2007)
17. Vafadar, M., Behrad, A.: Human Hand Gesture Recognition Using Motion Orientation Histogram for Interaction of Handicapped Persons with Computer. In: IEEE Image and Signal Processing, pp. 378–385 (2008)
18. Wilson, A., Shafer, S.: UI for Intelligent Spaces. In: XWand. ACM (2003)
19. Wu, Y., Huang, T.S.: View-independent recognition of hand postures. In: IEEE Computer Vision and Pattern Recognition. Beckman Inst. for Adv. Sci. & Technol., vol. 2, pp. 88–94. Illinois Univ., Urbana (2000)
20. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting Faces in Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002)
21. Zhu, X., Yang, J., Waibel, A.: Segmenting Hands of Arbitrary Color. In: Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition. Interactive Syst. Labs., pp. 446–453. Carnegie Mellon Univ., Pittsburgh (2000)
22. Bretzner, L., Laptev, I., Lindeberg, T.: Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering. In: Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition, pp. 423–428. IEEE, Washington, D.C (2002)
23. Shi, J., Malik, J.: Motion segmentation and tracking using normalized cuts. In: Sixth International Conference on Proc. Computer Vision, January 4-7, pp. 1154–1160. IEEE, Berkeley (1998)
24. Quek, F.K.H.: Unencumbered Gestural Interaction. IEEE Multimedia 4(3), 36–47 (1996)
25. Stiefelhagen, R., Yang, J.: Gaze tracking for multimodal human-computer interaction. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1997, Karlsruhe University, April 21-24, vol. 4, pp. 2617–2620. IEEE (1997)
26. Kjeldsen, R., Kender, J.: Finding Skin in Color Images. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 312–317 (October 1996)
27. Zarit, B.D., Super, B.J., Quek, F.K.H.: Comparison of Five Color Models in Skin Pixel Classification. In: Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Illinois University, Chicago, IL, pp. 58–63. IEEE (September 1999)
28. Jones, M.J., Rehg, J.M.: Statistical Color Models with Application to Skin Detection. Int. Journal of Computer Vision 46(1), 81–96 (2002)

29. Comaniciu, D., Ramesh, V., Meer, P.: Real-Time Tracking of Non-Rigid Objects Using Mean Shift. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 142–149. IEEE (2000)
30. Zhu, Q., Cheng, K.T., Wu, C.T., Wu, Y.L.: Adaptive Learning of an Accurate Skin-Color Model. In: Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition, Santa Barbara, CA, USA, May 17-19, pp. 37–42. IEEE (2004)
31. Hamada, Y., Shimada, N., Shirai, Y.: Hand Shape Estimation Using Sequence of Multi-Ocular Images Based on Transition Network. In: VI 2002 (2002)
32. Gdalyahu, Y., Weinshall, D.: Flexible Syntactic Matching of Curves and its Application to Automatic Hierarchical Classification of Silhouettes. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(12), 1312–1328 (1999)
33. Heap, T., Hogg, D.: Towards 3D Hand Tracking Using a Deformable Model. In: Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition, Leeds University, October 14-16, pp. 140–145. IEEE (1996)
34. Athitsos, V., Sclaroff, S.: Estimating 3D Hand Pose from a Cluttered Image. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston University, MA, USA, June 18-20, vol. 2, pp. 432–439. IEEE (2003)
35. Thayananthan, A., Stenger, B., Torr, P.H.S., Cipolla, R.: Shape Context and Chamfer Matching in Cluttered Scenes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, USA, June 18-20, vol. 1, pp. 127–133. IEEE (2003)
36. Xie, J.: Optical Character Recognition Based on Least Square Support Vector Machine. In: Intelligent Information Technology Application, IITA 2009, School of Electronics, Jiangxi University of Finance and Economics, Nanchang, China, November 21-22, vol. 1, pp. 626–629. IEEE (2009)
37. Harrison, C., Benko, H., Wilson, A.D.: OmniTouch: Wearable Multi-touch Interaction Everyware, pp. 16–19. ACM, Santa Barbara (2011)

Adding Multi-Touch Gesture Interaction in Mobile Web Applications

Shah Rukh Humayoun[1], Franca-Alexandra Rupprecht[1],
Steffen Hess[2], and Achim Ebert[1]

[1] Computer Graphics and HCI Group
University of Kaiserslautern
Gottlieb-Daimler-Str., 67663, Kaiserslautern, Germany
{humayoun,ebert}@cs.uni-kl.de, frupprec@rhrk.uni-kl.de
[2] Fraunhofer IESE
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
steffen.hess@iese.fraunhofer.de

Abstract. This paper describes the MTGest framework, an open library for adding multi-touch gesture interaction to HTML-based mobile web applications. MTGest was used in a comparative study to evaluate the multi-touch gesture interaction in a mobile web application in comparison to a native iOS mobile application. The results indicates that in most cases the web based gestures efficiency is either approximately the same or higher than the iOS-based app. The study was carried out as an initial experiment using isolated gestures, targeting the iOS platform only. For generalizing the results there is a need to perform detailed user evaluation studies with different platforms and for more complex interaction scenarios.

Keywords: Smart Devices, Smartphones, Tablets, Mobile Apps, Web Apps, Multi-Touch Gesture, Interaction Design, Mobile Environments.

1 Introduction and Related Work

Due to the popularity of smart devices and mobile applications (also called mobile apps), companies are offering more and more their product support in them. In order to reach to a broad pool of potential users, companies need to develop their applications for many of the existing mobile platforms (e.g., Google Android [2], Apple iOS [1], Microsoft Windows Phone [5]). Developing mobile apps separately for each platform is costly and time consuming while keeping focus on just one platform reduces the number of accessible users. We found in our previous study [3] that developing mobile apps through cross-platform development frameworks (where the application is developed once and deployed on the possible target platforms) still lacks in many aspects such as: few platforms support, only partially support of the targeted platform's interaction schema, far behind the native development environments.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 48–57, 2014.

Mobile web applications (also called *mobile web apps*) are based on web technologies like HTML, CSS, and JavaScript. HTML5[1] as an upcoming standard by W3C[2] provides the possibility of offline browsing as well as accessing many device resources (such as localization services or sensors); hence, HTML5-based web apps can be used as an alternative to the native mobile apps in many cases. There are many benefits of this approach such as: requires less efforts and resources for developing, supports all platforms, provides a consistent user experience and interaction concept across all platforms. However, HTML5 lacks built-in tags for the functionality of current multi-touch gesture interaction paradigm, which reduces the web apps' applicability compared to the native mobile apps. HTML5 provides a set of interfaces for the basic touch events but does not have built-in functions to support directly most of the current multi-touch gestures. In this work, we focus on adding the current multi-touch gesture interaction paradigm support in HTML-based documents in order to provide the current multi-touch gestures (e.g., *double tap, swipe, flick, zooming, rotation*) in mobile web apps.

We provide this support through our developed library, called **MTGest** (**M**ulti-**T**ouch **Gest**ures) library. This library enables mobile web apps the provision of multi-touch gesture interaction inside them in order to give the expected user experience and interaction concept of the current mobile paradigm across all platforms. For checking the efficiency and user satisfaction with the provided multi-touch interaction support in mobile web apps by our MTGest library, we conducted a user evaluation study. In this conducted study, users from different backgrounds and expertise tried two simple apps, i.e., one mobile web app based on MTGest library and the other one a iOS-based mobile app based on iOS native gesture support. Users tried each gesture on both apps one-by-one and gave their feedback using a questionnaire form. Results show the same level of efficiency and user satisfaction in many cases, as well as better in few cases and lower in some other cases. Overall, results indicate that mobile web apps through MTGest kinds of libraries can be an alternative solution in the future.

Some other frameworks for the support of multi-touch gestures are: jQMultiTouch [7] web tool-kit, inspired by JQuery, for creating multi-touch interfaces; Gesture Coder [6] for generating code to recognize multi-touch gestures. However, we chose JQuery[3] and hammer.js[4] as foundations due to their powerful abstraction from low-level implementation details and their cross-browser compatibility. One way of working with the MTGest library has already been described in detail in [4], whereas this paper focuses on the study comparing MTGest with native gestures.

The remainder of the paper is structured as follows: In Section 2, we introduce our MTGest library. In Section 3, we provide details of the conducted user evaluation study. In Section 4, we present and discuss the results. Finally, we conclude in Section 5.

[1] W3C - HTML5. `http://www.w3.org/TR/html5/`

[2] World Wide Web Consortium. `http://www.w3.org/`

[3] The jQuery Foundation - `http://jquery.com/`

[4] Eight Media - `http://eightmedia.github.com/hammer.js/`

2 The MTGest Library

HTML5, the new specification of HTML by W3C, is still in working-in-progress process. HTML5 provides a set of interfaces[5] for touch events. However, it lacks built-in tags for the functionality of multi-touch gestures.

Our **MTGest** (**M**ulti-**T**ouch **Gest**ures Library) library, based on JavaScript and JQuery, enables the support of multi-touch gesture interaction in HTML5-based documents. It is built on top of the hammer.js library, which is also based on JavaScript, for controlling gestures on touch devices. It supports most of the single and multi-touch gestures in the current mobile domain. Moreover, it is possible to define own gestures in which the developer specifies the criteria for such a gesture, e.g., tapping three fingers together.

The standard gestures supported by our library are: *tap*, *double tap*, *hold*, *drag*, *swipe*, *transform* (*pinch*), *rotation*, *flick*, *zoom and rotation* together, and *shake*. Additional customized gestures (e.g., *three-fingers tapping* or *multi-fingers swiping*) are also provided for using in some specific interaction context.

The MTGest library works as follows. The provided functions, corresponding to each gesture, by the library are attached to a container representing a specific area in the HTML document. The *hammer.js* is also attached to the same container to get the touch events happen to this container. It is possible to attach more than one gesture to the same container. Then the specific area in the HTML document, representing the container, gives the interaction according to the attached gestures. Figure 1 provides the overall architecture.

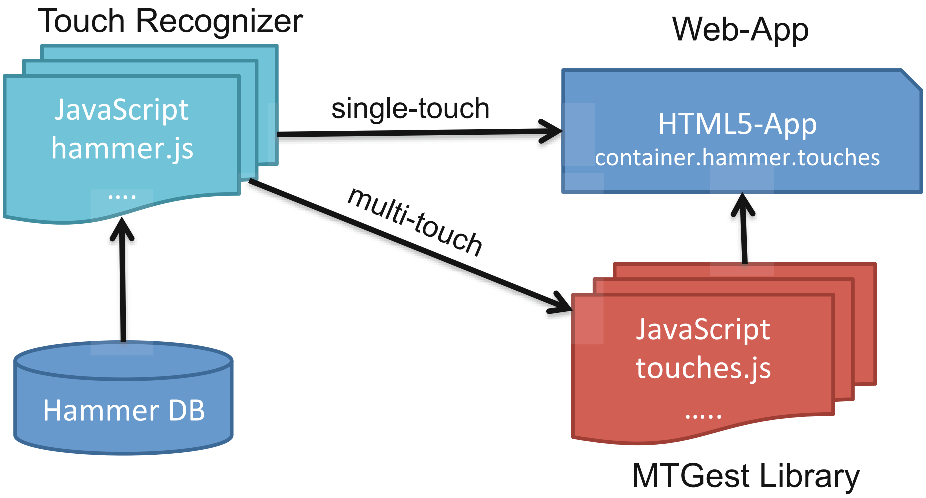


Fig. 1. The overall architecture of the working of MTGest library

[5] https://dvcs.w3.org/hg/webevents/raw-file/tip/touchevents.html

3 The User Evaluation Study

We performed a user evaluation study in a controlled environment, where the focus was on comparing the multi-touch gesture interaction support provided through our MTGest library and a native mobile platform. This was done through developing two simple mobile apps, one was a mobile web app that provides the desired multi-touch gestures through our MTGest library while the other one was a native iOS-based mobile app that provided these gestures' support through the native iOS library.

In the following, we provide details of the both developed apps (i.e., the mobile web app and the iOS-based mobile app), the study goal and hypothesis, and the experiment settings.

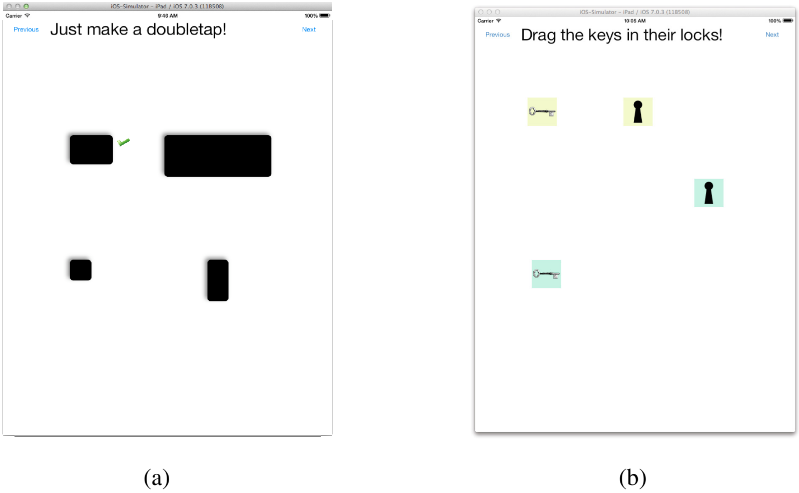
3.1 The Testing Apps

For testing the multi-touch gesture interaction support through our MTGest library and a native mobile platform library, we developed two simple apps. The first one was a mobile web app that used our MTGest library for providing the multi-touch gesture interaction support, while the second one was an Apple iOS-based native mobile app that provided the multi-touch gesture interaction support using the iOS native gestures support. Both apps provided the same level of functionality and there was no difference in the interface style or layout. This was done in order to avoid any biasedness in the user evaluation study.

Eight touch- and multi-touch gestures, mostly the standard ones provided by most of the current platforms, were implemented in both apps. These gestures include: *tap, double tap, hold, drag, swipe, flick, zoom* (both *zoom-in* and *zoom-out*), and *rotation*. In both apps, each gesture was covered up on one page where each page contained several (up to four) containers having the implementation of the underlying gesture. These containers were different in size and orientation with the same gesture support in order to provide a variety of user interaction with the underlying gesture. When a user interacts correctly with the container through the specified gesture, a feedback is shown to the user for this correct interaction; otherwise nothing is shown. The user can go to the next page for interacting with the next gesture any time or after finishing the interaction with all containers on the current page.

In the cases of *tap, double tap,* and *hold* gestures, we implemented four containers on each page having the underlying gesture support. Figure 2 (a) shows the screenshot of the native iOS-based app where the four containers have the *double tap* gesture interaction. The green correct mark indicates that the user has successfully interacted with this container with the *double tap* gesture. In the case of *drag* gesture, there were two sets of containers. One set was showing a key while the other one was showing the lock, as shown in Figure 2 (b). The container set showing the key shape were linked with the *drag* gesture. When a use drags this key container to a lock container, the app indicates a successful execution of the gesture.

In the case of *swipe* gesture, both apps provided four containers to provide the interaction support in four directions (i.e., left-to-right, right-to-left, up-to-down, and down-to-up), as shown in Figure 3 (a). The user needs to swipe the finger from the tail to the head of the arrow in order to execute the *swipe* gesture interaction correctly. In the case of *flick* gesture (here the *flick* gesture represents same as its representation in iOS), both apps provided one container that had the interaction of *flick* gesture in four directions same as with *swipe* gesture case, as shown in Figure 3 (b). Finally, the *zoom* and *rotate* gestures were implemented through two containers in both apps. In the case of *zoom* gesture, both containers were different in size and it was up to users to play with them for checking zoom-in and zoom-out interaction. While in the case of *rotate* gesture, two figures were given in the containers up-side-down orientation so that users can rotate them in normal orientation.



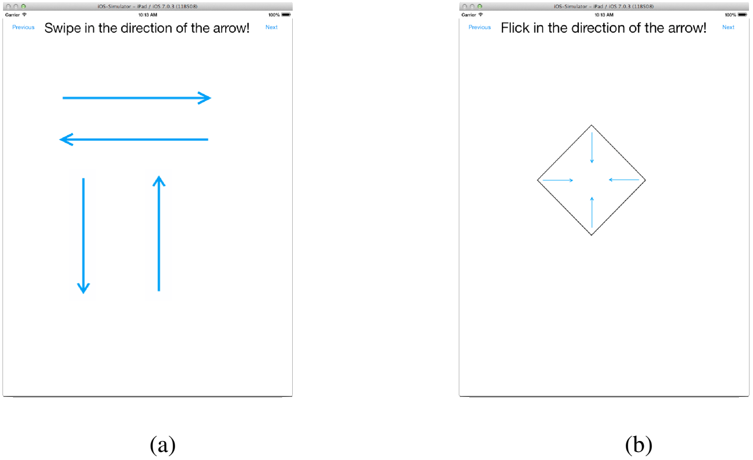
Just make a doubletap!

Drag the keys in their locks!

(a)

(b)

Fig. 2. (a) A screen-shot of the page with the *double tap* gesture support, *(b)* A screen-shot of the page with the *drag* gesture support



Swipe in the direction of the arrow!

Flick in the direction of the arrow!

(a)

(b)

Fig. 3. (a) Four directed arrows show the *swipe* gesture interaction in the same direction, *(b)* the arrows inside the container represent the *flick* gesture interaction in the same direction

3.2 Study Goal and Hypothesis

The goal of this user study was to analyze whether our developed MTGest library can provide the touch- and multi-touch gestures interaction support in mobile web apps compared to native mobile apps (here we target only the Apple iOS platform) from the perspective of efficiency of such interaction support and user satisfaction level with the underlying interaction. We compare the results from the following criteria:

- *Efficiency:* We check whether the underlying gesture worked accurately and the interaction-response time was appropriate. In this regard, we collect subjects' feedback for both apps and compare them.
- *User Satisfaction:* We collect subjects' feedback for both apps and compare them.

Our hypothesis is that in term of efficiency and user satisfaction, our proposed MTGest library provides approximately the same interaction support for the underlying gestures compared to the native platform (i.e., the iOS platform) support.

3.3 The Experiment Settings

We performed the evaluation study with 12 subjects (3 females and 9 males). We categorized them according to their experience with smart-devices and mobile platforms. Four subjects were experienced users of Apple iOS platform, three subjects were experienced of Android platform, while the remaining 6 subjects were without much expertise in any specific mobile platform. The age of subjects were between 20 and 36 years old with a mean of 27.5.

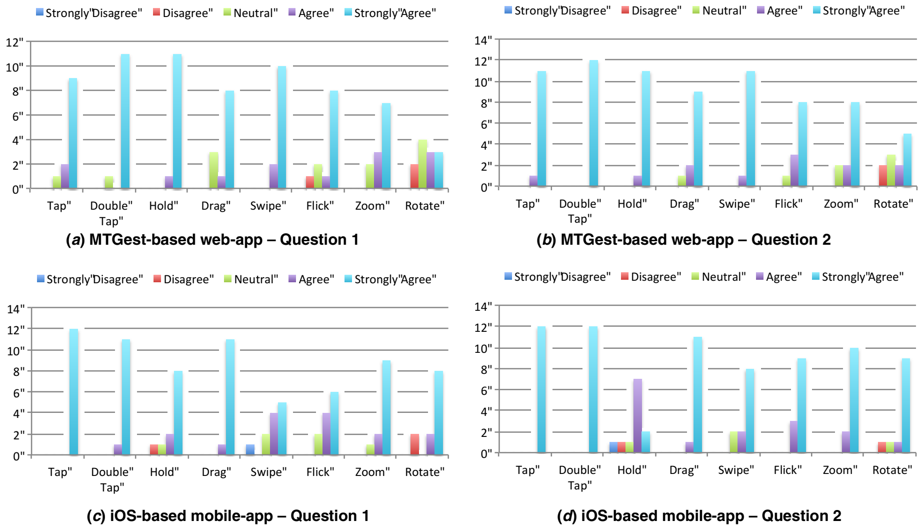
The test devices for both developed apps (i.e., the mobile web app based on our MTGest library and the iOS-based native mobile app) were Apple iPad 2 with the same specifications. We installed the web app on one device while the native app on the other device. Before start of the experiment, a brief tutorial was given to each subject about the goal of the experiment. For each tested gesture, subjects were asked first to try all the containers having the underlying gesture support on both devices. Then they were asked to fill a closed-ended questionnaire form with a likert scale from 1 to 5, where 1 meant *strongly disagree* and 5 meant *strongly agree*. There were total four questions in this mode, same for both apps separately. The aim of first two questions was to get the subjects' feedback for checking the efficiency of the underlying library (i.e., the MTGest library or the native iOS library) in providing the support of the tested gesture interaction. The aim of the later two questions was to get the subjects' feedback for checking their satisfaction level with the tested gesture for both apps. These four questions were:

1. *The gesture works accurately.*
2. *The interaction-response time of the gesture was appropriate.*
3. *Overall, I am satisfied with this gesture facility through the underlying app (i.e., the web app or the iOS-based app).*
4. *In future, I would like to use this gesture through the underlying app (i.e., the web app or the iOS-based app.*

At the end of closed-end questions for each gesture, subjects were also asked that which mode (i.e., the web app or the native app) for this gesture is preferable by them for the future usage, with the option of selecting one or both apps. In order to avoid any biasedness towards the second testing app due to the learning effects, half of the subjects were asked to test the web app first and then the iOS-based native app, while the other half were asked to test the iOS-based app first and then the web app.

4 Results and Discussions

In this section, we provide the results of our conducted user evaluation study and discuss them to check whether they reflect our initial hypothesis. After testing each gesture on both apps, subjects were asked to answer the set of questions regarding the tested gesture.



(*a*) MTGest-based web-app – Question 1

(*b*) MTGest-based web-app – Question 2

(*c*) iOS-based mobile-app – Question 1

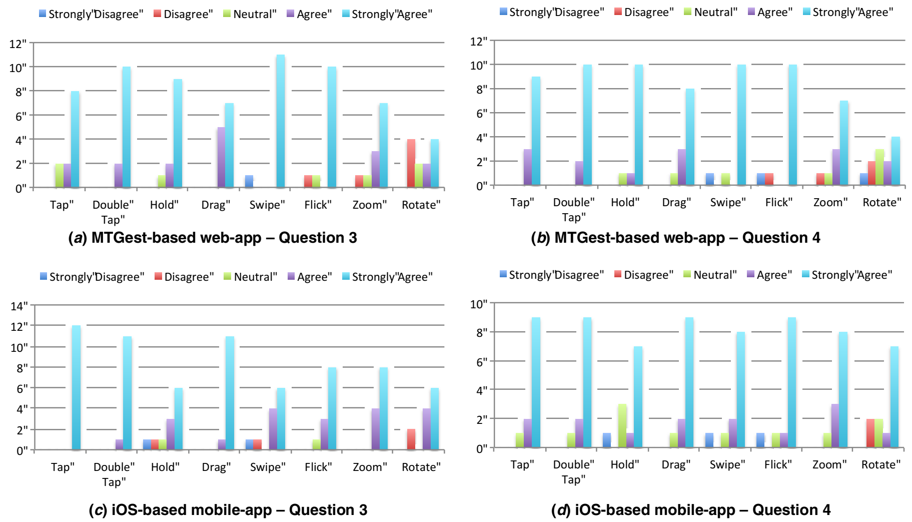
(*d*) iOS-based mobile-app – Question 2

Fig. 4. The subjects' feedback for questions 1 and 2, collected through the likert-scale

Figure 4 provides the subjects' feedback with regard to the first two questions for each of the tested gesture on both apps. For the *tap* gesture, all subjects strongly agreed for the accurately work of iOS-based app, while 9 subjects strongly agreed and 2 subjects just agreed for the web app. Regarding the second question, all subjects strongly agreed for both apps except one that rated agreed for the web app. Results for the *double* tap gestures are also nearly the same in both cases for both apps. This indicates that our MTGest library provides the same level of efficiency for these two gestures. The case of *hold* gesture is interested, as the subjects' feedback for the web app is far better than the iOS-based app. We observed that iOS gives a too quick interaction response, which the subjects might not expected from the *hold* gesture as they were expecting a little wait for keeping hold the touch. That might be the reason for this lower ranking by subjects. We also observed that subjects from the Android platform or non-experienced background were more reluctant in liking the iOS-based app response, while they felt happy with the web app response. In the case of *drag* gesture, subjects' feedback was a bit better for the iOS-based app compared to the web app. However, the difference was not very noticeable.

In the case of *swipe* gesture, subjects' feedback about the web app was much higher than the iOS-based app. We observed that again this is because the too quick response in iOS, as even if the subject swiped just little more than half of the swipe area length it started working. In the case of web app, it worked only when subjects swiped the whole length of the swipe area. Due to this, subjects felt more confident in web app compared to the iOS-based app. This is also indicated in the case of *flick* gesture, where subjects' feedback about the web app was slightly better than the iOS-based app. In the case of *zoom* gesture, subjects rated iOS-based app better than the web app. However, again the difference is not much significant. Finally, in the case of *rotate* gesture, subjects rated quite higher the iOS-based app compared to the web app. We observed that this is because of the image drawing performance issue in the web app, as the image is drawn on the page each time the user moves the fingers for the rotation.

Overall, the subjects' feedback of the first two questions indicates that in most cases the web app efficiency is either approximately same or higher than the iOS-based app. While in some complex gestures such as zooming and rotation, it is behind the iOS-based app. However, this can be improved in future. In summary, we can say that the overall feedback of these two questions confirm our hypothesis regarding the efficiency of our developed MTGest library.



(*a*) MTGest-based web-app – Question 3

(*b*) MTGest-based web-app – Question 4

(*c*) iOS-based mobile-app – Question 3

(*d*) iOS-based mobile-app – Question 4

Fig. 5. The subjects' feedback for questions 3 and 4, collected through the likert-scale

Figure 5 provides the subjects' feedback with regard to the later two questions for each of the tested gesture on both apps. The aim of these two questions was to check the user satisfaction level with the tested gesture interaction. For the *tap* gesture, the subjects' feedback with the iOS-based app was a bit higher than the web app. Moreover, 11 subjects preferred for using this gesture through iOS-based app and 9 subjects also went for web app too. It is noted that in the case of future usage of the tested gesture, subjects were free to choose one or both. In the case of *double tap* gestures, subjects' feedback was approximately the same. Also, 11 subjects choose the iOS-based app while 10 mentioned the web app for the future usage of this gesture. We observed that subjects' feedback improved towards positive with the web app after getting experience. In the case of *hold* gesture, the subjects' feedback reflects the feedback of question 1 and 2, as their satisfaction trend for the web app was quite higher than the iOS-based app. They significantly also preferred the web app for the future usage of this gesture (12 compared to 4). In the case of *drag* gesture, the subjects' feedback was a bit higher for the iOS-based app compared to the web app. However, 10 subjects choose web app while 9 subjects choose iOS-based app for the future usage of this gesture.

In the case of *swipe* gesture, subjects' feedback about the web app was much higher than the iOS-based app. Moreover, 10 subjects preferred web app and 5 preferred iOS for the future usage of this gesture. We observed that subjects' feedback for the web app was increased because the web app provided better the expected interaction (i.e., working when user swipes through the whole area rather than just a part of it) in this gesture implementation. The same also went for the *flick* gesture, where subjects' feedback again was more towards the web app. However in this case, the subjects' preference for the future usage of this gesture was nearly the same for both the web app and the iOS-based app, i.e., 8 and 7. In the case of *zoom* gesture, subjects were a bit higher satisfied with the iOS-based app than the web app. However, the difference is unnoticeable. Finally, in the case of *rotate* gesture, subjects significantly rated higher the iOS-based app compared to the web app. Also, only 2 subjects preferred the usage of this gesture in web app compared to 9 for the iOS-based app.

Overall, the subjects' feedback of the later two questions reflects their experience with the tested gesture on both apps and approximately the same as of the previous two questions. Except in the cases of *drag* or *rotate* gestures, the subjects' feedback about the later two closed-ended questions for the web app was either approximately the same as for iOS-based app or higher than it. However, the subjects' feedback regarding their satisfaction level has many limitations. There are many factors (e.g., users' expectations, curiosity, their interests in new experiences, their expertise with gestures, their positive attitude towards Apple, their low expertise with MTGest library, etc.) that can affect users' satisfaction level. In spite of this, results of the study provide an indication that the web apps have the potential of providing an alternative to the native mobile apps if they get support by multi-touch gestures libraries. MTGest library is one of the candidate libraries; however, it needs to be improved for providing better performance in the cases of some complex gestures (e.g., *zoom* and *rotation*). Moreover, as the study targeted only the iOS platform and the tested scenarios were quite simple; hence, for generalizing the results there is a need to perform detailed user evaluation studies with different platforms and with more complex interaction scenarios.

5 Conclusion

In this paper, we showed that the MTGest library is feasible to be used for multi-touch gestures with regard to efficiency and user satisfaction. We performed an initial evaluation study with 12 subjects comparing the MTGest library with the native iOS gestures. A main issue of the study was to compare single multi-touch interaction in an isolated scenario.

The results showed, that MTGest works accurately for *tap*, *double tap*, *hold*, *drag*, *swipe*, *flick* without showing a significant difference to the native iOS gestures. Indeed, *swipe* and *flick* gestures were rated significantly better. The rotation gesture worked significantly more accurate on the native iOS implementation. In general, the study indicates that in most cases the web app efficiency is either approximately same or higher than the iOS-based app.

With regard to user satisfaction, the results of the study indicated not a clear result. The subjects' satisfaction level reflects their experience with the tested gesture on both apps. In general, there is no significant difference between web based gestures and native gestures. This leads us to the conclusion, that using MTGest is reasonable although it needs to be improved for providing better performance in the cases of some of the more complex gestures.

Future work will deal with the performance of a detailed user evaluation considering the usage of gestures within a concrete scenario and also do a comparison with other mobile operating systems, such as Google Android or Microsoft Windows Phone.

References

1. Apple iOS, `http://www.apple.com/uk/ios/`
2. Google Android, `http://www.android.com/`
3. Humayoun, S.R., Ehrhart, S., Ebert, A.: Developing Mobile Apps Using Cross-Platform Frameworks: A Case Study. In: Kurosu, M. (ed.) HCII/HCI 2013, Part I. LNCS, vol. 8004, pp. 371–380. Springer, Heidelberg (2013)
4. Humayoun, S.R., Hess, S., Kiefer, F., Ebert, A.: i2ME: A framework for building interactive mockups. In: MobileHCI 2013, pp. 606–611. ACM, New York (2013)
5. Microsoft Windows Phone, `http://www.windowsphone.com/`
6. Lü, H., Li, Y.: Gesture coder: A tool for programming multi-touch gestures by demonstration. In: CHI 2012, pp. 2875–2884. ACM, New York (2012)
7. Nebeling, M., Norrie, M.: 2012: jQMultiTouch: Lightweight toolkit and development framework for multi-touch/multi-device web interfaces. In: EICS 2012, pp. 61–70. ACM, New York (2012)

Harmonic Navigator: An Innovative, Gesture-Driven User Interface for Exploring Harmonic Spaces in Musical Corpora

David Johnson[1], Bill Manaris[1], and Yiorgos Vassilandonakis[2]

[1] Computer Science Department, College of Charleston,
66 George Street, Charleston, SC 29424, USA
manarisb@cofc.edu, dsjohnso1@g.cofc.edu
[2] Music Department, College of Charleston,
66 George Street, Charleston, SC 29424, USA
vassilandonakisy@cofc.edu

Abstract. We present Harmonic Navigator (HN), a system for navigating and exploring harmonic spaces extracted from large musical corpora, to be used in music composition and performance. A harmonic space is a set of harmonies (chords) and transitions between harmonies found in a music corpus. By navigating this space, the user can derive new harmonic progressions, which have correct voice leading. HN is controllable via a Kinect gesture interface. To aid the user, the system also incorporates stochastic and evolutionary techniques. HN offers for two primary modes of interaction: a harmonic transition selector, called *harmonic palette*, which utilizes a GUI to navigate harmonic transitions in a front-to-back manner; and a harmonic-flow scrubber, which presents a global overview of a harmonic flow and allows the user to perform common audio scrubbing and editing tasks. Both GUIs use colors to indicate harmonic density based on Legname's density degree theory.

Keywords: harmonic navigation, computer music, graphical user interface, gesture language, Kinect sensor, harmonic space, music composition, music performance.

1 Introduction

The use of computation in music composition and performance has emerged from advancements in music technology, such as MIDI interface and synthesized instruments, explorations in the use of mathematic and aleatoric principles in composition by composers like Iannis Xenakis, Gyorgy Ligeti and John Cage [1,2], and the application of artificial intelligence tools to music analysis and generation.

Several systems have emerged in recent decades to assist with music performance and composition, including Cope's *EMI* [3], Biles' *GenJam* [4], and Pachet's *Continuator* [5]. These are discussed in more detail in the next section. We present a novel system that provides an innovative, gesture-driven user interface for navigating

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 58–68, 2014.

harmonic spaces of music from large corpora. This system combines stochastic and evolutionary techniques and is an extension of *Monterey Mirror*, an interactive system for melodic exploration [6].

Harmonic Navigator allows for user-guided generation of new harmonic (chord) material from an existing musical corpus (currently we explore the Riemenschneider collection of *371 J.S. Bach Chorales*). This corpus is used to train a Markov model, a stochastic model that represents the transition probabilities of chords in the corpus. The Markov model is capable of rapidly generating material that is similar to the provided corpus. In practice, the generated material often contains only short-term similarities (event-to-event) and lacks long-term coherent structure. We utilize a genetic algorithm to search the Markov model for high-quality material. Using power-law metrics as a fitness measurement allows the genetic algorithm to select material that is similar to target material, such as a user-provided melody or harmonic flow [7-8].

Finally, the system allows saving of a generated chord progression, for further processing and use in music composition projects.

This paper focuses on the user interface aspects of the Harmonic Navigator. The remaining sections are organized as follows: section 2 presents related background research; section 3 describes the target audience and presents a high-level task analysis for the system; section 4 describes the user interface in more detail; section 5 provides an overview of the system architecture and major software components; finally, section 6 discusses future work.

2 Background

Within the last 50 years there have been numerous applications of computing and artificial intelligence in analysis, generation, composition, and performance of music. While these results are sometimes judged by how well they model human intelligence (strong AI), the real contribution lies in how they may enhance human creativity and facilitate artistic exploration and expression.

GenJam generates jazz improvisations for real time performance [4]. *GenJam* is trained using an interactive genetic algorithm, which determines fitness through a human mentor. The trained population is used to “trade fours” with a human performer.

The *Corpus-Based Harmonic Progressions Generator* [10] mixes stochastic selection via Markov models and user input to generate harmonic progressions in real time. The user enters information to specify harmonic complexity and tension, as well as a bass-line contour. This is used by the system to influence the selection of harmonies from the trained Markov models, and to generate a harmonic progression.

Experiments in Music Intelligence (EMI) is the most comprehensive work in automated computer music generation to-date [3]. EMI analyzes a corpus of musical works and trains Markov models. EMI can then automatically compose pieces in a style similar to the corpus. EMI works offline and has been used to generate numerous pieces in the style of various composers.

Continuator is an interactive music performance system which accepts musical input from a human performer. It completes musical material in the same style as the user input [5]. Using a musical corpus, the system trains several Markov models. Human performer input is matched against the various Markov models until a match is found. The corresponding Markov model is used to generate a musical continuation. This makes the system sometimes generate perfect reproductions of earlier musical input, and other times less accurate repetitions (introducing interesting variations).

NEvMuse [11] is an experiment in using genetic programming to evolve music pieces based on examples of desirable pieces. *NEvMuse* uses power-law metrics as fitness functions. In an evaluation experiment, these metrics were able to predict the popularity of 2000 musical pieces with 90.7% accuracy. The system was used to autonomously "compose" variations of J.S. Bach's Invention #13 in A minor (BWV 784). Similarly to *NevMuse*, the Navigator's genetic algorithm uses power-law metrics to determine fitness.

Monterey Mirror [6] uses Markov models, genetic algorithms and power-law metrics to generate musical phrases in real-time, based on musical input from a human performer. Markov models are used to capture short-term correlations in melodic material. The genetic algorithm is then used to explore the space of probable note combinations, as captured by the Markov model, in search of novel, yet similar melodic material. Similarity is measured using power-law metrics, which capture long-term correlations in melodic material, i.e., the statistical balance between expectation and surprise across various musical parameters [8].

Harmonic Navigator is implemented in Jython and Java using custom GUI, MIDI and OSC libraries. It incorporates computational elements from *NevMuse* and *Monterey Mirror* to allow human performers to navigate the space of musical harmonies using a gesture-based interface [12].

In this paper, we present a new user interface for the Harmonic Navigator that allows composers and performers to create new music by modifying musical output of a generative system in real-time.

3 Target Audience

The Harmonic Navigator (HN) is a gesture-based interactive system for exploring harmonic spaces of distinct (or composite) musical styles (see Fig. 1). Also, it may be used to generate music in real-time, in collaboration with human performers. The UI has been designed for users with, at least, basic training in functional tonality (first-year college harmony, or equivalent). However, as we collect usability feedback from more users, this UI may evolve (e.g., to provide more or, even, less information).

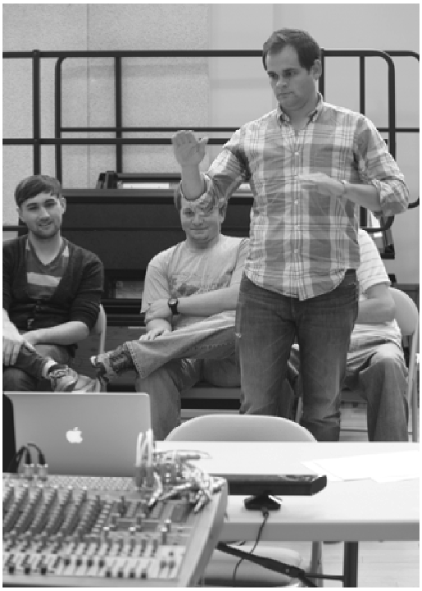


Fig. 1. One of the authors interacting with the system

3.1 Music Composers

HN can be incorporated by music composers, in a computer-aided composition context. In particular, a composer may use it to explore compositional ideas in harmonic spaces derived from various musical corpora. These can consist of pre-existing libraries of established musical (and therefore harmonic) styles, or could be a collection of the composer's previous own body of work. By employing these corpora, traditional harmonies may be derived and be evaluated on a consonance/dissonance scale [14]. More dense harmonies may also be explored, and may be similarly evaluated on a consonance/dissonance scale [9]. In order for this to work well, the music corpora loaded to the system must contain enough musical pieces (for harmonic variety) and should be stylistically consistent (e.g., consist only of Baroque pieces, or Impressionist pieces). By combining two stylistically inconsistent groups of pieces, this would create a rather disjoint harmonic space, consisting of two mostly isolated "islands" (although it would be possible to "travel" from one to the other, via, some common basic harmonies, which may appear in pieces from both styles, but function in different ways in each).

3.2 Music Educators

HN may also be used to enhance traditional classroom pedagogy in tonal harmony. Professors may engage students through tonal harmony games implemented on an HN platform. “Players” could interactively assign appropriate tonal function and hierarchy to each important pitch in a melody: tonic, predominant or dominant [13], and then select from a variety of available chords in various inversions. In the end, users may gain a much deeper appreciation for harmonic functions quicker and retain it for a much longer time.

It could also be used to explore pitch-set generated harmony or 12-tone and serial harmonic styles in a more advanced 20th century harmony course. In this context, HN will provide even more insight to the student, as it would be able to offer suggestions that take into account pitch usage and cycling, in addition to the harmonic spacing and density.

3.3 Music Performers

Finally, HN may be used in musical performances. Musicians and non-musicians (e.g., members of the audience, or passers-by), may utilize MIDI and OSC controllers (e.g., iPhone TouchOSC client), as well as traditional instruments, to create harmonic contexts for improvised performances. Another related possibility is to compose/design musical games (e.g., the system could be driven through audience participation) to engage, inspire, and possibly challenge musicians in various performance environments, or to allow non-musicians to create musical performances in unconventional settings (such as art galleries with HN sound installations).

4 User Interface

The Harmonic Navigator offers two primary modes of interaction: a gesture-based harmonic transition selector, called the *harmonic palette*, and a harmonic-flow scrubber, which presents a global view of a flow being generated. The first UI provides a tree-level view, and thus allows for localized control and inter-harmony navigation. The second UI provides a forest-level view, and supports scrubbing and editing actions. Herein we focus mainly on scrubbing actions (such as playback in arbitrary speed). Both views use colors to indicate harmonic density calculated using Legname's density degree theory [9].

4.1 The Harmonic-Palette View

The Harmonic-Palette View presents available harmonies as a dynamic navigable space. It utilizes a 3D front-to-back approach. The interface presents users with a *harmonic palette*, from which to choose a follow-up harmony (see Fig. 2). The palette contains a number of circles, each representing a harmony. The current harmony is located in the center of the display. Follow-up harmonies are determined by the current harmony (as dictated by the training corpus), and are placed in a clockwise fashion, around a clock face labeled with the 12 tones. Pieces are normalized to the tonic, so pitch C is always positioned at 0. We use vertically stacked numbers to denote harmonic intervals. This is consistent with the vertical placement of notes on a staff. We have considered using Western musical notation, however, this representation provides more direct information, i.e., users can see the intervals right away - they do not have to derive them from the musical notation.

Moreover, the size (radii) of follow-up circle-harmonies corresponds to transition probabilities from the current harmony (the larger, the more probable).

In the case of multiple follow-up harmonies having the same root pitch (e.g., see E and A root pitches, in Fig. 2, they are arranged around a smaller clock face. The size (radius) of this clock face corresponds to the sum of the enclosed harmonies' probabilities. Hovering the cursor over this clock face zooms in to display a larger version of the clock face, which presents more information about the contained harmonies, and allows the user to select one. These harmonies are arranged inside the smaller clock face based on the second pitch in the harmony.

When dealing with multiple harmonies that have the same second pitch, these harmonies will also be placed inside an even smaller clock. This hierarchical grouping continues until all harmonies can be represented individually.

The HN engine is capable of making recommendations for what it considers possibly aesthetic choices for follow-up harmonies. This is accomplished via a genetic algorithm which runs continuously (in the background) to suggest interesting harmonic flows. The follow-up harmony (or harmonies) selected by the genetic algorithm is (are) identified by a special bright ring around a suggested harmony. Since the genetic algorithm is running continuously, it is possible for the suggested harmony to change (by the genetic algorithm discovering a better choice) as the user is contemplating.

Circle-harmonies are assigned color based on intervallic tension. Since intervallic tension is already visible on the interface, through the displayed harmonic intervals, the assigned color representation is redundant. This emphasizes the existing information, and makes it more visible to non-experts.

Intervallic tension of a chord is determined by two factors. One is the intervallic content of the chord - a chord with more tense intervals has a higher tension factor, and thus sounds more dissonant. The relaxation vs. tension of the chord is mapped to cool vs. warm colors on a color wheel, i.e., blues are cool (relaxed) and reds or yellows are warm (tense).

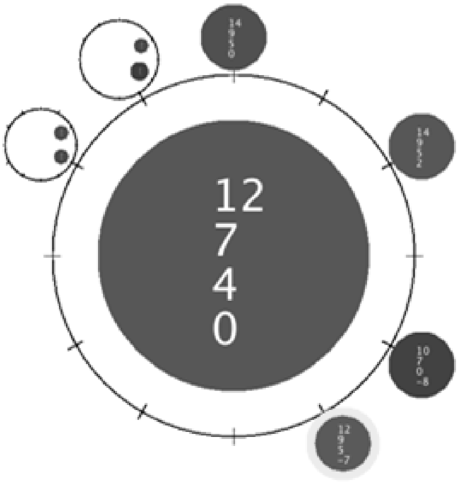


Fig. 2. The *harmonic palette* interface is used to select a follow-up harmony. The current harmony is in the center. *Follow-up harmonies* are arranged in a clockwise fashion, around a *clock face* corresponding to the 12 tones. *Numbers* represent harmonic intervals in a chord. *Color* (here reproduced in grayscale) denotes chord harmonic density.

4.2 Gesture Language for the Harmonic Palette

The Harmonic Palette UI has been designed to support three main user tasks for harmonic navigation. These are:

1. "Explore the current harmony palette";
2. "Select a follow-up harmony"; and
3. "Backtrack" (i.e., unselect current harmony and return to the previous palette).

Our current prototype is implemented using the Kuatro system. The Kuatro system is a new architecture for supporting a multitude of sensors and wireless controllers for audio/visual interactive installations. The main objective behind its design is to hide the complexities of communicating with such devices, and allow the UI developer to focus on the higher-level aspects of designing an effective UI for audio/visual control of a computer music application. The Kuatro architecture will be reported elsewhere.

We have designed a Kinect-based gesture language to implement the above user tasks. (We are also exploring gesture languages for other controllers, such as the Leap Motion sensor and OSC control via smartphones.) The Kinect gesture language utilizes only one hand via three gestures (freeing the second hand for other activities, such as interacting with MIDI and OSC controllers):

- **Hand Movement in the X-Y Plane** – Moving the hand left-to-right and up-to-down moves the cursor around the display. This action supports exploration of the current harmony palette (e.g., hovering over a secondary clock face to enlarge it).
- **Hand Push** – Pushing towards a follow-up harmony selects it. This moves the selected circle-harmony to the center, begins sounding the corresponding harmony (via MIDI), and displays the next harmony palette. This action supports moving forward in the harmonic space.
- **Hand Wave** – Waving over the current circle-harmony (center of the display), stops sounding it, and returns to the previous harmonic palette (to possibly try something else). This action supports moving backward in the harmonic space.

In particular, moving backwards allows the user to step back to previous harmony selection points, and try other alternatives. While this may seem peculiar during live performance, it may be utilized creatively (not unlike sound looping, and/or "scratching" by DJs). On the other hand, this is quite natural for composition tasks (i.e., "should I use this harmony or that?" or "what harmonic choices would I have here, had I gone to a relative minor three chords ago?").

4.3 Harmonic-Flow View

The Harmonic-Flow View presents a global, forest-level view of a harmonic flow generated by the user through the harmonic palette UI (or automatically by the harmonic generator engine). Through this view, the user may explore and update the different harmonies comprising the harmonic flow as they desire. As seen in Fig. 3, harmonies are placed horizontally across the display. For each harmony being selected, alternate harmonies, as determined by the Markov model, are displayed vertically.

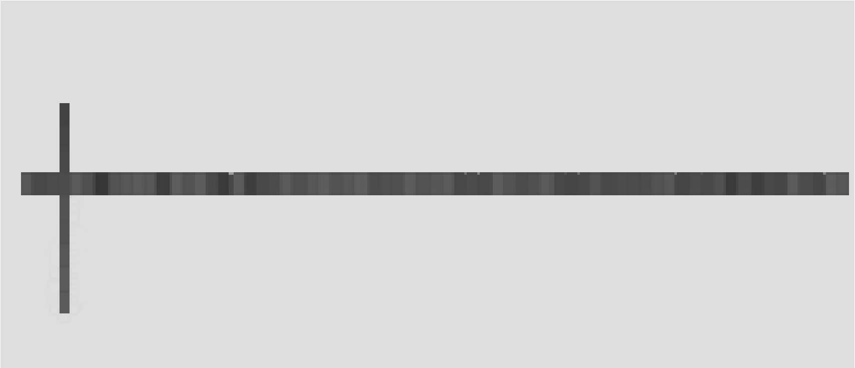


Fig. 3. The *harmonic-flow scrubber* interface is used to view a complete harmonic flow, as constructed through the lower-level interface (see Fig. 2). The harmonic flow appears on the horizontal. Individual harmonies are displayed as *rectangles*. Hovering over a rectangle presents alternative harmonies (on the vertical). *Color*, again, denotes chord harmonic density.

4.4 Gesture Language for the Harmonic Flow

The Harmonic Flow UI supports three main user tasks. These are:

1. "Forward and backward scrubbing";
2. "Explore alternative harmonies"; and
3. "Select a new harmony".

As mentioned earlier, through the Kuatro architecture, users may utilize various gesture and motion controllers to interact with the UI. Herein, we present a Kinect-based language for users to control the system via their location in a room (many other possibilities exist for other controllers and sensors). By viewing the room from above, we use an X-Y coordinate system to track a user through the room and map their location to specific tasks.

- **Movement along the x-axis** – The x-axis runs parallel to the display and controls the scrubbing capabilities. By moving parallel to the display, the user identifies which harmonies are played across the flow. The tempo of scrubbing is controlled by how fast or slow the user is moving in this direction.
- **Movement along the y-axis** – The y-axis is perpendicular to the display. By moving along the y-axis, the user plays the harmonies presented in the vertical list of harmonic alternatives. As the user moves closer to the display, they play harmonies upward in the transition list; these are harmonies with increasing tension. As the user moves away from the display, they play harmonies downward in the transition list; these are harmonies with decreasing tension.

A user selects an alternative harmony by beginning to move again across the x-axis. Also, selecting an alternative harmony triggers HN to regenerate the flow based on their new selection, if opted by the user, via the genetic algorithm. The genetic algorithm and the corresponding automated generation of harmonic flows is presented extensively in [15].

5 System Architecture

The Harmonic Navigator system uses a Model-View-Controller architecture based on the Kuatro system. This reduces complexity from the UI design and implementation, while allowing for a multitude of controllers, such as a mouse, a Kinect, and smartphones using OSC clients (e.g., TouchOSC). In Fig. 4, the View in this architecture is the UI, the controller is the Gesture Engine, and the Model is the Harmonic Generator. To support a wide range of controllers we have implemented a protocol for the Gesture Engine to communicate via OSC. (This will be presented in a future publication.)

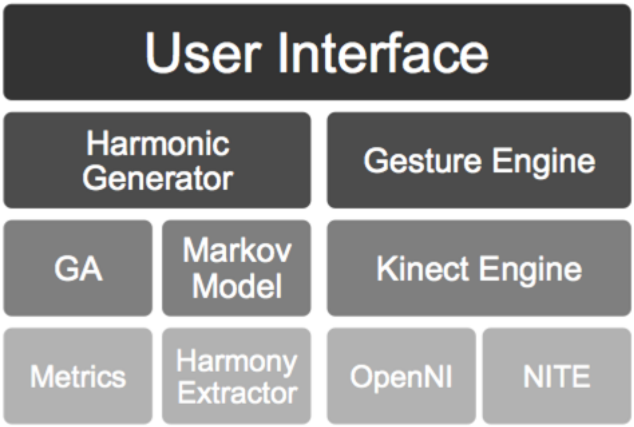


Fig. 4. Harmonic Navigator architecture using a Microsoft Kinect

6 Discussion and Future Work

The Harmonic Navigator is a powerful tool for exploring harmonic spaces in a direct, physical, and accessible manner. As new gesture control devices are introduced, its power will only increase. The possibility of allowing non-expert musically users to experience harmonic flows in such an intimate manner presents various possibilities for further work. We are currently exploring an application that will introduce the novice theory student to the notion of tonal function in common practice music. The system attaches a T, PD or D label to each suggested chord in the *harmonic flow scrubber*, and the user can quickly develop their listening ability to recognize tonal function and navigate harmonically through a musical phrase using harmonic implications alone. They can then harmonize a given melody or bass line in a more musically intelligent way by selecting chords with the appropriate function label among the ones suggested by HN, thus gaining a deeper understanding of tonal harmony. This deeper understanding would normally take several years of study, as well as keyboard proficiency. This type of learning could be further enhanced by creating a physical space larger than the user, so that he can navigate through it by walking around the space, "scrubbing" through the functional harmonic space, as possible via the Kuatro architecture discussed above.

We have presented Harmonic Navigator, a system for navigating and exploring harmonic spaces extracted from large musical corpora, to be used in music composition and performance. This system is currently being evaluated with actual users, in order to improve its usability and possibly improve its UI.

In closing, video demos of the system are available here:

- A demo of the harmonic palette UI being controlled via a Kinect - **http://goo.gl/ni7ZVl**.
- A demo of the harmonic flow view - **http://goo.gl/hpXk2G**.

References

1. Xenakis, I.: Formalized Music: Thought and Mathematics in Music. Pendragon Press, Hillsdale (1992)
2. Cage, J.: Silence: Lectures and Writings of John Cage. Wesleyan University Press, Middletown (1961)
3. Cope, D.: Virtual Music: Computer Synthesis of Musical Style. MIT Press, Cambridge (2004)
4. Biles, J.A.: Performing with Technology: Lessons Learned from the GenJam Project. In: Musical Metacreation Worksop, 9th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2013), pp. 14–19. AAAI Press, Palo Alto (2013)
5. Pachet, F.: Playing with Virtual Musicians: The Continuator in Practice. IEEE Multimedia 9(3), 77–82 (2002)
6. Manaris, B., Hughes, D., Vassilandonakis, Y.: Monterey Mirror: Combining Markov Models, Genetic Algorithms, and Power Laws. In: 2011 IEEE Congress on Evolutionary Computation (CEC 2011), 1st Workshop in Evolutionary Music, pp. 33–40. IEEE Press, New York (2011)
7. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Hafner Publishing Company, New York (1949)
8. Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., Davis, R.B.: Zipf's Law, Music Classification and Aesthetics. Computer Music Journal 29(1), 55–69 (2005)
9. Legname, O.: Density Degree of Intervals and Chords. 20th Century Music 4(11), 8–14 (1997)
10. Eigenfeldt, A., Pasquier, P.: Realtime Generation of Harmonic Progressions Using Controlled Markov Selection. In: 1st International Conference on Computational Creativity (ICCC-X), pp. 16–25. ACM Press, New York (2010)
11. Manaris, B., Roos, P., Machado, P., Krehbiel, D., Pellicoro, L., Romero, J.: A Corpus-Based Hybrid Approach to Music Analysis and Composition. In: 22nd Conference on Artificial Intelligence (AAAI 2007), pp. 839–845. AAAI Press, Palo Alto (2007)

12. Manaris, B., Johnson, D., Vassilandonakis, Y.: Harmonic Navigator: A Gesture-Driven, Corpus-Based Approach to Music Analysis, Composition, and Performance. In: 9th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2013), pp. 67–74. AAAI Press, Palo Alto (2013)
13. Berry, W.T.: Structural Functions in Music, pp. 40–57. Prentice Hall, Upper Saddle River (1976)
14. Hindemith, P.: The Craft of Musical Composition, Schott, Mainz, Germany, pp. 87–89 (1945)
15. Manaris, B., Johnson, D., Vassilandonakis, Y.: A Novelty Search and Power-Law-Based Genetic Algorithm for Exploring Harmonic Spaces in J.S. Bach Chorales. In: 3rd International Conference on Evolutionary and Biologically Inspired Music, Sound, Art and Design. Springer, Berlin (2014)

HandyScope: A Remote Control Technique Using Circular Widget on Tabletops

Takuro Kuribara, Yusaku Mita, Kazusa Onishi,
Buntarou Shizuki, and Jiro Tanaka

University of Tsukuba, Japan
{kuribara,mita,onishi,shizuki,jiro}@iplab.cs.tsukuba.ac.jp

Abstract. A large multi-touch tabletop has remote areas that the users might not touch by their hands. This forces users to move around the tabletop. In this paper, we present a novel remote control technique which we call HandyScope. This technique allows users to manipulate those remote areas. Moreover, users can move an object between the nearby area and the remote areas using a widget. In addition, users can precisely point a remote area quickly because this system includes our proposed control-display ratio changing system. To evaluate the performance of HandyScope, we compared HandyScope with direct touch manipulation. The results show that HandyScope is significantly faster in selection.

Keywords: bimanual interaction, multi-touch, gesture, dynamic control-display gain, pointing, target acquisition, pull-out.

1 Introduction

A large multi-touch tabletop allows users to surround the tabletop and touch the tabletop from their respective positions. However, it has remote areas that users might not touch by their hands; for example, touching a distant object displayed on the opposite side of the tabletop is difficult due to the large size of the touch screen. This forces users to move around the tabletop.

To solve this problem, we present a novel remote control technique which we call HandyScope (Figure 1). This technique allows users to manipulate remote

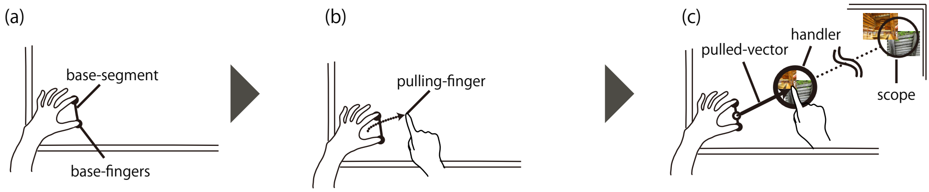


Fig. 1. HandyScope allows users to point and manipulate the remote area. a) When users put two fingers, and b) drag their finger to cross the segment between the two fingers, then c) HandyScope is activated.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 69–80, 2014.

areas (e.g., move, rotate, and resize distant objects) and move an object between the nearby area and the remote areas. In addition, users can precisely point a remote area quickly by using the widget because this system includes the control-display (C-D) ratio changing system which we have already proposed [21].

2 Related Work

Remote pointing techniques have been intensively investigated to facilitate especially pointing on large wall displays. Such techniques are device-based pointing [6,14], gesture-based pointing [19], and gaze-based pointing [8]. In contrast, our technique allows users to point remote areas on tabletops, which adopt a bimanual gesture. Therefore, we focus on pointing techniques for tabletops and studies of bimanual interaction.

Pointing Techniques for Tabletops

Parker et al. used the stylus tip's shadow to point at a remote position [15]. In the work of Banerjee et al. [3], users could point at a remote position on tabletops and dynamically change C-D ratio using one hand while performing a pointing manipulation with the other hand. The above techniques required additional devices that obtain the position of users' hands to realize direct-pointing. Bartindale et al. [5] developed an onscreen mouse for multi-touch tabletops that allows users to point at a remote position, similar to a conventional physical mouse. However, this technique required to use tabletops that allow for a measurement of the area of hand's contact. In contrast, our technique can be applied to tabletops that detect multi-touch points without additional devices and recognizing the shape of hands. Matejka et al. [13] also developed an onscreen mouse, while its activation method is still open in the literature.

I-Grabber [1] is an onscreen widget controlled by bimanual multi-touch interaction. Our technique is also controlled by using bimanual multi-touch interaction. However, our technique allows users to change the C-D ratio and to use only a single multi-touch gesture from activation to pointing. Therefore, users can point precisely and quickly.

Bimanual Interaction

There was some research on bimanual interaction such as 3D operation [16,20], modeling [2,10], and precise selection [7] . In contrast, our technique allows users to point remote areas using bimanual interaction.

Tokoro et al. presented a pointing technique that utilized two acceleration sensors, and postures of both hands pointing at a remote position [17]. Furthermore, Malik et al. developed a bimanual pointing technique by using image processing [12]. In contrast with these techniques, our technique is performed by using touch based gestures.

3 HandyScope

HandyScope allows users to manipulate remote areas using a circular widget. The widget is composed of two parts, a scope and a handler. The scope is sent to remote areas to select an area manipulated; the handler is used to manipulate the remote area by users. The scope area is displayed in the handler; and all events onto the handler area are sent to the scope area. Therefore, users can manipulate (e.g., move, rotate, and resize) the remote objects within the scope, using the handler. Moreover, this technique uses pull-out, a bimanual multi-touch gesture [22]. This gesture allows multiple users to, without conflicting with other touch gestures, simultaneously manipulate remote areas. Below we describe the interaction of HandyScope and its advantages.

3.1 Activation and Control Technique

Figure 1 shows the procedure of HandyScope. Users put two fingers of their non-dominant hand (base-fingers) on a tabletop as shown in Figure 1a. When users drag a finger of their dominant hand (pulling-finger) to cross the segment between the base-fingers (base-segment) as shown in Figure 1b, a circle (scope) is displayed on the ray between the midpoint of the base-segment; another circle (handler) is displayed around the pulling-finger as shown in Figure 1c. If users arrange the pulled-vector, the scope position is updated according to the change. Users can quit control anytime by detaching both of the base-fingers from the tabletop.

3.2 Deciding the Position of Scope with Dynamic C-D Ratio

Suppose that $P_i(x, y)$ and k_i are the i-th scope position and the i-th C-D ratio after i frames have passed since users placed their base-fingers on the tabletop as shown in Figure 2, respectively. Then P_i and k_i are given by the following formulas:

$$\begin{aligned} P_i &= G_0 + \sum_{j}^{i} k_j \mathit{\Delta} V_j, \\ \mathit{\Delta} V_i &= V_i - V_{i-1}, \\ k_i &= \alpha \times \frac{|S_i|}{|S_0|}. \end{aligned} \tag{1}$$

S_0 and S_i are the length of base-segment when base-fingers were placed on the tabletop, and the length of i-th base-segment, respectively. Then, the C-D ratio k_i is calculated as the ratio of the two lengths with α which is a constant. Furthermore, G_0 is the midpoint of base-segment, and V_i is the pulled-vector from G_i to the pulling-finger. Then P_i is calculated using k_i and $\mathit{\Delta} V_i$ (the difference of V_i) caused by moving dominant or non-dominant hand. Both P_i and k_i are calculated in each frame.

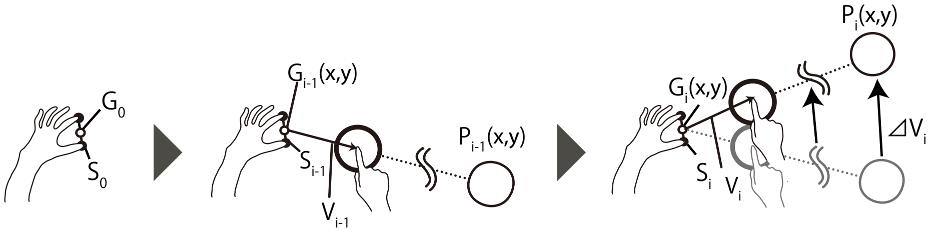


Fig. 2. Moving the circular widget using a simple gesture

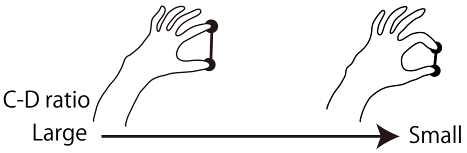


Fig. 3. Dynamic C-D ratio according to the length of base-segment

As (1) shows, k_i, the C-D ratio in our technique, changes depending on the length of the base-segment. Figure 3 shows the relation between the C-D ratio and the length of base-segment. When users pinch out the base-fingers, k_i becomes large. Similarly, when users pinch in the base-fingers, k_i becomes small. This design allows users to selectively perform rough control with a large C-D ratio or precise control with a small C-D ratio, because they can point while controlling the C-D ratio simultaneously. For example, users can move scope roughly and quickly with a large C-D ratio, then they can move the scope precisely and slowly with a small C-D ratio as shown in Figure 4.

3.3 Remote Manipulation Using the Widget

Users can manipulate remote objects using the handler, e.g., resize the remote objects (Figure 5a) and rotate the remote objects (Figure 5b). To achieve this, the scope area is displayed in the handler and all events onto the handler area are sent to the scope area. Therefore, users can manipulate remote objects without walking to remote areas or bringing remote objects to nearby area.

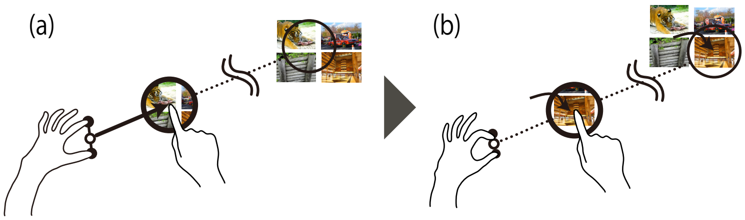


Fig. 4. Usage of dynamic C-D ratio. Users a) roughly point at a distant position quickly with a large C-D ratio, and then b) precisely point at an object with a small C-D ratio.

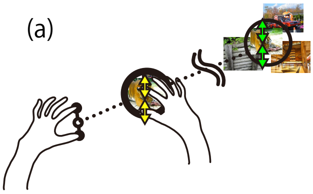


Fig. 5. Manipulating remote objects from nearby area: a) resizing the remote objects and b) rotating the remote objects

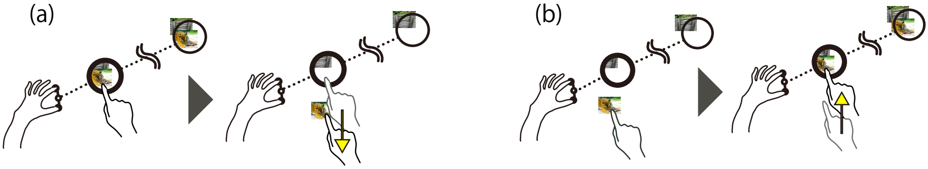


Fig. 6. Transferring objects between a nearby area and a remote area, namely, a) from the remote area to the nearby area and b) from the nearby area to the remote area

3.4 Transferring Objects between Nearby and Remote Area

If users select a remote object in the handler and drag it outside the handler, the remote object is transferred to the nearby area as shown in Figure 6a. Correspondingly, if users select a nearby object and drag it into the handler, the nearby object is transferred to the remote area as shown in Figure 6b. In this way, users can transfer the objects quickly between the nearby and the remote area.

3.5 Adjusting the Widget

Users can adjust the widget by interacting with the edge of the handler. To move the circular widget again to manipulate other remote areas, users drag the edge of the handler as shown in Figure 7. By pinching in and out on the edge of the handler, users can resize the circular widget as shown in Figure 8. In this way, users can manipulate larger objects at the remote areas.

3.6 The Advantages of HandyScope

HandyScope allows users to manipulate remote areas. This is similar to Frisbee [11] or Dynamic Portals [18]. However, Frisbee requires users to determine the remote area in advance; Dynamic Portals needs collaborator(s) to select the remote area. In contrast, HandyScope allows users to activate it from any position and decide the remote area quickly by dynamically changing C-D ratio. Furthermore, it is possible adjust the position and the size again.

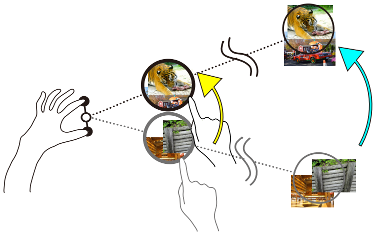


Fig. 7. Moving the circular widget again

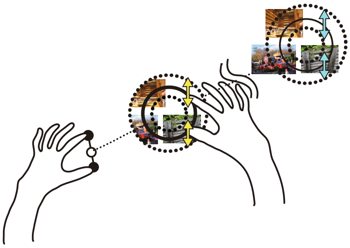


Fig. 8. Resizing the circular widget

4 Evaluation

We conducted experiment to measure the performance of HandyScope. In this experiment, we compared HandyScope (HandyScope condition) with the existing direct touch (Touch condition) in terms of typical three manipulations on tabletops. These three manipulations were Selecting, Rotating, and Resizing.

4.1 Participants and Evaluation Environment

Ten undergraduate and graduate students ranging in age from 20 to 24 (M=22.8, SD=0.5) participated in this experiment. One of them was left-handed. All of them had never used HandyScope.

We show the evaluation environment in Figure 9. As the tabletop in this evaluation, we used a 1470 mm × 80 mm 60-inch display (PDP-607CMX[1]) with a multi-touch function by attaching a multi-touch frame (PQ Lab, Multi-Touch G^{3}[2]). We adjusted the height of the tabletop to 93 cm. This height was selected to be consistent to those of the tabletops in studies on tabletops such as [4,9,23], ranging from 91 cm to 105 cm. We assigned 12 to α of (1) in Section 3.2, so that participants did not need to change the C-D too frequently in this environment.

4.2 Task

We asked the participants to perform Selecting task, Rotating task, and Resizing task, in this order. The design of these tasks follows the ones used in evaluating the pointing technique for tabletops by Banerjee et al. [3]. We asked them to complete a practice task before performing the real ones. The amount of the practice task was 1/4 of the real task. We divided the participants into two groups to counterbalance the order effect between two technique conditions. One group performed the Touch condition first, and the other performed HandyScope condition first. Participants could use each hand freely in this experiment. We asked them to answer a questionnaire after having finished all tasks. The experiment lasted approximately one and a half hour per participant, including

[1] http://pioneer.jp/biz/karte/PDP-607CMX.html

[2] http://multitouch.com/product.html

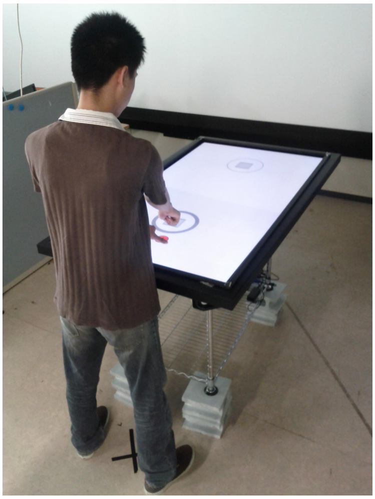


Fig. 9. Experimental environment

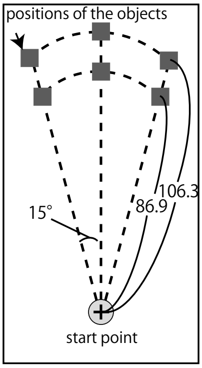


Fig. 10. Positions of start point and target objects

answering the questionnaire. A participant was paid 1640 JPY (approximately 16 USD) for her/his participation.

4.3 Selecting Task

We asked the participants to select a target object displayed at some position. First, a participant stand at the center of one short side of the tabletop (the spot marked by black tape as shown in Figure 9) before each trial. From this position, she/he selected a target object displayed at some position. Figure 10 illustrates the position of both the start point and the target objects displayed on the tabletop. The start point and a target object were displayed before each trial.

In HandyScope condition, a participant started the Selecting task by starting HandyScope on the start point. Then, she/he moved the scope to the target object, and tapped it. When the target object was tapped, the trial was completed and a beep was played. In Touch condition, a participant started the Selecting task by tapping the start point. Then, she/he moved (i.e., walked or ran) to the position where she/he could touch the target object, and tapped the target object.

In this task, independent variables were: target distance (900 and 1100 pixels, i.e. approximately 922 and 1127 mm, respectively), target angle (-15, 0, and 15 degree), target size (40, 60, and 80 pixels, i.e. approximately 41, 61, and 82 mm, respectively), and technique (HandyScope and direct touch).

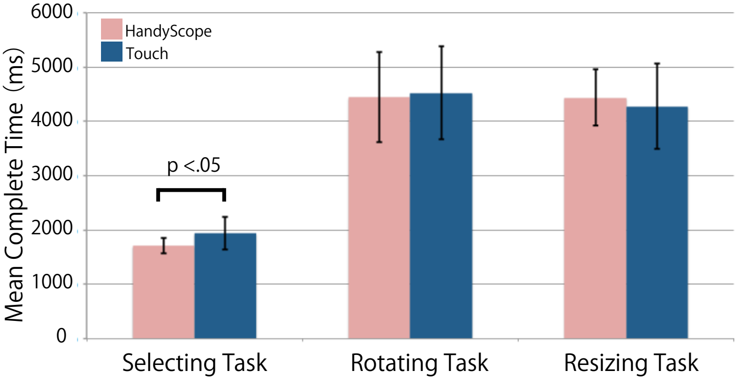


Fig. 11. Mean of the trial-times for each task

Each participant performed 3 trials in each combination of factors, thus performed 108 (2 × 3 × 3 × 2 × 3) trials in total. Independent variables for each technique were presented in randomized order.

Results. We measured the time to complete a trial (trial-time). The left two bars in Figure 11 show the mean of the trial-times with each technique. The mean time was 1942.6 ms in Touch condition, and was 1715.2 ms in HandyScope condition. The result of t-test between the two mean times was t(9)=2.72, p=.011<.050. This result suggests that selecting in HandyScope condition was significantly faster than that in Touch condition.

4.4 Rotating Task

We asked the participants to rotate an object to fit a dock displayed at some position. The object was displayed at the same position as the dock, while its angle was different, to make the participants just rotate the object in this task. The start point, the positions, and the action to start the task were the same as those of Selecting task.

In HandyScope condition, a participant rotated an object to fit the dock by HandyScope. If the angle of the object and the dock were the same (i.e., within ± 5 degree), the color of the object's border became red. In this condition, when she/he finished manipulation, then one trial was completed and a beep was played. In Touch condition, she/he moved to a position where she/he could touch the target object, and then rotated the target object.

In this task, independent variables were: target distance (900 and 1100 pixels, i.e. approximately 922 and 1127 mm, respectively), target angle (-15, 0, and 15 degree), dock size (60 and 80 pixels, i.e. approximately 61 and 82 mm, respectively), rotate angle (-45 and 45 degree), and technique (HandyScope and Touch). Each participant performed 2 trials in each combination of factors, thus performed 96 (2 × 3 × 2 × 2 × 2 × 2) trials in total. Independent variables for each technique were presented in randomized order.

Results. The middle two bars in Figure 11 show the mean of the trial-times with each technique. The mean time was 4520.4 ms in Touch condition, and 4443.5 ms in HandyScope condition. The result of t-test between the two mean times was t(9)=.267, p=.397>.050. There was no significant difference in mean time between each technique.

4.5 Resizing Task

We asked the participants to resize an object to fit the dock displayed at some position. The object was displayed at the same position as the dock, while its size was different. The start point, the positions, and the action to start the task were the same of those of Selecting task.

In HandyScope condition, a participant resized an object to fit the dock by HandyScope. If the size of the object and the dock were same (i.e., within ± 5 pixel), the color of the object's border became red. In this condition, when she/he finished the manipulation, then one trial was completed and a beep was played. In Touch condition, she/he moved to a position where she/he could touch the target object, and then resized a target object.

In this task, independent variables were: target distance (900 and 1100 pixels, i.e. approximately 922 and 1127 mm, respectively), target angle (-15, 0, and 15 degree), dock size (60 and 80 pixels, i.e. approximately 61 mm and 82, respectively), resize direction (expanding and decreasing), and technique (HandyScope and Touch). Each participant performed 2 trials in each combination of factors, thus performed 96 ($2 \times 3 \times 2 \times 2 \times 2 \times 2$) trials in total. Independent variables for each technique were presented in randomized order.

Results. The right two bars in Figure 11 show the mean of the trial-times with each technique. The mean time was 4277.9 ms in Touch condition, and 4438.2 ms in HandyScope condition. The result of t-test between the two mean time was t(9)=-.935, p=.187>.050. There was no significant difference in mean time between each technique.

4.6 Consideration

The mean of the trial-times in HandyScope condition was significantly faster in Selecting task. However, there was no significant difference between techniques in Rotating task and Resizing task. From these results, HandyScope is considered to be useful for selecting a remote area.

In contrast, there was no significant difference between techniques in Rotating task and Resizing task. The possible cause of this derives from the fact that restarting HandyScope took time. In this experiment, there were situations where the participants accidentally detached their fingers before finishing the trial. In this case, they needed extra time to restart HandyScope to manipulate again. In contrast, in Touch condition, they needed little time to manipulate again in such situations, because they had already moved near the target object.

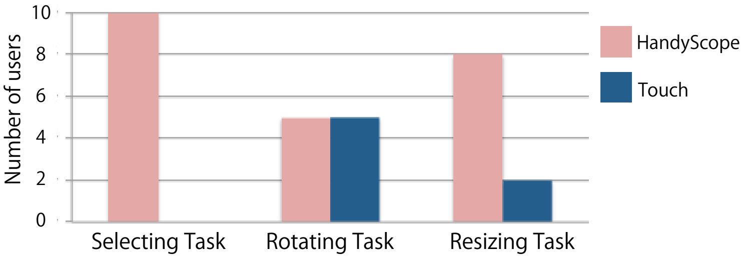


Fig. 12. Questionnaire of preferred technique

Because of this, we considered that HandyScope took time to Rotating task and Resizing task. To avoid accidentally quitting HandyScope, we modify the design of HandyScope to remain activated even if users detach their base-fingers. In this case, we will place an additional button for quitting HandyScope around the edge of the handler; users push this button to quit HandyScope instead of detaching their base-fingers.

4.7 Questionnaire

Figure 12 shows the results of questionnaire asking a favorite technique by task.

In Selecting task, all of participants preferred HandyScope. In addition, in Resizing task, eight out of ten participants preferred HandyScope. As the reason of these results, all of these participants said that they could manipulate remote objects without moving, by using HandyScope.

In Rotating task, five participants preferred HandyScope; other five participants preferred direct touch. Two of the participants said that they prefer direct touch because they could use both hands. In addition, two of the other participants said that they had some trouble in keeping the base-fingers touched on the tabletop. Another of the participants also commented that he had serious troubles in restarting HandyScope when he missed the trial.

In Resizing task, two of the participants who preferred direct touch also commented that they had troubles in keeping their base-fingers on the tabletop.

5 Discussion

To investigate whether multiple users simultaneously manipulate remote areas without conflict with other touch gestures using HandyScope, we conducted a collaborative task which arranged cluttered photos as shown in Figure 13. In this task, twenty photos were displayed on the tabletop. The size, angle, and location of the photos were random. Two of the authors arranged the photos cooperating with each other. We stood around the tabletop and did not walk. If we could touch the photos, we manipulated the photos using direct touch. In contrast, if we could not touch the photos, we manipulated the photos using HandyScope. We continued this task five times.

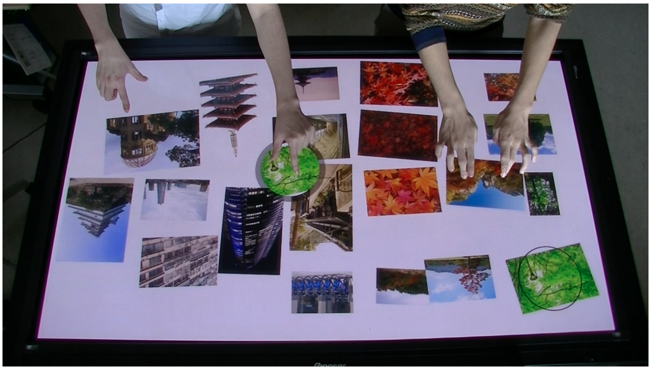


Fig. 13. Collaborative work of multiple users

As a result of this task, we did not observe any accidental activation of HandyScope. Therefore, HandyScope has potential for avoiding conflict with other touch gestures. As future work, we would like to perform a detailed evaluation of collaborative work using HandyScope.

6 Conclusion

We designed and implemented a remote control technique, HandyScope. The technique allows users to manipulate remote areas that users might not touch with their hands. In addition, users can move an object between the nearby area and the remote areas using the widget. The evaluation using the prototype revealed that HandyScope is a useful technique for selecting a remote area. Moreover, the questionnaire results showed that HandyScope is liked by users. In our future work, we plan to investigate the performance of transferring the objects using HandyScope. Moreover, we would like to use HandyScope on large wall multi-touch displays.

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. Balakrishnan, R., Fitzmaurice, G., Kurtenbach, G., Buxton, W.: Digital tape drawing. In: Proc. of UIST 1999, pp. 161–169 (1999)
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. Banovic, N., Li, F.C.Y., Dearman, D., Yatani, K., Truong, K.N.: Design of unimanual multi-finger pie menu interaction. In: Proc. of ITS 2011, pp. 120–129 (2011)
5. 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)

6. Baudisch, P., Sinclair, M., Wilson, A.: Soap: A pointing device that works in mid-air. In: Proc. of UIST 2006, pp. 43–46 (2006)
7. Benko, H., Wilson, A.D., Baudisch, P.: Precise selection techniques for multi-touch screens. In: Proc. of CHI 2006, pp. 1263–1272 (2006)
8. Bolt, R.A.: Gaze-orchestrated dynamic windows. In: Proc. of SIGGRAPH 1981, pp. 109–119 (1981)
9. Furumi, G., Sakamoto, D., Igarashi, T.: SnapRail: A tabletop user interface widget for addressing occlusion by physical objects. In: Proc. of ITS 2012, pp. 193–196 (2012)
10. Grossman, T., Balakrishnan, R., Kurtenbach, G., Fitzmaurice, G., Khan, A., Buxton, B.: Creating principal 3D curves with digital tape drawing. In: Proc. of CHI 2002, pp. 121–128 (2002)
11. Khan, A., Fitzmaurice, G., Almeida, D., Burtnyk, N., Kurtenbach, G.: A remote control interface for large displays. In: Proc. of UIST 2004, pp. 127–136 (2004)
12. 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)
13. 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)
14. Myers, B.A., Bhatnagar, R., Nichols, J., Peck, C.H., Kong, D., Miller, R., Long, A.C.: Interacting at a distance: Measuring the performance of laser pointers and other devices. In: Proc. of CHI 2002, pp. 33–40 (2002)
15. 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)
16. 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: Proc. of CHI 2012, pp. 1297–1306 (2012)
17. Tokoro, Y., Terada, T., Tsukamoto, M.: A pointing method using two accelerometers for wearable computing. In: Proc. of SAC 2009, pp. 136–141 (2009)
18. Voelker, S., Weiss, M., Wacharamanotham, C., Borchers, J.: Dynamic Portals: A lightweight metaphor for fast object transfer on interactive surfaces. In: Proc. of ITS 2011, pp. 158–161 (2011)
19. Vogel, D., Balakrishnan, R.: Distant freehand pointing and clicking on very large, high resolution displays. In: Proc. of UIST 2005, pp. 33–42 (2005)
20. Wang, R., Paris, S., Popović, J.: 6D hands: Markerless hand-tracking for computer aided design. In: Proc. of UIST 2011, pp. 549–558 (2011)
21. Yoshikawa, T., Mita, Y., Kuribara, T., Shizuki, B., Tanaka, J.: A remote pointing technique using pull-out. In: Proc. of HCI 2013, pp. 416–426 (2013)
22. Yoshikawa, T., Shizuki, B., Tanaka, J.: HandyWidgets: Local widgets pulled-out from hands. In: Proc. of ITS 2012, pp. 197–200 (2012)
23. Zhang, H., Yang, X.D., Ens, B., Liang, H.N., Boulanger, P., Irani, P.: See Me, See You: A lightweight method for discriminating user touches on tabletop displays. In: Proc. of CHI 2012, pp. 2327–2336 (2012)

Comparing Hand Gesture Vocabularies for HCI

Alexander Mehler, Tim vor der Brück, and Andy Lücking

Text Technology Lab, Department of Computer Science and Mathematics,
Goethe-University Frankfurt am Main

Abstract HCI systems are often equipped with gestural interfaces drawing on a predefined set of admitted gestures. We provide an assessment of the fitness of such gesture vocabularies in terms of their learnability and naturalness. This is done by example of rivaling gesture vocabularies of the museum information system WikiNect. In this way, we do not only provide a procedure for evaluating gesture vocabularies, but additionally contribute to design criteria to be followed by the gestures.

1 Motivation

Hand gestures are of great interest for HCI applications, since they are considered to help "to develop more natural and efficient human-computer interfaces." [1] There are two kinds of prevalent HCI gestures: *manipulators* and *semaphores* [2]. Manipulators are actions that manipulate some entity provided by the display – for instance, pushing a button or moving a slider. Therefore, manipulators are largely driven by the displayed entity and its functionality. This "tight relationship between the actual movements of the gesturing hand/arm with the entity being manipulated" [2, p. 172] is not a defining feature of semaphores. Rather, semaphoric gestures are hand/arm forms that are organized as a predefined, often stylized vocabulary, or lexicon [2, p. 173]. Such gesture vocabularies can be designed in a better or worse way. Semaphores are considered to be better, if they are more "intuitive" or "motivated". Motivatedness is accomplished if the form (hand shape, movement trajectory) of a gesture "imitates the referent by selecting one or more of its visually perceivable features" [3, 49]. In other words: intuitive gestures resemble their object, they are *iconic*. However, it is well known now that iconic gestures do not signify or refer on their own. Rather, other means are required for establishing signification, for instance, a conventional one [4]. Conventionality involves arbitrariness that has to be mastered by learning. Of course, users favor gesture vocabularies that can be learned easily [5, p. 33]. Accordingly, a second dimension for evaluating sets of semaphores has to be their learnability.

Both lines of assessing the fitness of gesture vocabularies have been pursued in previous research by different methodologies, for example:

- the naturalness of gesture vocabularies has been investigated by [6] by means of user studies;
- the learnability of semaphores (including an empirically specified intuitiveness index) have been studied as an analytical optimization problem by [7].

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 81–92, 2014.

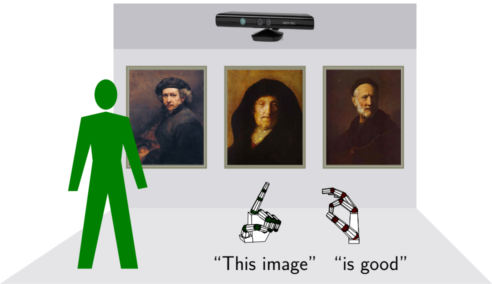


Fig. 1. WikiNect application scenario: rating of an image (taken from [9])

$$\left[\begin{array}{ll} \textit{one.handed.gesture} & \\ \text{Two.Handed.Conf} & 0 \\ \text{Mov.Relative} & 0 \\ \text{RH} & \left[\begin{array}{ll} \textit{right.hand} & \\ \text{Handshape.B} & \\ \text{Handshape.Mov} & 0 \\ \text{Handshape.Path} & 0 \\ \text{Palm.Orient} & \text{PAB} \\ \text{Palm.Mov} & 0 \\ \text{Palm.Path} & 0 \\ \text{BoH.Orient} & \text{BUP} \\ \text{BoH.Mov} & 0 \\ \text{BoH.Path} & 0 \\ \text{Wrist.Loc} & \text{CC} \\ \text{Wrist.Dist} & \text{D-O} \\ \text{Wrist.Mov} & \text{MDR>MUR} \\ \text{Wrist.Path} & \text{LINE>LINE} \\ \text{Mov.Extent} & \text{M>L} \\ \text{Temporal.Sequence} & 0 \end{array}\right] \end{array}\right]$$

Fig. 2. Representation of Checkmark gesture from Table 3

The latter work deals with static hand configurations from one robotic arm control vocabulary. The present paper further develops optimization procedures for gesture vocabularies, mainly in two respects:

1. firstly, in addition to static gesture, also dynamic gestures are accounted for;
2. secondly, evaluation is not only based on one gesture vocabulary, but is carried out as a comparison between different sets of semaphores.

The testing environment for the comparison of gesture vocabularies is the WikiNect system [8] (see also `www.hucompute.org/ressourcen/wikinect`). WikiNect is a platform for the gestural writing of wikis in the context of museums. Using the Kinect technology, WikiNect allows for a non-contact, gesture-based segmentation, linkage, attribution and rating of (segments of) images. As an on-site museum information system, WikiNect aims at enabling museum visitors to describe, evaluate and comment images of the corresponding exhibition. In Figure 1 (taken from [9]) a typical WikiNect application scenario is given where a user selects an image by means of a pointing gesture and appreciates it using a semaphoric, codified "OK" gesture.

Being an HCI application that is addressed to the diverse audience of museum visitors, WikiNect itself has an interest in natural and learnable gesture-based interactions. Accordingly, the gesture vocabularies to be evaluated are taken from two prototype implementations of WikiNect [10,11]. To this end, Section 2 describes the gesture vocabularies in conjunction with a subset of tasks accomplished by WikiNect. Section 3 accounts for task-gesture mappings in terms of a quadratic optimization problem. It starts from a quantitative analysis of Wikipedia-based image descriptions which results in a corresponding set of soft constraints. The evaluation rationale and experimentation for assessing gesture

Table 1. Selected tasks accomplished by WikiNect

Navigation Tasks	Segmentation Tasks	
Scrolling backward	Select image	Circular segment
Scrolling forward	Segment image	Rectangular segment
Close, back to Main	Save image	Polygonal segment
Undo	Display segments	Free-hand segmentation

Table 2. Spatial expressions partitioned according to three spatial modalities *Direction*, *Relations*, and *Form*

Direction	Relation		Form
left	above	behind	circle
right	below	through	rectangle
up	by	at	triangle
down	in	on	cornered
front	around	between	bent
back	in front of	along	random
			straight

vocabularies is finally presented in Section 4, while Section 5 provides a concluding discussion.

2 WikiNect Gestures, Tasks and Annotations

The usage of WikiNect is subdivided into a navigation and a segmentation component [8]. Navigation gestures are used for selecting WikiNect's functional modules, while segmentation gestures are operative in the segmentation mode. Table 1 lists 12 of these tasks which have been implemented in two prototype systems according to different design strategies [10,11]. The first prototype, hereafter called WN-1, provides a set of controlling gestures taken from the InkCanvas class of the `.NET` Framework and mapped onto the system's operations [10]. The second prototype, WN-2, can be operated mainly by manipulation gestures (e.g., by pushing buttons that trigger a certain operation) [11].

Any gesture used to implement WikiNect has been represented in terms of spatial predicates. The rationale behind this is to allow for task-gesture mappings: gestures are preferably mapped to tasks with which they share many predicates. In order to obtain a set of spatial predicates, we use the list of the spatial predicates collected by [12, p. 97]. This list has been extended by (1) the directions spanned along the body axes and (2) basic form-related predicates. The spatial predicates are partitioned according to the spatial modalities *direction*, *relation* and *form* – see Table 2. They are used to label both the tasks and the gestures for spatial properties, either quite literally or associatively. Some notes on the application of the predicates:

Table 3. Navigational gestures used in [10] for implementing WikiNect

Gesture Annotation	Image	Movement	Task	Task Annotation	Naturalness
right, straight, along	→	towards right	Scrolling backward	back, below, left	0
left, straight, along	←	towards left	Scrolling forward	front, up, right	0
through, cornered, down, up, right	✓	Checkmark, towards down-left, towards up-right	(1) Select image, (2) Segment image, (3) Save image	(1) around, through; (2) in, through; (3) in, random	(1) 0.077; (2) 0.1; (3) 0
right, through, cornered, up, around, above	⬏	towards right, upward	Close active window, back to main	back, below, left	0
through, right, cornered, down, around, below	⬎	towards right, downward	(1) Undo, (2) Display segments of an image	(1) back, down, left, random; (2) in, around, by, random	(1) 0.033; (2) 0.031

- If a movement comprises a change of direction, it is understood as to run *through* the turning point and the predicate “through” is chosen.
- If a task contains a temporal aspect like *backwards* (i.e., going back in the system’s history), three conceptualizations are acknowledged:
  1. Stack – orientation along longitudinal axis (“up”, “down”);
  2. Tape – orientation along transversal axis (“right”, “left”);
  3. Gaze – orientation along sagittal axis (“front”, “back”).
- Closed forms give rise to containment indicated by “in”.

We emphasize that the annotation so far has the status of a working hypothesis. We aim at demonstrating that our approach is feasible and provides useful results without claiming that the predicate list is the only possible one.

For illustration, the description and annotation of gestures and tasks of WN-2 is given in Tables 3 and 4. The columns “Movement” and “Image” contain a shorthand and a pictorial representation of the gestures. The column “Naturalness” shows the naturalness index calculated according to the procedure explained in Section 4.1. To make the gestures’ forms objects of quantitative analyses, they are coded according to the kinematic-oriented representation format of [13] – see Figure 2 for an example. Based on text-based representations of this kind, we apply distance measures in optimizing task-gesture mappings.

Table 4. Segmentation gestures used in [10] for implementing WikiNect

Gesture Annotation	Image	Movement	Task	Task Annotation	Naturalness
circle, around, bent	○	Circle	Cut out circular segment	circle, bent, in, around	0.18
around, rectangle, cornered	□	Rectangle	Cut out rectangular segment	in, cornered, around, rectangle	0.15
around, triangle, cornered	△	Triangle	Cut out polygonal segment	triangle, in, around, cornered	0.15
through, left, right, down, cornered, between	<	Towards down-left, towards down-right	Activate free-hand segmentation	random, in, around, circle, rectangle, triangle	0

3 Towards Optimal Task-Gesture Mappings

The task of image description is schematized to a certain degree [14]. WikiNect deals with four such routinized tasks: *rating*, *segmenting*, *linking* and *attributing* images (e.g., with information about painters or techniques). Our aim is to find gestural representations of theses tasks so that users can make image descriptions by using WikiNect, that is, by *gestural writing* [9]. A naïve way to realize this would be to select from an artificial lexicon of prespecified gestures. The problem is rather how to justify any mapping of image description tasks onto gestures. An iconic gesture, for example, is a natural candidate to manifest a gestalt-related image description, while a deictic gesture is a better candidate for selecting images on the screen.

Our approach to solve this problem is twofold: firstly, we analyze Wikipedia as the biggest sample of image descriptions to learn about the frequency distributions of the actions involved in such descriptions. Secondly, we utilize this information to derive constraints that any procedure of gesture selection should fulfill to provide both efficiently producible and learnable gestures for gestural writing. This approach follows a twofold optimization criterion: we select gestures for actions of image descriptions such that the more frequent the action the more easily producible the gesture while preserving a certain amount of discriminability (i.e., learnability) among gestural manifestations of different actions.

Information about the frequency distributions of image description tasks is not directly accessible for lack of large-scale annotations of corresponding speech acts. However, the English Wikipedia offers a range of data to approach this information. To learn about the frequency distribution of linking images, for example, we can explore hyperlinks between articles about these images (see Table 5 for a

Table 5. Statistics of image description articles in the English Wikipedia

Attribute	Value
articles	2,862
instances of painting/artwork template	2,926
links among the 2,862 articles	62,725
corpus size	14.7 MB
average size (per article)	5.3 KB
date of extraction	2014=2012 November 1, 2014

statistics of the underlying corpus; see Figure 3 for the resulting distribution (distributions have been shifted by one, to account for zero frequencies)). Likewise, to get information about the frequency distribution of image attributions, we explore every instance of `Template:Infobox_artwork`[1] (Figure 4). Next, since there is no matching template for segmenting images, we need to assess the corresponding frequency distribution indirectly. This is done by exploring the frequency distribution of section headers like `Composition`, `Analysis` or `Details` within the corpus of image articles (Figure 5). Likewise, because of the lack of directly accessible ratings of images, we explore the ratings of their corresponding articles (as manifested by the `Rate this page`-section). In this way, we approximate a frequency distribution of image-related ratings (Figure 6). As can be seen by Figures (3–6), each of the four tasks (*linking*, *attributing*, *segmenting* and *rating*) results in a power-law-like frequency distribution being reminiscent of Zipf's law of least effort [15]. Only a couple of images is, for example, linked to many other images while most images are linked only once (Figure 3). Likewise, there is a small set of predominant attributes while most attributes are rarely used if at all. Further, the frequency distribution of section headers shows a small set of predominant sections (Figure 5) that leave behind a huge set of rarely used ones: apart from conventional sections in Wikipedia (e.g. `References` or `External links`), the former set is exemplified by headers like `Artist`, `Description` and `Composition`. That is, when writing about the content of images, Wikipedians follow a power law according to which they prefer a small range of topics of highest probability. Analogously, the distribution of the numbers of ratings strictly follows a power law (Figure 6) in any of the four dimensions considered by Wikipedia: a few images have many ratings while most images have few ratings or none at all.

In sum, image descriptions follow a highly skewed distribution such that the frequencies of the underlying actions decay according to a power law. Thus, when looking for gestural manifestations of such actions we can follow the example of natural languages [15]: the more frequent an action the simpler its manifestation should be. Since we need to manifest different actions simultaneously, we additionally need to preserve discriminability among neighboring ranks in the

[1] We also explore `Template:Infobox_Painting` which redirects to `Template:Infobox_artwork`.

frequency distribution of gestural manifestations. As a rule of thumb: *optimizing along the criterion of least effort should not happen at the expense of discriminability and thus learnability among highly frequent gestures.* In what follows, we represent this finding in terms of a *quadratic integer programming problem* whose solution leads to the optimal task-gesture mapping – subject to the operative constraints (number of tasks, gesture repertoire etc.).

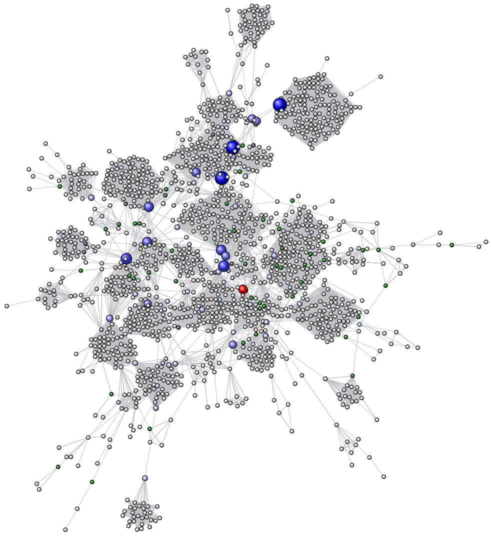


Fig. 3. The largest weakly connected component of the article graph of image descriptions in the English Wikipedia that covers 56% of the descriptions. The distribution of the node degrees of this graph follows a power law with exponent 1.55 (according to [16], we fit the complementary cumulative distribution $P(X \geq x)$ that yields an exponent of 0.55 ($\overline{R}^2 = 94\%$); according to [17] this corresponds to an exponent of 1.55 in terms of $P(X = x)$. The same procedure is applied in all fittings).

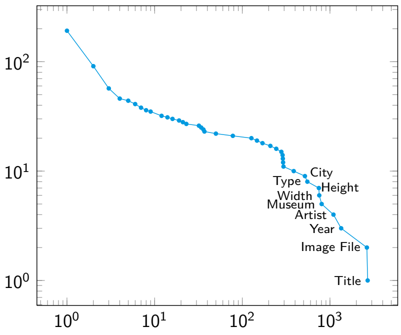


Fig. 4. 1-shifted complementary cumulative distribution of attribution templates for images (exp. 1.26 (2.26), $\overline{R}^2 = 97.7\%$)

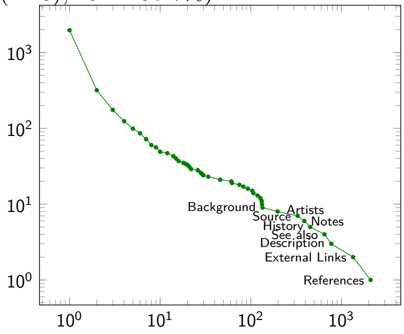


Fig. 5. Complementary cumulative distribution of section headers (exponent 1.55 (2.55), $\overline{R}^2 = 99\%$)

Generally speaking, a quadratic integer programming problem requires all decision variables to be integer, while its constraints are required to be linear and the objective function to contain a quadratic term. To reformulate this in terms of gesture modeling, we proceed as follows: let n be the number of gestures and m the number of tasks. Assume that gestures and tasks are all numbered so that the set of tasks is given by $T = \{t_1, \ldots, t_n\}$ and the set of gestures by $G = \{g_1, \ldots, g_m\}$. The decision variables in this mapping problem

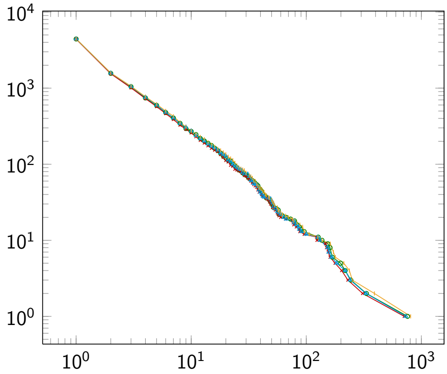


Fig. 6. 1-shifted complementary cumulative distributions of ratings show four distributions of the rating template (*trustworthy* (green circle, exp. 0.7 (1.7), $\overline{R}^2 = 99\%$), *well-written* (orange bars, exp. 0.72 (1.72), $\overline{R}^2 = 99\%$), *objective* (red crosses, exp. 0.7 (1.7), $\overline{R}^2 = 99\%$), and *complete* (blue triangles, exp. 0.7 (1.7), $\overline{R}^2 = 99\%$))

Table 6. Frequency distribution of tasks by predicates

Task	**Freq.**	**Perc. %**
Circular segment	1,180	10.37
Close, back to Main	719	6.32
Display segments	1,399	12.29
Free-hand segmentation	1,189	10.45
Rectangular segment	1,172	10.3
Save image	1,121	9.85
Scrolling backward	719	6.32
Scrolling forward	827	7.27
Segment image	1,132	9.95
Select image	62	0.54
Polygonal segment	1,171	10.29
Undo	691	6.07
Sum	11,382	100

are binary features x_{ij} that are 1 if gesture g_i should be mapped to task t_j and zero otherwise. A hard constraint is to require that each task is always mapped to a single gesture, i.e., synonymous gestures and not-assigned tasks are not allowed. We formalize this by means of equality constraints:

$$\sum_{i=1}^{n} x_{ij} = 1 \text{ for } j = \{1, \ldots, m\} \; ; \tag{1}$$

Since the number of gestures exceeds the number of taks, some gestures have to be polysemous and are assigned to several tasks. For the gestures, we only require that each gesture is assigned to at least one action:

$$\sum_{j=1}^{m} x_{ij} \geq 1 \text{ for } i = \{1, \ldots, n\} \tag{2}$$

In addition to hard constraints, three soft constraints are encoded into the objective function:

1. *The simpler the gesture, the more frequent the action to which it is mapped.*
2. *The more frequent an action, the more motivated the gesture mapped onto it.* Since the mapping of gestures to actions has to be memorizable, it should be motivated as much as possible (as explained in Section 4.1).
3. *The more frequent two actions, the easier the discriminability of their gestural manifestations.*

We represent Constraint 1 and 2 by a linear model and Constraint 3 by a squared term as part of the objective function. Given two sets of tasks and gestures (Section 2), an assessment of the motivation of any candidate task-gesture-relation (Section 4.1), a frequency distribution of tasks (Section 4.2), and a measure of the discriminability of gestures (based on their matrix representations – see Figure 2 and [13]), we finally get an optimization problem whose solution, henceforth called **gesture optimizer**, leads to an optimal task-gesture mapping subject to the operative constraints.

4 Experimentation

In this section, we compare two instantiations of the gesture optimizer and contrast them with their corresponding null-models of random task-gesture assignments. To this end, we utilize both implementations of WikiNect (see Sec. 2).

For instantiating the optimizer, we first need to specify two boundary conditions: the motivation of task-gesture relations and the frequency distribution of image description tasks.

4.1 On the Naturalness of Task-Gesture Relations

In order to find an optimal mapping of gestures onto tasks, one needs to know the degree of motivation by which a candidate gesture fits as a manifestation of the tasks. If a user wants to move, for example, something to the left of the display, it is a bad choice to signal this by moving the hand to the right. We provide a simple quantification of this sort of naturalness in terms of bipartite graphs whose bottom mode comprises the candidate gestures and whose top mode is spanned by the tasks under consideration. For any pair $\{g, t\}$ of gestures g and tasks t, an edge occurs in the graph whose initial weight equals the overlap of the predicate descriptions $P(g)$ and $P(t)$:

$$w_1(\{g, t\}) = \frac{|P(g) \cap P(t)|}{\min(|P(g)|, |P(t)|)} \tag{3}$$

Next, we account for diversification in the bipartition. The reason is to prefer unifying task-gesture mappings (in terms of $1:1$ mappings). To see this, think of a system of n tasks, $n \gg 2$, mapped onto one or two gestures. Because of the polysemy of the gestures (as a function of the predicates assigned to them), this system tends to be unnatural: it leads to a semantic overload of the gestures in question. Thus, we re-weight edges as follows (d_v is the degree of vertex v in the bipartition):

$$w_2(\{g, t\}) = w_1(\{g, t\}) \cdot \frac{2}{d_g + d_t} \tag{4}$$

Obviously, a $1:1$-mapping does not alter w_1. Conversely, if the gesture is polysemous or the task is manifested by different gestures, then $w_2 < w_1$. Finally, for any gesture (task), we get a rank order of tasks (gestures) according to their decreasing degree of naturalness. Note that the edge weights are ordinally scaled.

Table 7. Assignments determined by the optimizer for scenarios 1 and 2

Task	Gesture (Scenario 1)	Gesture (Scenario 2)
Scrolling backwards	Left	Grab and drag left
Close window	Left	Grab and drag left
Save Image	Left	Grab and drag right
Display image segments	Right below	Grab and drag right
Scrolling forward	Right	Push forward
Selection	Right above	Push forward
Segment Image	Checkmark	Push forward
Circular segment	Circle	Push forward
Free-hand segmentation	Circle	Push forward
Rectangular segment	Rectangle	Set image point
Polygonal segment	Triangle	Set image point
Undo	Open triangle	Grab and drag left

4.2 Towards a Frequency Distribution of Image Description Tasks

In order to provide a frequency distribution of image description tasks for implementing the *gesture optimizer*, we cannot rely on the Wikipedia data explored in Section 3. The reason is that we focus on the specific task list of WikiNect (see Table 3 and 4). Thus, we alternatively analyze a specialized corpus of image descriptions [18]. The aim is to estimate the probability by which the tasks of Table 1 are conducted in sessions of image description. Since the Wally corpus [18] does not annotate this information and since some of the focal tasks are even not observable in the corpus, we account for this probability indirectly. Following the former sections, we relate tasks and gestures by the predicates they share in their descriptions (see Table 3 and 4). As we map a range of expressions onto these predicates (e.g., *round* and *around* are explored as manifestations of the same-named predicate `around`), the mapping is done by observing the corpus frequencies of the predicates' verbal manifestations. The result of this mapping is shown in Table 6. In contrast to our findings of Section 3, this distribution does not fit a power-law. This may hint at insufficient or even erroneous descriptions of tasks and gestures. For example, though we additionally accounted for multi-word expressions (e.g., *in the front of*), we did not resolve paraphrases of spatial descriptions. Thus, Table 6 has to be understood as a first attempt to estimating the frequencies in questions.

4.3 Results

We tested our approach on two scenarios: given the set of tasks listed in Table 1, the scenarios are distinguished by using the WN-1 and WN-2 set of gestures, respectively. For both scenarios, we determined the optimal assignment for the decision variables by means of the Gurobi optimizer[2] and therefore the optimal mapping

[2] `http://www.gurobi.com`

Table 8. Values of the objective function as determined by the optimizer and the base line method

	Scenario 1	Scenario 2
Optimized value	−2.09	−1.14
Baseline value	−0.98	−0.55

from tasks to gestures that minimizes the objective function (see Table 7 for the optimal mappings).[3]

As a base line, we estimated the expectation value of the objective function by generating 1,000 random assignments of tasks to gestures that fulfill the hard constraints of the optimization problem. The evaluation shows that the optimizer determines assignments for both scenarios for which the objective function values are lower than the base line values (see Table 8 – recall that the lower the objective value the easier to learn and more natural the assignment). Furthermore, the optimal value of the objective function of scenario 1 is below the optimal value of scenario 2, which indicates that scenario 1 is the superior one in terms of learnability and naturalness. Since the number of gestures exceeds the number of tasks in both scenarios, some gestures have to be assigned to more than one task. As can be seen in Table 7, for instance, the *Circle* gesture from scenario 1 is assigned to both the tasks *circle* and *free-hand segmentation*, since both tasks can be chosen in the same context. Thus, the gesture *Circle*, which intuitively is strongly related to circular segmentation mode, gets ambiguous under this assignment. This observation hints at context as a further parameter for improving our model in future work.

5 Conclusion

Based on the notions of *learnability* and *naturalness*, we provide the *gesture optimizer*, a method to assess the fitness of HCI gesture vocabularies to a set of tasks. Optimization is expressed as a quadratic integer programming problem sensitive to a number of constraints. The method is tested in a gesture vocabulary comparison of two WikiNect implementations. Given frequency information of the tasks, a discriminability order between the gestures and a naturalness index based on spatial annotations for gesture-task mappings, we found that the gesture optimizer not only distinguishes gesture vocabularies from a random baseline, but also ranks the vocabularies in the intuitively correct way. Thus, in order to provide an assessment for HCI gestures, the gesture optimizer fuses information and considerations from different sources. Not all of these sources are fully developed yet. However, even given these conditions, we could show that *naturalness*, *frequency* and *learnability* are effective design criteria for devising good HCI gesture vocabularies. This result shows that existing vocabularies

[3] For the second scenario, the optimizer was able to determine the optimal value, for the first scenario we used the best solution found before reaching a time limit.

(think, e.g., of touch gestures!) can be evaluated and, possibly, improved. The gesture optimizer also delineates criteria for designing new vocabularies, so that the method proposed here has many practical applications and provides a test bed for further studies on the fitness of HCI gestures.

References

1. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 677–695 (1997)
2. Quek, F.K.H., McNeill, D., Bryll, R.K., 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)
3. Poggi, I.: Iconicity in different types of gestures. Gesture 8(1), 45–61 (2008)
4. Burks, A.W.: Icon, index, and symbol. PPR 9(4), 673–689 (1949)
5. Baudel, T., Beaudouin-Lafon, M.: Charade: Remote control of objects using free-hand gestures. Communications of the ACM 36(7), 28–35 (1993)
6. Grandhi, S.A., Joue, G., Mittelberg, I.: Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. In: Proceedings of SIGCHI, pp. 821–824 (2011)
7. Stern, H., Wachs, J., Edan, Y.: A method for selection of optimal hand gesture vocabularies. In: Sales Dias, M., Gibet, S., Wanderley, M.M., Bastos, R. (eds.) GW 2007. LNCS (LNAI), vol. 5085, pp. 57–68. Springer, Heidelberg (2009)
8. Mehler, A., Lücking, A.: WikiNect: Towards a gestural writing system for kinetic museum wikis. In: Proceedings of UXeLATE, pp. 7–12 (2012)
9. Mehler, A., Lücking, A., Abrami, G.: WikiNect: Image schemata as a basis of gestural writing for kinetic museum wikis. Universal Access in the Information Society (2014) (accepted)
10. Inceoglu, M.R.: WikiNect. BA thesis, Goethe University Frankfurt (2013)
11. Asir, A., Creech, B., Homburg, T., Hoxha, N., Röhrl, B., Stender, N., Uslu, T., Wiegand, T., Kastrati, L., Valipour, S., Akemlek, D., Auth, C., Hemati, A.: Korchi, Said Omari, S., Schöneberger, C.: Praktikum WikiNect (2013)
12. Klein, W.: Raumausdrücke. Linguistische Berichte 132, 77–114 (1991)
13. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, LREC 2010, pp. 92–98 (2010)
14. Jörgensen, C., Jaimes, A., Benitez, A., Chang, S.: A conceptual framework and empirical research for classifying visual descriptors. Journal of the Am. Soc. f. Inf. Sci. a. Tech. 52(11), 938–947 (2001)
15. Zipf, G.K.: Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology. Hafner Publishing Company, New York (1972)
16. Newman, M.E.J.: Power laws, Pareto distributions and Zipf's law. Cont. Phys. 46, 323–351 (2005)
17. Adamic, L.A.: Zipf, power-law, Pareto — A ranking tutorial (2000), `http://www.hpl.hp.com/research/idl/papers/ranking/`
18. Rohde, H., Clarke, A., Elsner, M.: Wally referring expression corpus (wrec) v0.1.0, (dataset). University of Edinburgh. School of Philosophy, Psychology, & Language Sciences. Linguistics & English Language (2013)

Effectiveness of Virtual Hands in 3D Learning Material

Tetsufumi Mikami and Shu Matsuura

Tokyo Gakugei University, Faculty of Education, 4-1-1 Nukuikita, Koganei,
Tokyo 184-8501, Japan
a080359f@st.u-gakugei.ac.jp, shumats0@gmail.com

Abstract. A virtual reality model for a motional electromotive force physics experiment, “Fleming’s rail,” was designed and developed. A hand gesture interface was constructed to control a virtual simulation using a Microsoft Kinect sensor and a finger-gesture interface SDK. A gesture-based object tracking test was performed to examine the effects of virtual hand visualization. In addition, motion trajectories of real hands with and without hand visualization were analyzed. Trajectories obtained with hand visualization exhibited higher Hurst exponent values compared with those obtained without virtual hand visualization. This suggests that the displacement change was more persistent with positive fluctuation feedback, indicating sensory feedback for real hand motions. For comparison, the effects of the model on learning Fleming’s left- and right-hand rules were experimentally tested. Results exhibited that knowledge acquisition from the model was almost equivalent to that from the real experiment.

Keywords: Hand gesture interface, virtual reality learning material, Hurst exponent.

1 Introduction

In introductory physics, it is important to understand invisible physical quantities, such as force, energy, electric current, and voltage, because certain visible phenomena are explained from invisible physical quantities; e.g., a visible standing wave is explained as a superposition of two invisible waves traveling in opposite directions. To this end, mathematical simulations have been useful for generating visualizations of such physical mechanisms, and computer simulations have become popular as resources for learning [1].

Attempts have been made to apply virtual reality (VR) technology to the visualization of physics simulations in order to merge theory with real experiments and phenomena [2]. Such attempts have been partially intended for the correction of common beginner misconceptions [3]. The augmented reality technique is effective to visualize invisible components in real objects and phenomena [4]. In addition, projection mapping techniques are applicable to represent virtual components directly onto real objects and to make objects interactive using augmented reality.

With these approaches, one can interact with real physical objects by using virtual components that supplement these with physical properties. This will be effectively

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 93–101, 2014.

achieved by designing a natural user interface[5, 6] for the virtual components to be manipulated as real objects.

The purpose of this study is to develop a VR model for a physics experiment of motional electromotive force, i.e., "Fleming's rail." A hand gesture interface was introduced using a Microsoft Kinect sensor to manipulate the model. By using the hand gesture interface SDK provided by 3Gear Systems Inc., user hand gestures were exhibited as virtual hand gestures. The virtual hands manipulated the model's components according to the user's real hand motions.

This visualization is expected to help the user manipulate the VR model and explore the physical phenomena shown in it. In addition, this virtual experiment is conducted simultaneously with a real (i.e., non-virtual) version of the same experiment. Thus, the user can learn from both the real and virtual experiments at the same time.

In this paper, the motion trajectories of real hands are analyzed both with and without virtual visualization. Results showed that higher positive feedback for the motions was observed with the visualization of virtual hands. In addition, the effect of the model in attaining the basic knowledge of Fleming's left- and right-hand rules was compared using the virtual and real experiments separately. It was concluded that the effectiveness of the virtual experiment was equivalent to that of the real experiment.

2 Methods

2.1 Development Environment

The gestural interface SDK that tracked hand and finger motion was provided by 3Gear Systems [7]. An application was constructed using Light Weight Java Game Library 2.9.0 with OpenGL for 3D graphics. The motion sensors used were a Microsoft Kinect for Xbox 360 and an ASUSTek Xtion PRO. The development and tests were performed on an Apple Mac mini with a 27-inch display. The Kinect sensor was set 65 cm above the surface of the desk where hand gestures were performed.

2.2 Model

Fleming's rail consists of two long parallel rails on which a mobile conducting bar is mounted. The virtual model has a scale of 510 × 100 arbitrary units, which correspond to 510 × 100 mm as detected by the Kinect sensor.

The model is switched between generator mode and motor mode. In generator mode, the left sides of the rails are connected, thereby forming a circuit (Fig. 1 left). The mobile conducting bar can be moved according to the motion of the user's right hand. As the conducting bar is moved in the magnetic field, the electrons of the bar are driven by magnetic force to generate an electric current within the circuit.

In motor mode, the left sides of the rails are connected with an external battery to form a circuit (Fig. 1 right). The electric current provided by the battery is again driven by magnetic force, and the mobile bar moves in the direction of the rail. In this mode, the voltage of the battery varies as per the height of the left hand.

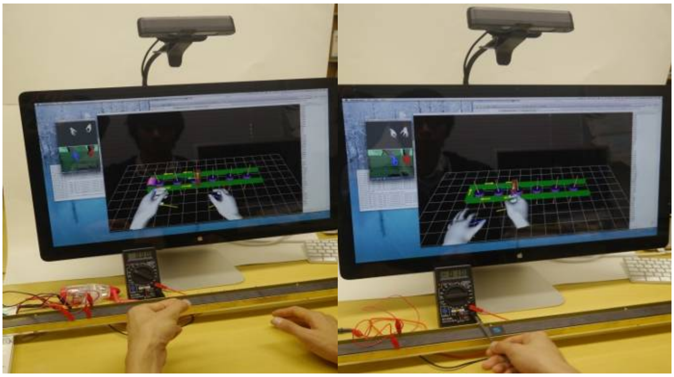


Fig. 1. Real Fleming's rail experiment and the VR model: (left) motor mode; (right) generator mode

Generator mode is initiated by a right-hand pinch gesture, and motor mode is initiated by a left-hand pinch gesture. For the motor, the directions of the electric current, magnetic field, and mechanical force are assigned to the middle finger, first finger, and thumb of the left hand, respectively. Similarly, electric current, magnetic field, and mechanical force are assigned to corresponding fingers of the right hand. In our model, the directions of these vectors are presented on the corresponding virtual hands; thus, the user can verify the relationships of these vectors by comparing them with those shown in the computer display.

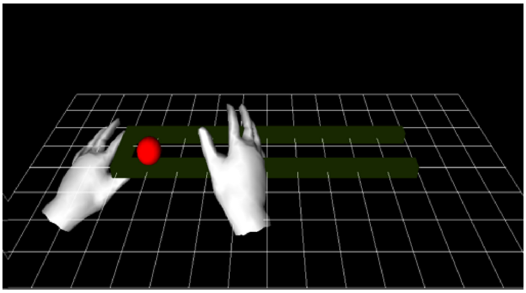


Fig. 2. 1D tracking test

2.3 Tracking Test

To compare the participant's hand motion with and without virtual hand visualization, a simple tracking test was introduced. A spherical object is generated at a randomly chosen position in the virtual space where the model rails are set (Fig. 2). When the participant's right hand point position collides with the sphere, it is moved to another randomly selected position. The position of the real hand was defined as the joint base of the middle finger. The participant tracks the sphere, and the Kinect sensor detects the trajectory of the user's hand motions. The sphere was generated in one, two, and three dimensions. For 1D tracking, the sphere was placed within $-230 \leqq x \leqq 230$

(arbitrary unit; corresponds to mm in real space), $y = 30$, and $z = 0$. The y-axis corresponds to a vertical line, and the z-axis corresponds to depth. For 2D tracking, the area for sphere motion was within $-230 \leqq x \leqq 230$, $y = 30$, and $-50 \leqq z \leqq 50$. For 3D tracking, the sphere appeared within $-230 \leqq x \leqq 230$, $30 \leqq y \leqq 130$, and $-50 \leqq z \leqq 50$. Under these conditions, the target sphere was captured approximately once per 20 frames of position detection.

During the tracking test, participants were asked to practice tracking in 3D for 30 s. Then, they attempted 1D tracking, followed by 2D and 3D tracking with virtual hand visualization. For each dimension, tracking continued for approximately 20 s and was repeated three times. The same procedure was repeated without the virtual hands. There were seven participants: four male and three female undergraduate university students.

For analysis, the following Hurst exponent H was calculated for hand motion trajectories. Let $x(t)$ be the value of a fluctuating variable at time t. Then, for arbitrary time difference Δt, the standard deviation $h(\Delta t)$ of the difference of the variable $X_t(\Delta t) = x(t) - x(t + \Delta t)$ tends to exhibit a power law, which is expressed as follows.

$$h^2(\Delta t) \approx \Delta t^{2H} . \qquad (1)$$

The exponent H is the Hurst exponent. If the fluctuation is a non-correlated Brownian fluctuation, $H = 0.5$. Thus, as H increases, the change of fluctuation tends to sustain with the positive feedback. In turn, as H decreases below 0.5, development of fluctuation is suppressed with negative feedback.

2.4 Classroom Practice

To compare the effectiveness of the virtual experiment with the real one, a classroom practice study was performed. Thirty students from a literature class were divided into groups A and B and were made to take a pretest before the main instructions were provided. Group A, which consisted of 14 students (10 female; four male), was instructed using the VR model. Group B, which consisted of 12 students (eight female; four male), was instructed using the real Fleming's rails experiment. After the initial instruction, a posttest was carried out. For the second instruction, instruction materials were exchanged between Groups A and B. Finally, another posttest was performed. Each participant attempted to conduct the virtual experiment once.

The pretest and posttests consisted of two questions regarding the use of Fleming's left- and right-hand rules, and a 5-point Likert scale questionnaire was used to determine the confidence level of the answers. The degree of difficulty was slightly increased from the pretest to the second posttest.

3 Results

3.1 Tracking Test

Figure 3 shows examples of 1D hand motion tracking with and without virtual hand visualization. The range of hand motion in the x-direction is rather large for the case

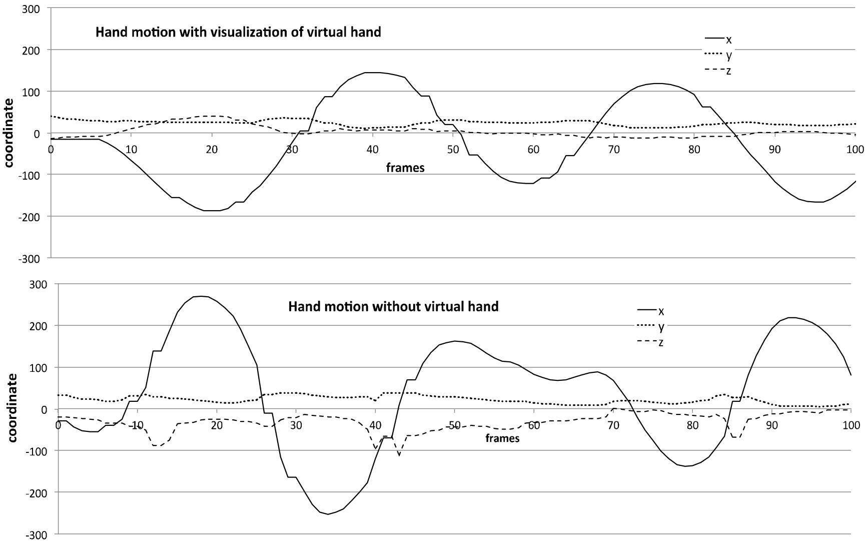


Fig. 3. Real hand motion trajectories for 1D tracking in x-, y-, and z-directions: (top) with virtual hand visualization; (bottom) with virtual hand visualization

with no hand visualization, which indicates that the rendering of the virtual hand reduces excessive motion. In addition, fluctuations in the y- and z-directions are smaller with the virtual hand, which indicates that the real hand motions are smoother and less excessive, and the user may be more careful when tracking motions.

Figure 4 shows log–log plots of Eq. (1) for 1D tracking with virtual hands. In the x-direction, the motion naturally persists and h_x increases. In the y- and z-directions, deviation is suppressed and motions are smooth with high H values.

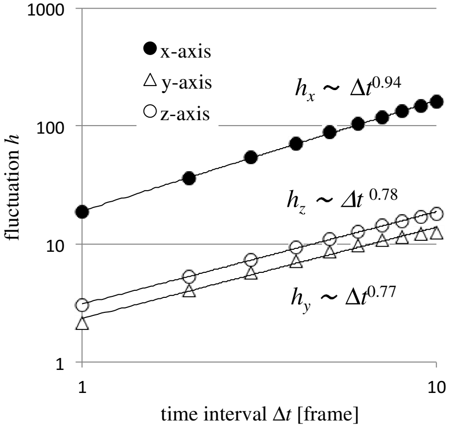


Fig. 4. Log–log plots of displacement h vs. time difference Δt for x-, y-, and z-directions for 1D tracking with virtual hand visualization

Figure 5 shows a comparison of Hurst exponent H of the hand motions with and without virtual hand visualization. This figure shows the results of trajectories for all three dimensions. Hurst exponent H for the motion with virtual hands showed higher values than without the visualization. This implies that both acceleration and deceleration toward the target were smoother and sustained when the virtual hands were displayed. Paired t-tests did not support the hypothesis that there would be no difference in the mean values of H with and without the virtual hands (significance level of 0.05); p-values were 2.0×10^{-5} for 1D tracking, 1.7×10^{-4} for 2D tracking, and 1.1×10^{-6} for 3D tracking.

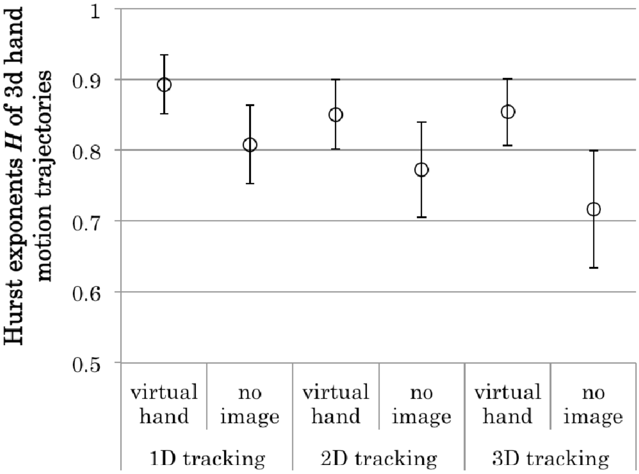


Fig. 5. Comparison of Hurst exponents of hand motion trajectories with and without virtual hand visualization for 1D, 2D, and 3D tracking (error bars show standard deviations)

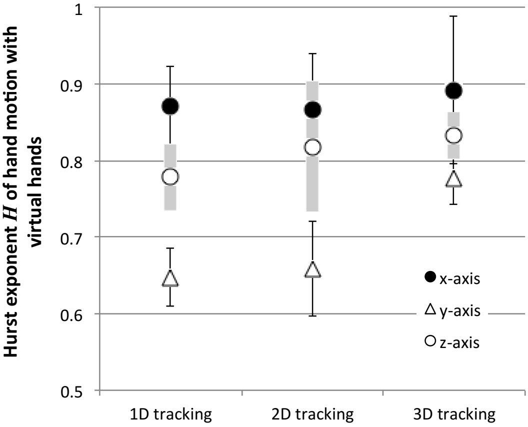


Fig. 6. Spatial dimensions dependency of Hurst exponents of hand motion with virtual hand visualization. The target sphere appears on the x-axis in 1D tracking, and it appears on the xz-plane in 2D tracking. The error bars show standard deviations; error bars for the z-axis are thick and light gray.

Figure 6 shows the dependence of Hurst exponent *H* to the spatial dimensions of tracking with virtual hand visualization. For 1D and 2D tracking, hand motion was primarily restricted to the xz-plane, which corresponds to the surface of the desk. In these cases, motion in the y-direction was not persistent; thus, more non-correlated randomness was observed, as is shown by the *H*-values that are close to 0.5. In 3D tracking, the range of motion was extended vertically, and the motion was also smooth, as is shown in the increased *H*-value as compared with the 1D and 2D tracking.

3.2 Classroom Practice

Figure 7 shows the rate change of correct answers. In addition, Fig. 8 shows the change of the self-confidence histograms for the participants' answers. Before instruction, 51% of participants strongly denied self-confidence, and the total rate of correct answers was 39%.

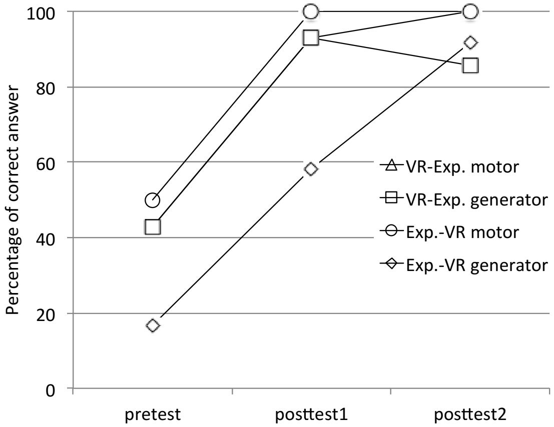


Fig. 7. Change of the percentage of correct answers for two types of instructions and two types of questions

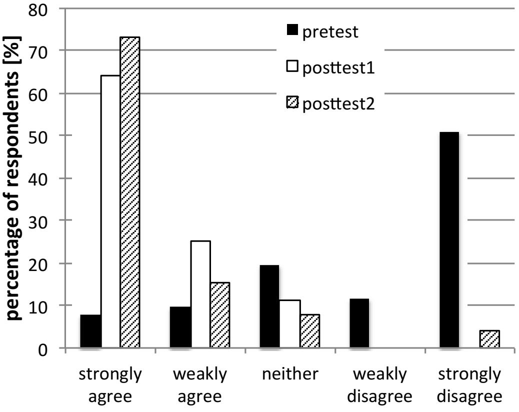


Fig. 8. Changes in confidence levels for student answers from the pretest and two posttests

After the first instruction, 64% of participants strongly agreed that were confident, and the total rate of correct answers was 87%. For the total seven failed answers, six cases were for the question on the generator, and five were from group B, which was first instructed with the experiment. This may partially reflect the fact that Fleming's right hand rule is not always taught in Japanese high schools. After the second instruction, 78% of participants strongly agreed that they were confident in their answers, and the total rate of correct answer was 94%.

The results, which show that the rate of correct answers increased and the distribution of confidence reversed, indicate that both the VR experiment and the real experiment were effective for learning Fleming's left and right rules. In addition, several subsequent tests were conducted for the same participants. The test results show that the rate of total correct answers obtained the following week was 83% and was 98% the week after that.

Figure 9 shows a comparison of confidence level histograms between the two instruction orders, i.e., from VR experiment to real experiment type (light color) and from real experiment to VR experiment type (hatched dark color). No obvious deviation was found between the confidence distributions of these instruction orders. A Welch two sample test showed that the hypothesis, i.e., the difference in means for the first posttests was expected to be zero, was supported at a significance level of 0.05 (p-value of 0.78). In addition, the test showed that the hypothesis for the second posttest was rejected with a p-value of 0.53.

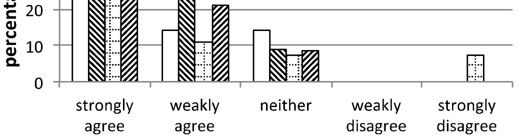


Fig. 9. Comparison of the student self-confidence histograms for two types of instruction through two posttests

These results suggest that the effectiveness of the VR model for learning Fleming's left- and right-hand rules does not differ between the real and virtual Fleming's rails experiments. However, from the descriptions of the impression of this entire session, many students commented that the virtual experiment was effective for confirming what was learned in the real experiment. This suggests that it is helpful in general to use VR material during or after a real experiment. In addition, some participants commented that this VR content was memorable because the virtual space manipulation was similar to the real experiment.

4 Conclusion

VR learning material for Fleming's rails was constructed, and a natural interface for manipulation was produced using hand gesture input through a Microsoft Kinect sensor. A virtual space object-tracking test demonstrated that the Hurst exponent of trajectories of real hand motions was higher when the virtual hands were visualized. This suggests that visual hand motion feedback results in smoother physical motion, which in turn facilitates more effective tracking.

Classroom practice revealed that the virtual experiment is almost equally effective as a real experiment for learning Fleming's left- and right-hand rules. By simulating hand manipulation, the VR material was a relatively natural and effective supplement to the real experiment. This may be partially due to the strong relationship between hand movements and Fleming's left and right hand rules. In this sense, the gestural interface may be applicable to learning materials that are related to somatosensory stimulation.

Acknowledgments. This study has been funded by a Grant-in-Aid for Scientific Research (C) 24501042 from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors would like to thank Enago (www.enago.jp) for the English language review.

References

1. Wieman, C.E., Adams, W.K., Perkins, K.K.: PhET Simulations that Enhance Learning. Science 31, 682–683 (2008)
2. Yang, K.-Y., Heh, J.-S.: The Impact of Internet Virtual Physics Laboratory Instruction on the Achievement in Physics, Science Process Skills and Computer Attitudes of 10th-Grade Students. J. of Sci. Edu. Tech. 16(5), 451–461 (2007)
3. Trindade, J.E.: Improving Physics learning with virtual environments: An example on the phases of water. Interactive Educational Multimedia 11, 212–236 (2005)
4. Liarokapis, F., Petridis, P., Lister, P.F., White, M.: Multimedia Augmented Reality Interface for E-learning (MARIE) 1(2), 173–176 (2002)
5. Wigdor, D., Wixon, D.: Brave NUI World, Designing Natural User Interfaces for Touch and Gesture. Morgan Kaufmann, Burlington (2011)
6. Oviatt, S.: The Design of Future Educational Interfaces. Routledge, New York (2013)
7. 3Gear Systems Inc., http://www.threegear.com

Proposal of the Effective Method of Generating Characteristic Gestures in Nonverbal Communication

Toshiya Naka[1] and Toru Ishida[2]

[1] Panasonic Corporation R&D, Japan
[2] Department of Social Informatics, Kyoto University, Japan
naka.tosiya@jp.panasonic.com,
naka.toshiya.25w@st.kyoto-u.ac.jp

Abstract. According to the rapid spread of the Internet, the new devices and web applications using the newest multimedia technologies are proposed one after another and they become commodity in an instant. In these new web communications, the natural and intelligible interaction corresponding to the user's various demands is required. In the communication in which persons do the direct dialogue in the interaction not only on the web but also in real world, it is widely known by the psychology field that the nonverbal information which is hard to express in words such as expression of face and gesture is playing the important role. In our research, the new analysis method of interaction using the dynamical model is proposed and paid our attention to the characteristic gestures especially. These gestures are the special motions such as lively or powerful actions which used effectively in Kabuki, anime, dance and the special gestures in the speech and presentation of attracting audiences. By analyzing the mechanisms of these characteristic gestures mathematically, we can design the new interactive interfaces easily which are natural and familiar for all users.

Keywords: Nonverbal Communication.

1 Introduction

The conventional researches about the role of nonverbal information such as the facial expression and gestures have been studied by cognitive and social psychologists for many years. It was advocated by A. Mehrabian in 1981 that there were many rates which was occupied by nonverbal information farther than verbal one [1]. M. Wagner found out in 2004 that the gesture would play very important role when forming the place which shared early stage of communications in elementary school education [2]. Furthermore, D. McNeil discovered in 2005 that the language was assumed to constitute the independent communication channel, although gesture and language were the same growing points [3]. In the research on the interaction of human and computers, B. Reeves pointed out in 1996 that people tends to treat computers and other media as if they were either real people or real places, it was called by Media equation [4]. And B. Shneiderman in 1997 advocated that there was two poles of dialog with direct communication and agent (including 3D character), and he pointed out that the effect on the dialog with the humanoid agent was skeptical [5]. On the other hand, since it is

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 102–112, 2014.

the natural interaction and does not need special operation, there are many researches for the dialog with 3D character. There is the interesting research of expression of body language using an interactive robot by T. Nishida [6], he showed that it was insufficient just to reproduce motions and facial expression but also needs to express the higher order expression such as emotion called Conversational informatics. But many of old researches based on psychology have some problems that they are used subjective evaluation and lack in reproducibility.

There is the field of mechanical robot control where prosperous research of quantitative analysis of human's motions is proceeding. In this field, S. Kudoh in 2006 analyzed that the adaptive control of balance in a walk of humanoid robot by defining the moment of robot's arm in four-musculars model [7]. And Y. Uno showed in 1989 that there was the relation between four-muscular model in motion of human's upper arm and torque as the bell type velocity change [8]. In the kinematical analysis of sports, M. Feltner analyzed in 1986 the movements of shoulder and arm in pitching of baseball [9] and C. Putnam conducted to show rules of pitching motion quantitatively between upper arm and torque of joint in football [10]. But these conventional researches aimed the specific sports motions, so there was little research which paid its attention to the gesture in communication. Under these circumstances, we tried to make modeling the mechanism of characteristic gesture communication by referring to these motion evaluating methods.

2 Definition of Nonverbal Information

In this section, the nonverbal information which is the main theme of our research is defined and categorized. There is the following well known nonverbal information in human communication. "**Facial expressions"**: They not only express person's individuality but also include the much information such as emotions, internal feeling and intention. "**View direction"**: Many feeling information is included in the movement of eyes such as turned away ones eyes, winks and gazing and so on. "**Pose and gesture"**: Many communicating information is included like gesture, pose and motion of hands and figures. Furthermore adding the **Individual distance** to them, people are taking various communications by selecting them according to each situation appropriately. Among this nonverbal information, we tried to quantify the structure of gesture communication because of the numerical analysis was not performed until now. The characteristic gestures can be classified into the following two categories like **Reality** and **Actuality** shown in Table1.

Table 1. Definition and classification of gestures

Characteristics of gestures	
Reality	Actuality
Smooth and flowing motion Not awkward motion	Vivid performance, powerful and persuasive action
Correctness and continuity of accuracy	Creation of presence, sense of closeness and persuasion. ex. Gesture of emphasis and **exaggeration**

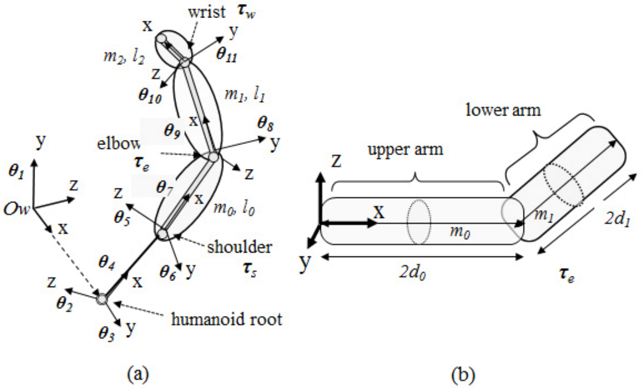
“Reality”: Correctness of motion, continuity and smoothness of accuracy are required. It is the level in which numerical evaluation, analysis and reappearance are possible.

“Actuality”: Gestures which are intentionally used by persons although not necessarily the natural motion but they give some strong impression for recipient. It is the motion that is intentionally used for emphasis (exaggeration) in the remarkable speech and attracting audiences, too. We estimated that people could feel the sense of closeness to 3D characters when this actuality was realized.

3 Definition of Mathematical Analysis Model

We selected the **kinematic dynamics** of articulated structure as the mathematical model in order to analyze gestures quantitatively. The torque value τ which arises at each joint (it expresses the strength of exaggeration in motion) can be given using the angle data θ of each joint defined by multiple skeletal structure shown in Figure 3. It can be given by the equation of motion in Equation 1. In Equation 1,θ is the time series data of each joint angle $[\theta_1,\theta_2,\cdots\theta_{11}]$, M is inertia matrix, C is coriolis force, also g shows the gravity andθ' andθ'' are the angular velocity and angular acceleration for every joint, respectively.

$$\tau = M(\theta)\,\theta'' + C(\theta, \theta')\,\theta' + g(\theta) \tag{1}$$



(a) (b)

Fig. 1. Three dimensional skeletal model of 3D character (right arm)

Furthermore, Lagrange function L is defined by Equation 2 when we expressed Equation 1 by generalized coordinate system using joint angle θ(i = 0～n), also the equation of motion Q_i is given by the Equation 3 using L.

$$L = \sum_{0<i<n} \{(\text{Kinetic energy of link i}) - (\text{Potential energy of link i})\} \tag{2}$$

$$Qi = \frac{d}{dt}\left(\frac{\partial L}{\partial \theta'_i}\right) - \frac{\partial L}{\partial \theta_i} \qquad \text{where } i = 0 \sim n \tag{3}$$

The following nonlinear ordinary differential equations can generally be described by the equation of motion Q_i as Equation 4. In the right-hand side of Equation 4, the first term is the angular velocity, second term shows the force of coriolis and centrifugal force, and the third term is gravity, respectively. In case of rotational movement, Q_i turns into torqueτ which arises at each joint as shown in Equation 5. Furthermore, T_j shows the conversion matrix which translates into the world coordinate system from the local coordinate of the j-th joint, J_i is the inertia tensor of *j*-th link and m_i shows the mass of *i*-th link, g^T is the gravity vector and S_j expresses the position vector of center of gravity of the *j*-th link.

$$Qi = \sum_{j=i}^{n}\sum_{k=0}^{j} trace \left[\frac{\partial T_j}{\partial \theta_k} J_j \frac{\partial T_j}{\partial \theta_i}\right] \ddot{\vartheta}_k + \sum_{j=i}^{n}\sum_{k=0}^{j}\sum_{l=0}^{j} trace \left[\frac{\partial 2 T_j}{\partial \theta_l \partial \theta_k} J_j \frac{\partial T_j}{\partial \theta_i}\right] \ddot{\vartheta}_k \ddot{\vartheta}_l$$

$$- \sum_{j=i}^{n} m_j \, g^{\mathrm{T}} \frac{\partial T_j}{\partial \vartheta_i} S_j \qquad \text{where } \mathrm{i} = 0 \sim \mathrm{n} \tag{4}$$

$$\tau = \mathrm{J}\,(\vartheta_i)^T \begin{pmatrix} f \\ N \end{pmatrix} \tag{5}$$

In this paper, when calculating the inertia tensor *Ji*, each link is approximated with the elliptic cylinder shown in Figure 1-b, therefore the density distribution inside of each link sets constant. We estimate that d_{mp} is minute mass at the point P (mass center of gravity at each link is the point (x_p, y_p, z_p) in local coordinate system) in the rigid body, then inertia tensor *H* of the circumference of center of gravity at each link can be denoted by Equation 6.

$$\mathrm{H} = \begin{bmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{bmatrix} \tag{6}$$

Where $I_{xx} = \int (y_p{}^2 + z_p{}^2)\ dm_p$, $I_{yy} = \int (z_p{}^2 + x_p{}^2)\ dm_p$ $I_{zz} = \int (x_p{}^2 + y_p{}^2)\ dm_p$,

$$I_{xy} = I_{yx} = \int x_p y_p dm_p,\quad I_{yz} = I_{zy} = \int y_p z_p dm_p,\quad I_{zx} = I_{xz} = \int z_p x_p dm_p$$

If we approximate each link such as shown in Figure 1-b, the inertia procession *J* will be given by the Equation 7. However, the length of the elliptic cylinder shall be *2d* and center of gravity is at the starting point, further the length direction is defined by x-axis, y-axis and z-axis in the direction which intersects perpendicularly with them. The y-axis of the path of the ellipse of the section which intersects perpendicularly within the length direction, and the length of z shaft-orientations are set to *a* and *b*, respectively.

$$J = \begin{bmatrix} d^2m/3 & 0 & 0 & d/2 \\ 0 & a^2m/2 & 0 & 0 \\ 0 & 0 & b^2m/2 & 0 \\ d/2 & 0 & 0 & 1 \end{bmatrix} \tag{7}$$

4 Gesture Evaluation System

Block diagram of gesture assessment system is showed in Figure 2. In this system, gestures of humans (actors) are captured using motion tracking [11], and they are changed into time-series-data θ(θ1, θ2, ... θn) of each joint angle shown in Figure 3. And torque τ is calculated which arises at each joint by using kinematical dynamics analysis. In case when the direct dynamical analysis is used, we can calculate the change of time-series-data θ(θ1, θ2, ... θn) of joint angle from each joint torque τ. Therefore, we can generate or correct the gestures of 3D character. In our proposed system, the special data of the body which required for the calculation of kinematical analysis is used such as standard Japanese body shape data like m0 is 1.49 kg, m1 is 1.08 kg and m2 is 0.24m, l0 is 0.28 m , l1 is 1.08m and l2 is 0.17m, respectively.

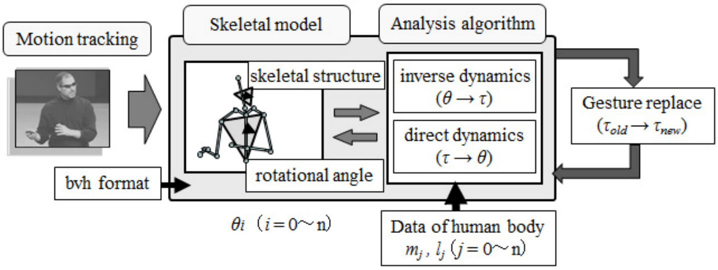


Fig. 2. Gesture analysis system and process using 3D character

Furthermore, the value T_n is defined by Equation 8 which is integrated with the sum of time squares derivative value from start time t_s of motion to end time t_e at each joint torque τ as the total amount torque value of each gesture. Generally, T_n expresses the size of time average torque change of each gesture. When T_n is quantitatively small, it means that few amounts of change of motion and the degree of emphasis of gesture will be small movement.

$$T_n = \int_{t_s}^{t_e} \left\{ \left(\frac{d\tau_s}{dt} \right)^2 + \left(\frac{d\tau_e}{dt} \right)^2 + \left(\frac{d\tau_w}{dt} \right)^2 \right\} \tag{8}$$

In this equation,τ_s ,τ_e andτ_w are torque values of each joint which are calculated by Equation 1 and they are raised at the shoulder joint, elbow and wrist of the dynamic model of 3D character shown in Figure 3, respectively.

5 Experiments and Results

Some experiments by using the characteristic gestures and results in order to verify the usefulness of our proposed model are described in this section. As the experimental gestures used for the proof, we selected twenty-five characteristic gestures carefully which were the motions such as using in Walt Disney's anime, dance, theater, Japanese Kabuki and the emphasizing (exaggeration) gestures of effective speech and presentations. The classification of the gestures which used for the experiments and the feature of each motion are listed in Appendix. In Appendix, Category A includes the action like "kime (finalized action)" and "tame (emphasis/exaggerating motion)" of Japanese Kabuki, Category B is the motion of dance movement of Laban's classification [12] and Category C is gestures of anime characters effectively used in Walt Disney’s movie [13] and characteristic exaggerating gestures used by the emotional expression technique [14]. Furthermore, we classified in Category D which is the exaggeration gestures used in the comedy of Japan called Manzai. In order to compare the effect of emphasis gestures in the communication, we selected the gestures in Category E which are often used for the method of persuading in the speech and presentations like Mr. S. Jobs and Mr. B. Obama who charm audience and attract attention. All experimental results showed that the motion of these gestures had strong correlation with the value of torque τ of main joints which were able to calculate by our proposed model (See Figure 3 and 4 [11]).

6 Gesture Generation and Compensation by Direct Dynamics

By using the direct dynamical analysis method , it is possible to calculate each joint angle $\theta(\theta 1, \theta 2, ...\theta n)$ from the amount of changes of joints torque. When the torque value τ of multiple joint skeletal structure of 3D character is generated or corrected, then we can estimate the movement of link of each skeletal structure using Newton-Euler method as follows. The angular acceleration θ” (θ”1, θ”2, ...θ”n) which arises at each joint of 3D character is given by following Equation 10 by transforming Equation 1.

$$\theta'' = M(\theta)^{-1}\{\tau - C(\theta, \theta')\theta' - g(\theta)\} \tag{10}$$

In case the displacement angle$\theta(0)$ and articular velocityθ' (0) of each joint of 3D character at the time t = 0 (which is starting position) and the torque value $\tau(t)$ (which arises at each joint from time *ts* of start time to target time *te*) are given, then we can obtain the numerical solution of $\theta(t)$ by defining some suitable value of Δt by solving $\theta(t)$ and θ'' (t) of each joint at the time t = 0,Δt, 2Δt, 3Δt ..., one by one until the purpose time *te*. And$\theta(t)$ and θ'' (t) at time t are obtained, the value of $\theta(t+\Delta)$ in time t+Δ andθ'' (t+Δ) can be calculated using the equation of motion in Equation 10, and by supposing the value of θ'' (t) of Equation 10 is still more nearly constant in the minute time interval [t and t+Δ] is drawn by Equation 11.

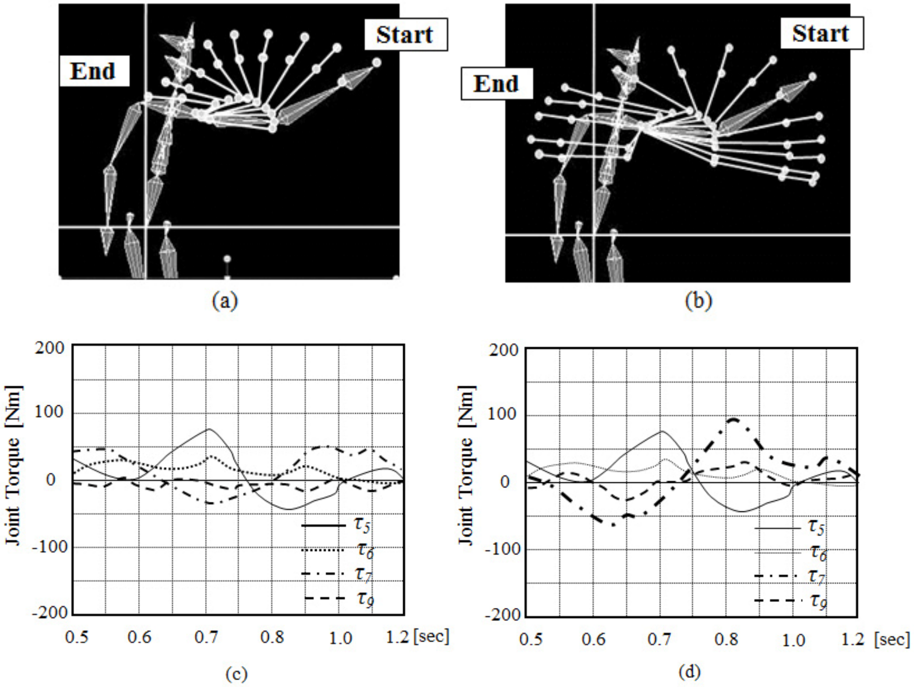
$$\theta(t+\Delta t) = \theta(t) + \theta'(t)\Delta t + (\Delta t)^2\frac{\theta''(t)}{2} \quad (11)$$

Furthermore, it is possible to omit the third term which is square of Δt of Equation 11 because it is minute, then we can obtain Equation 12. The solution of Equation 12 is approximated to the clause of the first item of Euler series expansion one by one to the last time te with Δt. We selected the calculating step Δt of Newton-Euler method to 0.0029 by considering the convergence time. We checked the convergent accuracy error of calculation from our exploratory experiment, even if it used minute Δt value beyond this value [11].

$$\theta(t+\Delta t) = \theta(t) + \theta''(t)\Delta t \quad (12)$$

7 Verification about Naturalness of Generated Gestures

In this section, the validity of our proposed model is verified by using twenty-five characteristic gestures which are listed in Appendix. We compared the naturalness (actuality) of generated gestures with the original motions by using the following method. As the basic experiments, some exaggerated gestures are generated by **Inverse dynamical** calculation since the difference appears clearly. For these gestures, we replaced the torque value τ of the natural motion (without exaggeration) by newly calculated torque value τ_{new} for each gesture which was classified from Category A to Category E in Appendix. They both move same start and target position (destination) correctly.

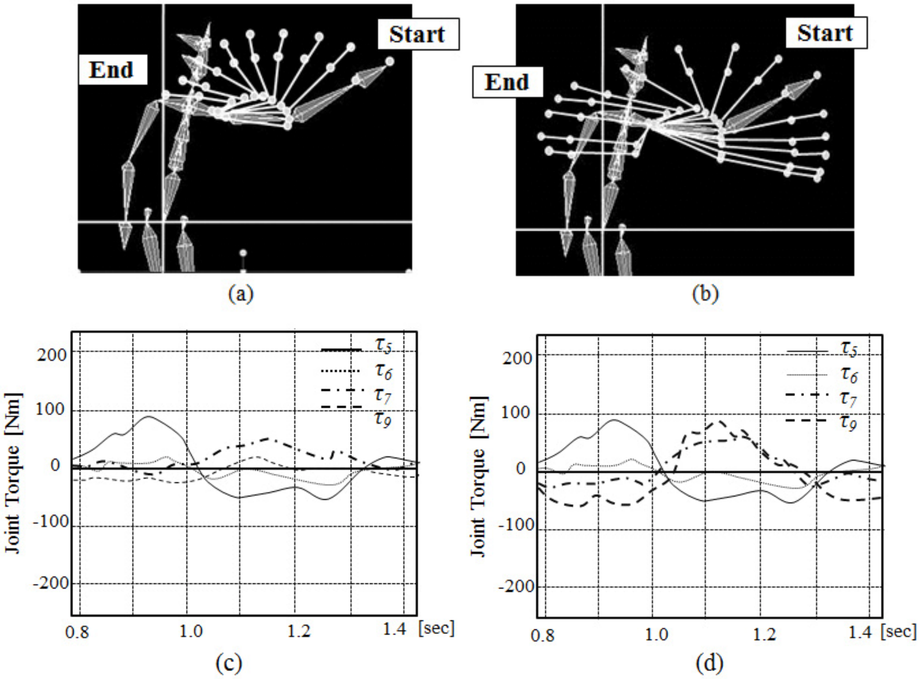


(a) (b)

(c) (d)

Fig. 3. (a) Natural gesture (b) Exaggerating gesture and each Torque (Gesture 18)

In Figure 3, we showed the typical example of (a) natural gesture without exaggeration and (b) gesture with exaggeration as representative case of Gesture 18 in Category D. Also in Figure 3 (c) and (d), the inward and outward rotational swing torqueτ_5 in horizontal plane (rotation of circumference of z-axis) of right shoulder is expressed with solid line, it is the main link of multiple joint skeletal structure. The external and inner rotation torque (rotation of circumference of x-axis) τ_7 is shown by dashed line and long dashed line shows the inward and outward rotational swing torqueτ_9 of left elbow, and outward swing and the adduction torqueτ_6 (rotational of circumference around y-axis), respectively.al In Figure 4, we showed another typical example of (a) natural gesture without exaggeration and (b) gesture with exaggeration as representative case of emphasis gesture in speech Geture 21 in Category E. The inward and outward rotational swing torqueτ_5 (rotation of circumference of z-axis) in horizontal plane of right shoulder were expressed with the solid line in Figure 4-(c) and (d). The external and inner rotation torque (rotation of circumference of x-axis) τ_7 showed by dashed line and long dashed line was the inward and outward rotational swing torqueτ_9 of right elbow and outward swing and the adduction torque (rotation of circumference around y-axis) τ_6, respectively. We tried to compensate by replacing the torque value of the main link of multiple joint skeletal structure of natural motion (without exaggeration) by the exaggerating torque value τ_{new}. Some typical cases are shown as follows.



(a) (b)

(c) (d)

Fig. 4. (a) Natural gesture (b) Exaggerating gesture and each Torque (Gesture 21)

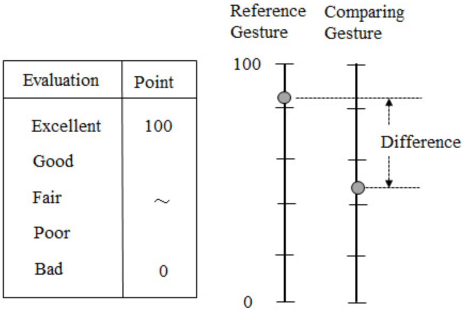
Case-1: As the natural Gesture18 in Category D, we tried to replace the torque value of external and inner rotation torqueτ7 of left shoulder and inward and outward rotational swing torqueτ9 of left elbow by each exaggerating torqueτnew7 andτnew9.

Case-2: Same as Case-1 of natural Gesture21 in Category E, we replaced the torque value of external and inner rotation torqueτ_7 of right shoulder and inward and outward rotational swing torqueτ_9 of right elbow by exaggerating torquesτ_{new}.

In both cases, new gestures of each joints anglesθnew （θ1,θ2・・・θn ） from new torque τnew are given by using the above mentioned direct dynamical method.

8 Results

We conducted to verify the reproducibility of actuality based on the subjective evaluation to the characteristic gestures which were newly generated by our proposed kinematical method. As for the evaluation, we used DSCQS (Double Stimulus Continuous Quality Scale) method of subjectively comparing the newly generated gestures with original ones. The flow of evaluation of DSCQS method is as follows. We showed evaluators the original exaggeration gesture as reference about 10 sec and placing interval of about 3 sec after that, we showed the newly generated exaggeration gestures about 10 sec. These trials were set into one pair and shown twice repeatedly. Each evaluator was requested to perform evaluation to both gestures at the time of second presentation. In these experiments, the order of presentation was changed at random without teaching each one which was the gesture of the original (reference). Twelve men and women (nine men and three women) of adult in twenties were selected for evaluator. Each evaluator was asked to mark the subjective evaluating value over each pair of gestures with continuation measure based on the five steps of quality as shown in Figure 5. Furthermore, the final score was normalized to 0 to 100 (maximum of measure is 100), and the evaluation value of the new exaggerating gesture from the difference of the reference was used as Evaluation difference (DE). Ten evaluators average value was adopted for this DE value as the last evaluation result (the maximum and minimum difference of evaluation result was accepted of each trial). This DE value shows the difference of subjectivity value of the nature of gestures. In case the impression of naturalness will be strong then DE become small (near the natural exaggerating gesture). It can be said to be one index of natural impression (actuality) when it has small value.



Evaluation	Point
Excellent	100
Good	
Fair	～
Poor	
Bad	0

Fig. 5. Measure and value of DSCQS evaluation

All evaluating results of above mentioned DSCQS method are shown in Table 2. The subjective evaluation values of the original exaggerating gesture used as reference, the value of newly generated exaggeration, and the evaluation difference DE value are listed, respectively. As for the result of Case 1, the average value of subjectivity evaluation of the original exaggeration gesture of Gesture 18 in Category D was 91.0, the average value of the newly generated exaggeration gesture became 69.8 and evaluation difference DE was set to 21.4. As for the result of Case 2, the average value of the original exaggeration gesture of Gesture 21 in Category E was set to 97.3, the average value of newly generated exaggeration gesture was 80.0 and evaluation difference DE became 17.3. Furthermore, total average value of subjectivity evaluation of the original exaggeration gestures became from 80.1 to 98.2 for all the gestures used for experiment from Category A to Category E, and the average value of the generated new exaggeration gesture was through 44.3 to 85.7 and each difference DE was set to 35.8 to 12.5. In all categories, the most natural exaggerating gesture was Category E with the subjectivity value of 80 to 90 percent. The DE value of the newly exaggerating gesture was the value from 60 to 70 percent of near impression (**actuality**) for other categories.

Table 2. Results of subjective evaluation of each exaggerating gestures

		Reference gesture	Generated gesture	DE value
Category A	Gesture 1	90.1	55.0	35.1
	Gesture 3	85.9	55.1	30.8
Category B	Gesture 6	78.3	47.1	31.2
	Gesture 10	80.1	44.3	35.8
Category C	Gesture 11	85.6	51.4	34.2
	Gesture 14	89.3	54.8	34.5
CategoryD	Gesture 18	91.0	69.6	21.4
	Gesture 19	92.1	64.8	27.3
Category E	Gesture 21	97.3	80.0	17.3
	Gesture 22	98.2	85.7	12.5

9 Conclusion

In this paper, the new quantitative evaluation technique of the mechanism of nonverbal communication which is especially paying attention to the characteristic gestures in the web communication was proposed. We analyzed the effect of the gestures using kinematic dynamical method by choosing the characteristic gestures carefully which were used for the purpose of exaggeration and emphasis in Kabuki, Disney's anime and the communication and presentation of attracting audiences and we obtained the following conclusions.

1. The exaggeration and emphasis degree of gesture has high correlation with the main joints torque value and it can be quantified by both inward and outward

rotational torque value and changing ratio of main joints such as shoulders, elbows and wrists of skeletal structure.

2. As for each characteristic gestures, it is possible to obtain the natural exaggeration gestures by correcting or replacing the torque values of natural motion withτ_{new} of the exaggeration ones. This calculation was conducted by using the **direct dynamical method**.
3. We performed to proof the **actuality** of those newly generated gestures based on the subjective evaluation, and the natural exaggerating gesture was generated with the result of subjective evaluation value was 80 percent or more near impression.

We can use these results for the wide range of the fields such that the development of user interface with 3D characters on the web with feeling **actuality** and the designing the natural and intelligible gesture interaction in real world.

References

1. Mehrabian, A.: Silent messages: Implicit communication of emotions and attitudes, 2nd edn. Wadsworth (1981)
2. Wagner, M., Mitchell, A., Nathan, J.: The role of gesture in instructional communication. In: 6th International Conference on Learning Sciences, pp. 35–43 (2004)
3. McNeill, D.: Gesture and thought. The University of Chicago Press (2005)
4. Reeves, B., Nass, C.: The media equation How people treat computers, television and new media like real people and places. University of Chicago Press (1996)
5. Shneiderman, B., Maes, P.: DEBATE: Direct manipulation vs interface agents. Interactions 4(6), 42–61 (1997)
6. Nishida, T., Jain, L., Faucher, C. (eds.): Modelling machine emotions for realizing intelligence: Foundations and applications. Smart Innovation, Systems and Technologies Series. Springer (2010)
7. Kudoh, S., Komura, T., Ikeuchi, K.: Stepping motion for a human-like character to maintain balance against large perturbations. In: IEEE International Conference on Robotics and Automation, ICRA2003 (2006)
8. Uno, Y., Kawato, M., Suzuki, R.: Formation and control of optimal trajectory in human multijoint arm movement-minimum torque-change model. Biological Cybernetics 61, 89–101 (1989)
9. Feltner, M., Dapena, J.: Dynamics of the shoulder and elbow joints of the throwing arm during a baseball pitch (1986)
10. Putnam, C.: Sequential motions of body segments in striking and throwing skills: descriptions and explanations (suppl. 1), 125–135 (2003)
11. Naka, T., Ishida, T.: Consideration of the effect of gesture exaggeration in web3D communication using 3DAgent. IEICE J96-D(8), 1925–1934 (2013) (in Japanese)
12. Bartenieff, I., Lewis, D.: Body Movement Coping with the environment. Gordon and Breach Publishers (1980)
13. Johnston, O., Thomas, F.: The Illusion of Life. Hyperion (1981)
14. Morris, D.: Gestures, Their origins and distribution. Stein and Day (1979)

Hand-Object Interaction: From Grasping to Using

Long Ni[1,2], Ye Liu[1,*], and Xiaolan Fu[1]

[1] State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
{nil,liuye,fuxl}@psych.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Evidence from psychology has shown that visual man-made manipulable objects can afford grasping actions even without the observers' intention to grasp them, and humans are able to use grasping information to recognize objects. But little is known if visual man-made objects, especially tools, can potentiate much more complex actions associated with using an object. In the present study, a priming paradigm was used to explore if passively viewing manipulable objects could be enough to activate specific action information about how to use them. The results showed that target objects with similar functional manipulation information to the prime objects were identified significantly faster than that with dissimilar manipulation knowledge to the prime objects. This is the first evidence by using behavioral study to indicate that just passively viewing a manipulable object is sufficient to activate its specific manipulation information that could facilitate object identification even without participants' intention to use them. The implications of manipulation knowledge in object affordances and object representation are discussed.

Keywords: Structural manipulation, Functional manipulation, Object recognition, Object affordances.

1 Introduction

How do humans interact with objects? One type of object-hand interaction is called structural manipulation (or volumetric manipulation), depending on online visual processing of objects' action-related properties such as object size or handle orientations [1 and 2]. Just imagine you grasp a cell-phone on your desk and move it from the left to the right. In order to implement these serious actions, you have to adjust your hand grip to the real size of the cell-phone and grasp it correctly. Another type of hand-object interaction that is more important in our daily life is functional manipulation, using the object with its function [1 and 2]. If you want to text a message with your phone, all you have to do is grasp it first and poke its keys or touch the screen. Though both types of object manipulations concern the ways we interact with objects [1, 2, and 3], learning how to use an object, especially a tool is of greater significance for individual development given the ubiquity of tool usage in human history.

[*] Corresponding Author.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 113–120, 2014.

The difference and dissociation between structural manipulation and functional manipulation have been supported by neuropsychological studies suggesting that there may exist two separate motor systems of hand-object interaction: dorso-dorsal pathway devoted to on-line translation of action-related properties of objects into motor program for reaching and grasping actions and a ventro-dorsal stream devoted to transforming object features into the appropriate object using action [1 and 2]. More importantly, psychological research has also shown that although both types of manipulation represent the possible interactions with objects, action associated with using an object seems more crucial to object representation [4].

But unfortunately, the bulk of psychological as well as human-computer interaction research aiming to examine hand-object interactions so far hasn't paid attention to functional actions, but only focused on simple and mechanical object grasping actions. For example, in robotics and automation field, robotic grasping has been an active research subject for decades, and a great deal of efforts has been spent on grasp synthesis algorithms to help robots grasp visually presented objects [5 and 6]. In addition, efforts have also been focused on object recognition using grasping cues [7 and 8] and grasp recognition by robots [9]. This is also the case in psychology. By far, examination of the interaction between perception and action has primarily centered on how visual features of an object can potentiate human's reaching and grasping actions toward the object. Psychological research has shown that visually presented graspable object can directly activate observer's structural action representation, which in turn influences both recognition [10] and grasping execution [11] toward that object.

These lines of evidence support Gibson's "object affordances" hypothesis [12 and 13] suggesting that humans perceive directly what tools afford in terms of meaningful actions, and visual objects can potentiate motor responses even in the absence of the observer's intention to implement an action. A growing body of evidence has already indicated that visual manipulable objects can automatically elicit action representation associated with grasping and moving an object without the observer's intention to act and even without their attention allocated to it [14 and 15]. For example, when participants were instructed to respond rapidly to the change in the prime objects' background color (either blue or yellow) by mimicking precision or power grip responses, they produced faster precision-grip responses to pinchable prime objects compared to the "graspable" ones, and faster power-grip responses when primed with graspable objects compared to pinchable objects, suggesting that the grip type of prime object irrelevant to the task affected participants' structural hand response (precision and power grasp) [16]. However, the potentiated action has been largely limited to structural manipulation associated with grasping and moving a manipulable object in terms of hand-object interaction.Therefore, little is known about if visual objects can also directly afford functional manipulation even when observer's attention is not allocated to the objects. Few psychological studies that touched on this issue provided inconsistent results. Evidence from neuroimaging studies showed that passively viewing a manipulable tool suffices to evoke its action information [17]. But due to the fact that most of the man-made objects in these experiments can be manipulated in both ways (e.g., we can structurally manipulate a calculator by grasping and moving it, and functionally manipulate it by poking its keys), we are not sure it is the structural manipulation or functional manipulation that leads to the activation of motor-related brains areas, including inferior parietal lobule, intraparietal sulcus and superior parietal lobule [17 and 18]. Brain imaging studies cannot help us to detangle the respective contribution of grasp-based action and function-based action. Several behavioral studies provided much more straightforward evidence suggesting that a manipulable object has to be processed to some degree before its functional manipulation information being evoked, and passively viewing the object is not enough to potentiate its functional manipulation information [19 and 20].

Based on the previous research, the goal of present study is twofold. We will examine: 1) if passively viewing a manipulable object is sufficient to activate its function-based action information; and 2) if so, is the function-based action information of a manipulable object able to affect its recognition? In order to address these issues, a priming paradigm modified from Helbig et al. [21] was used. Given the extensive experience we have interacted with common objects in terms of using them in our daily life, we hypothesized that function-based action knowledge could be a necessary component of object representation rather than a by-product of object processing. Therefore, it is predicted that passively viewing objects suffices to elicit their function-based actions.

2 Method

2.1 Participants

Participants consisted of a total of 16 undergraduate and graduate students (12 males and 4 females), ranging from 20 to 26 years of age (M = 20.1 years). All parti-cipants had normal or corrected-to-normal vision, and they were unaware of the purpose of the experiment.

2.2 Stimuli

We used 132 Gray-scale photographs of objects, including 86 man-made familiar manipulable objects and 46 animals, all of which were turned into a square of 280×280 pixels. Picture size on the screen was circa 9.7 cm by 9.7 cm, with a viewing distance about 85cm in order to keep the same visual angle about 6.5° with Helbig et al. [21]. All images were presented in the center of a 17-inch CRT computer monitor with a resolution of 1,024 by 768 pixels and a refresh rate of 80 Hz.

According to the functional manipulation actions of the prime objects and target objects, the experiment set up four conditions: congruent condition (the prime object and target object shared similar functional manipulation, e.g. a calculator and keyboard shared the same action of manipulation “poke” when using them); incongruent condition (functional manipulation of the prime object was different from that of target object, e.g., actions associated with using a keyboard and using an abacus were different); control condition (the prime objects were man-made object but hardly served a functional manipulation, e.g., tower) and unrelated condition (the prime objects were animals). Each of the four conditions contained 27 prime-target pairs with the same set of 27 target objects.

Because the participants were required to conduct object categorization task (judging if the target object is living or nonliving), we added another four filter conditions in each of which the same set of 27 living objects was used as target objects while the prime objects kept the same to the corresponding four experimental conditions. Therefore, the experiment was consisted of 27×4×2 trails in total.

In order to match object familiarity of the prime objects as well as visual similarity of prime-target pairs that might compound the expected functional manipulation congruency effect, we first asked 30 participants, none of whom took part in the experiment, to rate the familiarity and object manipulability of the original 184 objects pictures. Ratings were also obtained with regard to visual similarity and functional manipulation congruency between 196 pairs of prime and target objects, all of which were matched from the 184 objects. All the dimensions were rated on a five-point scale. We selected the final 27 prime-target pairs in each condition that repeated-measured ANOVA revealed no significant difference in the familiarity of the prime objects among the four critical conditions, but showed significant differences in functional manipulation congruency of the prime-target pairs among the four critical conditions, $F(3, 24) = 256.8$, $p < .001$. Post-hoc tests showed that functional manipulation congruency of prime-target objects in congruent condition (4.06) is much higher than that in incongruent (2.23, $p < .001$), control (1.25, $p < .001$) and unrelated conditions (1.09, $p < .001$). Although repeated measure ANOVA also revealed significant differences in visual similarity of the prime-target pairs among the four critical conditions, $F(3, 24)=55.16$, $p < .001$, Post hoc tests showed that the difference was attributed to higher visual similarity of prime-target pairs in congruent (3.25) and incongruent condition (3.12) than that in control (1.57, $p < .001$) and unrelated conditions (1.37, $p < .001$), while visual similarity between congruent and incongruent conditions revealed no difference.

2.3 Procedure

The experiment procedure was adapted from Helbig et al.'s research [21]. As schematized in Fig. 1, each trail started with a fixation point that remained 500ms on the center of the screen. After a blank white screen of 700ms, the prime object was present for 300ms and immediately replaced by another blank screen that was presented for 250ms. The blank screen was immediately followed by the target object that would not disappear until the response was made.

Different from the task in Helbig et al.'s research, the present experiment required participants to make object categorization task as quickly as possible without sacrificing accuracy. Object categorization responses were made by pressing A if the target object was man-made and L if it was a living object on the keyboard. All participants were right-handed and the dominant right hand was always used for responding to man-made targets. RTs were measured from the onset of the target objects.

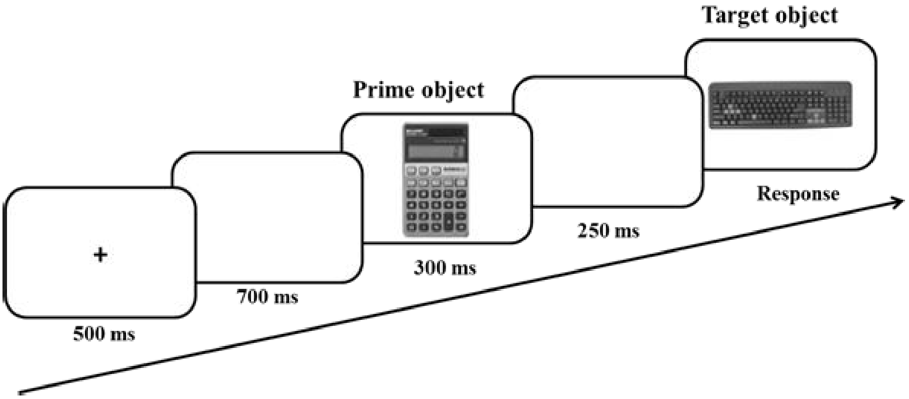


Fig. 1. Sequence of presentation in a typical trail for experiment

3 Results

All the data were analyzed using SPSS 17.0C (SPSS China). One participant was not included in the analysis because of his accuracy that is below three standard deviation of the mean. Response accuracy was not analyzed because it approached the ceiling that is higher than 99.7% in each condition. Reaction times more than 3 standard deviations above the mean were abandoned, as were trails with incorrect response. Totally, 1.54% trials were excluded. Because the primary goal of the present study is to examine if the action congruency effect would occur when the prime objects were merely passively viewed, therefore the data was analyzed with paired T-tests to directly compare participants' reaction times to target objects in congruent conditions with reaction times in incongruent conditions. Due to our prediction that the target objects in congruent condition would be identified much faster than in incongruent condition, one tailed paired t-tests was conducted.

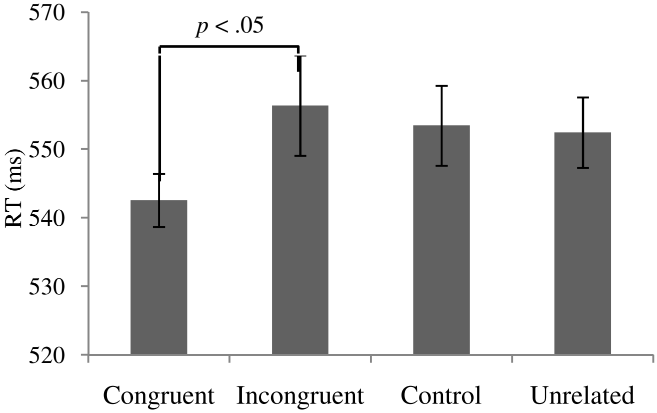


Fig. 2. Mean reaction times of target object classification in the four critical conditions

The response times for incongruent condition is the longest (RT = 556ms), while it is the shortest for congruent condition (RT = 543ms). Identification time for control condition (RT = 553ms) and unrelated condition (RT = 552ms) fell in between (Fig. 2). The results of the paired t-test showed that participants classified man-made target objects that shared with the prime objects similar functional manipulation significantly faster than target objects with different manipulation from the primes, $t(26) = 1.9$, $p < .05$. It also revealed faster responses for congruent condition as compared with the control condition, $t\ (26) = 1.73$, $p < .05$, as well as the unrelated condition, $t\ (26) = 1.8$, $p < .05$. No significant difference was found among incongruent condition, control condition and unrelated condition.

4 Discussion

Results of the experiment indicated that there was a reliable priming effect for functional manipulation. Specifically, when an object that afforded a specific functional action was primed by anther object with a similar functional manipulation, it would be processed and then identified more quickly. The action congruency effect occurred even when action-related information of the prime object was irrelevant to the experimental task and our participants had no intention to make any action response to the prime. More importantly, given the fact that the prime object per se was irrelevant to the categorization task, this result strongly demonstrated that passively viewing a manipulable object was sufficient to elicit its functional manipulation knowledge. Additionally, the action congruency effect couldn't be attributed to several potential variables due to the fact that both the visual similarity of the prime-target pairs in congruent and incongruent conditions and the familiarity of the primes among the four critical conditions were controlled in the experiment.

4.1 Hand-Object Interaction: From Grasping to Using

Since initially introduced by Gibson to explain how inherent "meanings" of objects in the environment can be directly perceived, and linked to the action possibilities offered to the agent [13], concept of object affordances has been developed by many other researchers and used in a variety of fields [see reviews, 22 and 23]. Though contemporary researchers still hold different views on affordances, most of them have been primarily focused on simple and mechanical actions associated with grasping objects, which relies on online processing of visual manipulation properties, such as object shape, size and orientation. As mentioned above, in robotics and automation field, scientists aim to create autonomous robotics not only capable of grasping a manipulable object, but also of categorizing and detecting objects according to their grasping affordances. However, our interaction with manipulable objects is not limited to simple and biomechanical reaching-out and grasping actions, it also involves complex functional manipulation that is more central to human's life dominated by tool use. Therefore, a more intelligent robot should be also capable of functionally manipulating a tool, and recognizing a tool according to its functional affordances.

The present study provided direct evidence for the notion that visual manipulable objects would also automatically afford human's action representation associated with using them, and the potentiated functional manipulation would in turn affect recognition of the objects. The extension of affordances in humans would be applied to robotics filed as well. New algorithms or methods would be explored to help robots functionally manipulate a visually presented tool or recognize the object by extracting its functional manipulation affordances.

4.2 Is Functional Action Knowledge A Key Part of Object Representation?

Though much evidence from behavioral , brain imaging and even neuropsychological research supported that action knowledge associated with an object's specific usage is an important component of the its representation, less of them are not without controversies [20]. Generally, we are not sure whether the activation of functional action information has a genuine causal role in object representation, or it is just a by-product of semantic or post-semantic object processing [for a review see 24]. Moreover, as for the issue if functional manipulation could be elicited under passively viewing condition, brain imaging research has not provided us a consistent picture. For example, while some studies have indicated that passive viewing of tools is sufficient to evoke a range of specific cortical activation associated with motor processes [17 and 18], others reported that brain regions specific to tools was evoked only when participants engaged in naming or object categorization task, not during passively viewing [25].

To our knowledge, the present study provided the first behavioral evidence that functional manipulation can be evoked under passive viewing condition. Due to the fact that participants in our experiment were neither required to attend to the prime objects nor biased to process action-related properties of the primes by asking them to make responses with prehensile actions, the action congruency effect we obtained excludes the possibility that the automatic activation of functional manipulation is an epiphenomenon of semantic or post-semantic object processing. On the contrary, it suggests that action representation associated with object use can pop out even when viewed passively. This result strongly supports the view that functional manipulation knowledge of man-made object is a necessary part of object representation.

Acknowledgments. This research was supported in part by grants from 973 Program (2011CB302201) and the National Natural Science Foundation of China (61375009).

References

1. Binkofski, F., Buxbaum, L.J.: Two Action Systems in the Human Brain. Brain and Language 127(2), 222–229 (2012)
2. Buxbaum, L.J., Kalenine, S.: Action Knowledge, Visuomotor Activation, and Embodiment in the Two Action Systems. Year in Cognitive Neuroscience 1191, 201–218 (2010)
3. Daprati, E., Sirigu, A.: How We Interact with Objects: Learning from Brain Lesions. Trends in Cognitive Sciences 10(6), 265–270 (2006)
4. Bub, D.N., Masson, M.E.J.: On the Dynamics of Action Representations Evoked by Names of Manipulable Objects. Journal of Experimental Psychology: General 141(3), 502–517 (2012)

5. Sahbani, A., El-Khoury, S., Bidaud, P.: An Overview of 3D Object Grasp Synthesis Algorithms. Robotics and Autonomous Systems 60(3), 326–336 (2012)
6. Gratal, X., Bohg, J., Björkman, M., Kragic, D.: Scene Representation and Object Grasping Using Active Vision. In: IROS 2010 Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics (October 2010)
7. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using Object Affordances to Improve Object Recognition. IEEE Transactions on Autonomous Mental Development 3(3), 207–215 (2011)
8. Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B.: Functional Object Class Detection Based on Learned Affordance Cues. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 435–444. Springer, Heidelberg (2008)
9. Kjellstrom, H., Romero, J., Kragic, D.: Visual Recognition of Grasps for Human-to-Robot Mapping. In Intelligent Robots and Systems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3192–3199 (September 2008)
10. Vainio, L., Symes, E., Ellis, R., Tucker, M., Ottoboni, G.: On the Relations between Action Planning, Object Identification, and Motor Representations of Observed Actions and Objects. Cognition 108(2), 444–465 (2008)
11. Tucker, M., Ellis, R.: Action Piming by Briefly Presented Objects. Acta Psychologica 116(2), 185–203 (2004)
12. Gibson, J.J.: The Senses Considered As Perceptual Systems. Houghton Mifflin, Boston (1966)
13. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
14. Symes, E., Ellis, R., Tucker, M.: Visual Object Affordances: Object Orientation. Acta Psychologica 124(2), 238–255 (2007)
15. Riggio, L., Iani, C., Gherri, E., Benatti, F., Rubichi, S., Nicoletti, R.: The Role of Attention in the Occurrence of the Affordance Effect. Acta Psychologica 127(2), 449–458 (2008)
16. Makris, S., Hadar, A.A., Yarrow, K.: Viewing Objects and Planning Actions: On the Potentiation of Grasping Behaviours by Visual Objects. Brain and Cognition 77(2), 257–264 (2011)
17. Chao, L.L., Martin, A.: Representation of Manipulable Man-made Objects in the Dorsal Stream. NeuroImage 12(4), 478–484 (2000)
18. Creem-Regehr, S.H., Lee, J.N.: Neural Representations of Graspable Objects: Are Tools Special? Cognitive Brain Research 22(3), 457–469 (2005)
19. Bub, D.N., Masson, M.E.J.: Gestural Knowledge Evoked by Objects As Part of Conceptual Representations. Aphasiology 20(9), 1112–1124 (2006)
20. 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)
21. Helbig, H.B., Graf, M., Kiefer, M.: The Role of Action Representations in Visual Object Recognition. Experimental Brain Research 174(2), 221–228 (2006)
22. Chemero, A.: An Outline of A Theory of Affordances. Ecological Psychology 15, 181–195 (2003)
23. Sahin, E., Çakmak, M., Dogar, M.R., Ugur, E., Üçoluk, G.: To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control. Adaptive Behavior 15, 447 (2007)
24. Mahon, B.Z., Caramazza, A.: A Critical Look at the Embodied Cognition Hypothesis and A New Proposal for Grounding Conceptual Content. Journal of Physiology-Paris 102(1-3), 59–70 (2008)
25. Devlin, J.T., Moore, C.J., Mummery, C.J., Gorno-Tempini, M.L., Phillips, J.A., Noppeney, U., Frackowiak, R.S.J., Friston, K.J., Price, C.J.: Anatomic Constraints on Cognitive Theories of Category Specificity. NeuroImage 15, 675–685 (2002)

Model-Based Multi-touch Gesture Interaction for Diagram Editors

Florian Niebling, Daniel Schropp, Romina Kühn, and Thomas Schlegel

Institute of Software- and Multimedia-Technology, Technische Universität Dresden,
Dresden, D-01062, Germany
{florian.niebling,thomas.schlegel,romina.kuehn}@tu-dresden.de,
d.schropp@gmx.de

Abstract. Many of todays software development processes include model-driven engineering techniques. They employ domain models, i.e. formal representations of knowledge about an application domain, to enable the automatic generation of parts of a software system. Tools supporting model-driven engineering for software development today are often desktop-based single user systems. In practice though, the design of components or larger systems often still is conducted on whiteboards or flip charts. Our work focuses on interaction techniques allowing for the development of gesture-based diagram editors that support teams in establishing domain models from a given meta-model during the development process. Users or groups of users are enabled to instantiate meta-models by free-hand or pen-based sketching of components on large multi-touch screens. In contrast to previous work, the description of multi-touch gestures is derived directly from the graphical model representing the data.

Keywords: Multi-touch gestures, model-based development.

1 Graphical Model-Driven Development

To allow for the graphical modeling of artifacts according to a given data model, graphical models can be used to represent features of the data model. These models contain shapes and containers providing a graphical description of data models and supporting the development of graphical diagram editors. One example of graphical modeling within the Eclipse framework is the Graphical Editing Framework (GEF) [13], which provides methods for the creation of graphical editors for the Eclipse Modeling Framework (EMF). The Graphiti Toolkit [7] based on GEF provides a graphical model for the representation of model instances, the Graphiti pictogram model. In our prototypical diagram editor, instances of a data-model can be created and manipulated by interacting with graphical representations specified using the Graphiti pictogram model, which are linked to the appropriate elements of the data model (see Figure 1).

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 121–130, 2014.

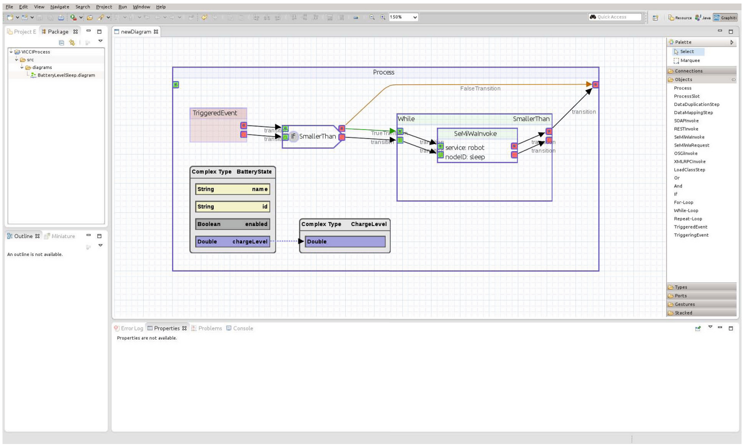


Fig. 1. Model-driven diagram editor based on Eclipse

1.1 Formal Representation of Gestures

Explicit methods for gesture recognition are based on patterns of strokes that are compared to the user input and evaluated regarding their similarity. To represent these strokes, previous developments use domain-specific languages to define multi-touch interaction such as the Gesture Description Language (GDL) [9] or the Gesture Markup Language (GML). In contrast, we propose to use the graphical representation of artifacts that is already present in the graphical model to derive multi-touch gestures. We identified three modes of gesture interaction that have been employed by users to sketch the various graphical items specified using the pictogram model: Single-touch / single-stroke, single-touch / multi-stroke, and multi-touch / multi-stroke (see Figure 2).

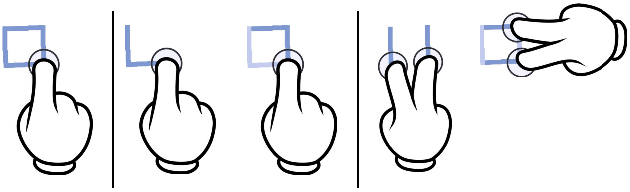


Fig. 2. (left) single-touch / single-stroke. (middle) single-touch / multi-stroke. (right) multi-touch / multi-stroke.

2 Related Work

Interaction with diagram editors was been simplified by enabling finger- or pen-based gestural sketching input. Plimmer et al. [12] give some overview about sketching tools developed for multiple application domains, such as UML modelling and UI generation, as well as on the gesture recognition algorithms that have been employed in the various works. Rubine's algorithm [14], a single-stroke pattern-matching algorithm, is one of the recognition techniques employed e.g. in InkKit [12], SUMLOW[5] and the Knight UML Designer [6]. Rubine's algorithm performs a comparison between features extracted from registered patterns and features extracted from user input. In the implementation in InkKit, the recognition process is started manually, while in SUMLOW, a timer is responsible for starting the recognition.

This is regarded as a drawback by Alvardo et al., who provide continuous recognition in their SketchREAD engine [1], using a gesture recognition algorithm based on bayesian networks.

In their SKETCH framework [15], Sangiorgi et al. employ a recognition algorithm based on the Levenshtein Distance [10], using string-based descriptions of gestures describing cardinal directions. The development effort on SKETCH seems to have ceased since 2010.

Scribble [16] is a GEF-based framework which allows for a seamless extension of GEF editors with gesture input. Since no pre-generated patterns are used, users are enabled to choose their own gestures, making the framework usable in many different application domains. The GEF editors that are augmented by Scribble have to be trained by the user to support their respective gestures of choice.

The related work shows the relevance of gesture-based input for diagram editors in multiple application domains. Multiple methods towards gesture recognition have been evaluated, with descriptions of gestures being either programmatic, pattern-based or feature-based. In contrast to the existing work, we propose methods for generating gesture description from the graphical models used to represent entities of the application domain.

3 Specification of Graphical Models

A *Graphical Model* contains graphical representations of the elements contained in a *Data Model* that represents concepts of the underlying application domain. In the context of workflow editing, a model-based graphical workflow editor contains a model of the workflow items (i.e. *activity*, *event*, *loop*, *connection*, etc.), and a graphical model containing graphical representations of these items (i.e. *rectangle*, *diamond shape*, *line*, etc.). By selecting graphical shapes in the editor, the user is enabled to instantiate concepts of the underlying data model.

In our prototypical application, we extended the graphical modeling framework Graphiti to allow for sketching of instances of the underlying *pictogram* model used by Graphiti. As can be seen in Figure 3, we extended Graphiti's *Diagram Editor* to make use of a *Gesture Recognizer*, that is able to detect shapes contained in the graphical model of the application. On detection of sketched fragments of the graphical model, instances of the underlying data model are created and the associated feature of the graphical model is added to the editor's scenegraph.

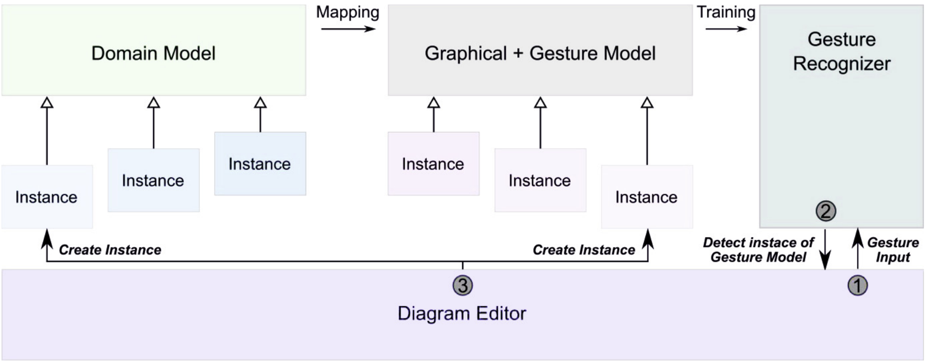


Fig. 3. Architecture of the prototypical diagram editing framework. Instances of the gesture model are recognized by a $N gesture recognizer, features of the data model and the graphical model are instantiated for display by the system.

3.1 Recognition of Sketched Graphical Models

To be able to recognize sketched user input, fragments of the graphical model, such as can be seen in Figure 5, have to be able to be detected by the *Gesture Recognizer*. We would like to give a short introduction about the methods that are employed in the *Gesture Recognizer* for sketch recognition, and for transformation of the *pictogram* model fragments to reference templates for the recognizer component.

The $1-Recognizer, a pattern-matching algorithm for single-stroke gestures, was introduced by Wobbrock et. al. [17]. It uses simple lists of coordinates as patterns for gesture recognition, which are compared to user input. The algorithm is implemented in four steps:

- Resampling. Because of different speed of user input, gestures contain different numbers of input points. In this step, the point path is resampled to contain a certain number of equidistant points, Wobbrock et. al. propose to use 64 points per point path.
- Rotation. Point paths are rotated in negative direction such that the *indicative angle*, the angle formed between the centroid of the gesture and the gestures first point, is 0.
- Scale and translation. After scaling the point path to a reference square, the centroid of the point path is translated to (0,0).
- Recognition. The point path is continuously rotated to find the minimum path-distance between the point path and supplied reference patterns.

The $1 recognizer's main benefits are simple implementation, high speed and that extensive training is unnecessary. This simplicity comes with several drawbacks. The algorithm is not able to distinguish input according to its orientation, aspect ratio or position, making it impossible to differentiate between e.g. squares and non-square rectangles. Also, $1 is not usable for multi-stroke gestures.

Listing 1.1. or.pictograms

```
<?xml version=”1.0” encoding=”UTF−8”?>
<pi:Diagram xmi:version=”2.0”
 xmlns:xmi=”http://www.omg.org/XMI”
     xmlns:xsi=”http://www.w3.org/2001/XMLSchema−instance”
     xmlns:al=”http://eclipse.org/graphiti/mm/algorithms”
     xmlns:pi=”http://eclipse.org/graphiti/mm/pictograms”
     visible=”true” active=”true” name=””>
  <children xsi:type=”pi:ContainerShape” visible=”true” active=”true”>
    <graphicsAlgorithm xsi:type=”al:Polyline” foreground=”//@colors.0”
        lineWidth=”2” width=”40” height=”40”>
      <points x=”0”  y=”20”/>
      <points x=”20” y=”0”/>
      <points x=”40” y=”20”/>
      <points x=”20” y=”40”/>
      <points x=”0”  y=”20”/>
    </graphicsAlgorithm>
    <children visible=”true”>
      <properties key=”gesture” value=”1”/>
      <graphicsAlgorithm xsi:type=”al:Ellipse”
          foreground=”//@colors.0” lineWidth=”3” filled=”false”
          width=”18” height=”18” x=”11” y=”11”/>
    </children>
    <children>
      <properties key=”gesture” value=”2”/>
      <graphicsAlgorithm xsi:type=”al:Polyline”
          foreground=”//@colors.0”>
        <points x=”0”   y=”0”/>
        <points x=”50”  y=”100”/>
        <points x=”100” y=”0”/>
      </graphicsAlgorithm>
    </children>
  </children>
  <colors red=”102” green=”102” blue=”255”/>
</pi:Diagram>
```

Fig. 4. Instance of Graphiti *pictograms* model representing the *or* data model instance of our prototypical workflow editor

Protractor [11] was developed to address some shortcomings of the $1 recognizer, the key difference being sensitivity to orientation, making it possible to distinguish eight base orientations. The $N recognizer [2] improves on the $1 algorithm, allowing for the representation of multi-stroke gestures as single-stroke gestures, combining the last point in a stroke with the first point of the following stroke. This allows for the recognition of a mixture of single- and multi-strokes. The $N Protractor [3] is a combination of the $N recognizer and the aforementioned Protractor algorithm.

The main reason for selecting the $N family of algorithms for our application framework is the simplicity of transforming fragments of the employed graphical model into patterns for the recognizer. The coordinate lists contained in the *al:Polyline* elements of the graphical model, as well as *al:Rectangle* and *al:Polygon* elements, as can be seen in Figure 5, can be easily converted into required coordinate lists for the $N recognizer. For shapes such as *al:Ellipse* and *al:RoundedRectangle*, we sample the shape to provide the appropriate coordinate lists.

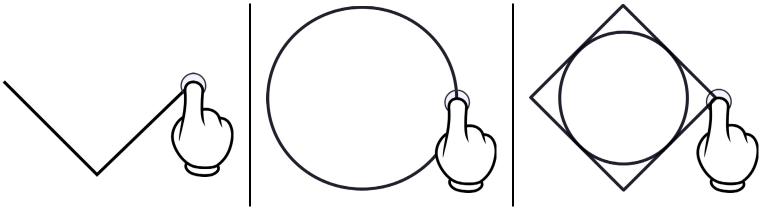


Fig. 5. Input possibilities for the *or* element: (left) mathematical symbol. (middle) circular gesture. (right) circular + diamond gesture.

3.2 Functionality of the Workflow Editor

We selected an existing EMF-based workflow editor to evaluate our prototypical gesture based sketching framework. The editor uses traditional mouse based interaction according to the WIMP concept. In addition to interactions performed using existing Eclipse interfaces, possible interactions with the editor are separated into two categories. As can be seen in Figure 1, the right side of the editor contains a list of workflow objects that can be dragged to the main diagram in the middle of the workspace, instantiating entities of the graphical model and placing them inside the diagram. Inside the diagram, existing workflow objects can be moved, deleted or connected using transitions between ports contained in workflow objects.

3.3 Integration of Gesture Recognition

The extended architecture of the workflow editor can be seen in Figure 6. Input, processing and detection of gesture based sketching are enabled inside Graphiti's *Diagram Editor* and our additional *Gesture Recognizer*. The link between Graphiti and gesture recognition is Graphiti's *Interaction Component*, where touch input is received and forwarded to the newly introduced gesture recognizer. Upon detection of one of the entities provided in the graphical model based on Graphiti's *pictogram model*, the *Diagram Type Provider* is used to instantiate the particular entities of the data model and the graphical model respectively.

4 Evaluation

The evaluation of our prototype was performed on a Dell XPS One 27 featuring a capacitive multi-touch display. The gestures that were automatically generated from the graphical model of our workflow editor were evaluated in a user study involving 15 participants. Furthermore, a comparison between the existing, traditional mouse-based interaction and gesture based usage was performed. This was done to gather evidence towards if expected advantages of gesture based sketching, such as higher intuitiveness, or expected disadvantages such as fat-finger problem or user fatigue, dominate the user experience. The evaluation

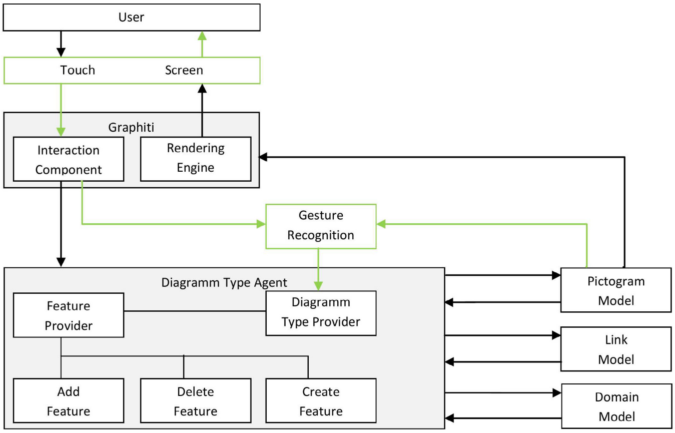


Fig. 6. Integration of gesture based sketching into the Graphiti architecture (adapted from Brand et al. [4])

was performed using two different quantitative methods, a formative user survey, and user transparent observation of behaviour. Participants in the study have not been involved in the development of the gesture recognition, although most of them belonged to the same faculty with similar background in software engineering, the application domain of the tasks in the user study.

To achieve the above mentioned goals, the following scenario was prepared. The participants were to sketch the graphical workflow elements *Process*, *If*, *Or*, *And*, *Loop* and *Ports*. *Ports* belonging to some of the elements were to be connected using *Transitions*. The workflow editor was to detect sketched workflow elements and position them at the respective position in the diagram.

Indicators that were rated were based on the NASA TLX evaluation [8] to assess cognitive and physical demands, overall effort, mental effort, physical effort, temporal effort, time pressure, performance and frustration levels. The observer that was monitoring the participants was mainly passive. Although the sequence of user tasks was arranged through the use of a survey sheet, the approach towards the solution of each task was presented to be open to the preferences of the user. The advances of the users were logged in the background and analyzed afterwards.

4.1 Evaluation Results

Following a short introduction of the evaluated categories are an evaluation of the most interesting results of the user study. Mental effort of gesture based sketching was perceived to be lower as traditional mouse interaction throughout the study. Further, a significant reduction in mental effort between the first and the following tasks leads to the impression that the method of interaction is learned after a very short period of familiarization, and can thus be characterized as intuitive.

Physical effort was perceived to be higher using gestures than using mouse interaction, a result that was to be expected since gesture interaction requires free movement of a stretched arm in mid-air in front of the display. Physical effort seems to be a fundamental weakness of gesture interaction, which is also seconded by results in the overall effort category. Temporal effort was also perceived to be higher for gesture interaction, even on tasks where measurements of the time requirements for mouse based and gesture based interaction where similar. Overall values for frustration where quite high when recognition of sketched objects failed. This happened mainly in the sketching of ports, with a recognition rate as low as 67.5%, where recognition rates for the other workflow elements reliably achieved between 90% and 100%. Further evaluation has since shown that the low rate of recognition of ports was due to problems with the positioning of the performed sketching. Multiple users have tried to sketch ports slightly on the outside of existing workflow objects, when ports were actually only added to workflow objects when the sketching was performed on the inside of an object, due to programming errors in the prototype.

Although multiple participants of the study voiced their discomfort with longer periods of gesture interaction on a desktop computer due to physical effort, overall evaluation has shown that users accepted the process of gesture based sketching of graphical representations as equal to mouse based interaction.

The decision to allow for multi-stroke gestures has to be reconsidered, as only two of the 15 participants made active usage of multi-stroke sketching for the *And* element, even after being explicitly advised towards the possibility of multi-stroke sketching.

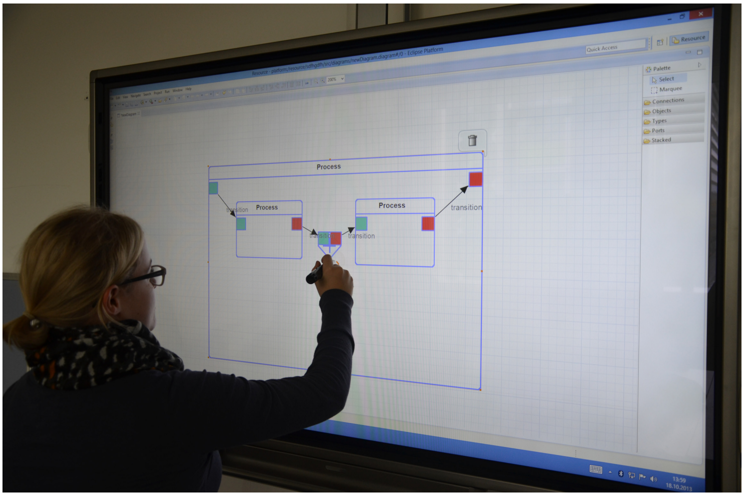


Fig. 7. Pen-input on interactive whiteboard

Multiple users intuitively reduced complex geometries to simpler gestures that represented subsets of the graphical representation of objects. E.g., the surrounding diamond of the graphical representations of the *And* and *Or* elements (see Figure 5) have been disregarded by most users, leaving just a simple *circle* gesture for the *Or* element and a *plus* gesture for the *And* element.

All participants but one have sketched transitions in a straight line between workflow objects, even when the final graphical representation of a transition was not a straight line to avoid cutting existing workflow elements.

5 Summary and Conclusion

We have evaluated a method for diagram sketching where gestures were automatically derived from the underlying graphical model of the application. A prototypical workflow editor based on Eclipse and Graphiti was augmented to support the generation of templates for a gesture recognizer from the Graphiti *pictogram* model. A formative user study was performed to evaluate user interaction with the modified editor.

As a fundamental difference towards previous work, the presented concept and prototypical implementation allows for collaborative multi-user interaction using multi-touch multi-stroke gestures. Evaluation with single user interaction on a desktop PC have shown that the sketching of workflows was accepted to be largely equivalent to mouse-based interaction concerning the preparation of workflow diagrams. Follow-up testing has shown tendencies that collaborative scenarios featuring digital whiteboards are promising targets for further user studies. Independent of the testing environment, our evaluations have shown that gestures derived from graphical models are accepted as input methods by users. The intuitive reduction of graphical representations by users towards simpler geometric subsets suggests further areas of research towards automatic generation of intuitive gestures from graphical models.

6 Future Work

Several different methods that support graphically similar objects need to be evaluated in future work. First, using context to allow the system to choose the item that is perceived to be of higher probability. Second, the support of similar objects using the same gestures, with additional pop-up menus allowing the user to choose one of the different objects. Third, further evaluation which parts of the graphical model are perceived by users to carry the most significance or relevance. Identifying parts that are perceived to be meaningful by users given a graphical representation is also necessary the more complex graphical models become.

Further work is also needed in the evaluation of introducing mobile multi-touch devices such as tablets into the software modeling process, expanding the collaborative user environment from single devices such as whiteboards to multiple devices.

References

1. Alvarado, C., Davis, R.: Sketchread: A multi-domain sketch recognition engine. In: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology, UIST 2004, pp. 23–32. ACM, New York (2004)
2. Anthony, L., Wobbrock, J.O.: A lightweight multistroke recognizer for user interface prototypes. In: Proceedings of Graphics Interface 2010, GI 2010, pp. 245–252. Canadian Information Processing Society, Toronto (2010)
3. Anthony, L., Wobbrock, J.O.: $N-protractor: A fast and accurate multistroke recognizer. In: Proceedings of Graphics Interface 2012, GI 2012, pp. 117–120. Canadian Information Processing Society, Toronto (2012)
4. Brand, C., Gorning, M., Kaiser, T., Pasch, J., Wenz, M.: Development of High-Quality Graphical Model Editors. Eclipse Magazine (2011)
5. Chen, Q., Grundy, J., Hosking, J.: An e-whiteboard application to support early design-stage sketching of uml diagrams. In: Proceedings of the 2003 IEEE Conference on Human-Centric Computing, pp. 219–226. IEEE CS Press (2003)
6. Damm, C.H., Hansen, K.M., Thomsen, M.: Tool support for cooperative object-oriented design: Gesture based modelling on an electronic whiteboard. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2000, pp. 518–525. ACM, New York (2000)
7. Fuhrmann, H.A.L.: On the Pragmatics of Graphical Modeling. Kiel Computer Science series. Books on Demand (2011)
8. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human Mental Workload 1(3), 139–183 (1988)
9. Khandkar, S.H., Maurer, F.: A language to define multi-touch interactions. In: ACM International Conference on Interactive Tabletops and Surfaces, ITS 2010, pp. 269–270. ACM, New York (2010)
10. Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10, 707 (1966)
11. Li, Y.: Protractor: A fast and accurate gesture recognizer. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 2169–2172. ACM, New York (2010)
12. Plimmer, B., Freeman, I.: A toolkit approach to sketched diagram recognition. In: Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1, BCS-HCI 2007, pp. 205–213. British Computer Society, Swinton (2007)
13. Rubel, D., Wren, J., Clayberg, E.: The Eclipse Graphical Editing Framework (GEF). Eclipse, Addison-Wesley (2011)
14. Rubine, D.: Specifying gestures by example. SIGGRAPH Comput. Graph. 25(4), 329–337 (1991)
15. Sangiorgi, U.B., Barbosa, S.D.J.: Sketch: Modeling using freehand drawing in eclipse graphical editors. In: Proceedings of the FlexiTools Workshop (May 2010)
16. Scharf, A.: Scribble - a framework for integrating intelligent input methods into graphical diagram editors. In: Software Engineering 2013 Workshopband (inkl. Doktorandensymposium), pp. 591–596 (February 2013)
17. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, UIST 2007, pp. 159–168. ACM, New York (2007)

Multi-sensor Finger Ring for Authentication Based on 3D Signatures

Mehran Roshandel[1], Aarti Munjal[2], Peyman Moghadam[3],
Shahin Tajik[1], and Hamed Ketabdar[4]

[1] Deutsche Telekom Innovations Laboratories, Ernst-Reuter Platz, 10587 Berlin Germany
{mehran.roshandel,shahin.tajik}@telekom.de

[2] Department of Biostatistics and Informatics, University of Colorado Denver,
13001 E 17th Pl Aurora CO 80045
aarti.munjal@ucdenver.edu

[3] Autonomous Systems, CSIRO Computational Informatics, 1Technology Court,
Pullenvale, QLD 4069
peyman.moghadam@csiro.au

[4] Quality and Usability Lab, TU Berlin Deutsche Telekom Innovation Laboratories
Ernst-Reuter-Platz 7, 10587 Berlin Germany
Hamed.Ketabdar@clickandbuy.com

Abstract. Traditional methods of authenticating a user, including password, a Personal Identification Number (PIN), or a more secure PIN entry method (A PIN entry method resilient against shoulder surfing [14]), can be stolen or accessed easily and, therefore, make the authentication unsecure. In this work, we present the usability of our multi-sensor based and standalone finger ring called Pingu in providing a highly secure access system. Specifically, Pingu allows users to make a 3D signature and record the temporal pattern of the signature via an advanced set of sensors. As a result, the user creates a 3D signature in air using his finger. Our approach has two main contributions: (1) Compared to other wearable devices, a finger ring is more socially acceptable, and (2) signatures created via a finger in the air or on a surface leaves no visible track and, thus, are extremely hard to forge. In other words, a 3D signature allows much higher flexibility in choosing a safe signature. Our experiment shows that the proposed hardware and methodology could result in a very high level of user authentication/identification performance.

Keywords: uman Computer Interaction (HCI), Touch less gestural interaction, Wearable device, Finger ring.

1 Introduction

Due to increased capability of a smartphone, users tend to store all of their personal information in their mobile devices. Smart technology, however, raises a serious threat to a user's credentials, unless the access to these devices is secured by information unique to each user. As an example, access to a user's smartphone may lead to his bank account, social security number, email accounts, or other personal information.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 131–138, 2014.

Traditional methods used for authenticating a login include entering a password, Personal Identification Number (PIN). Previous research shows that it's not difficult to replicate this information, thereby making it insecure. A more robust solution will be to provide users with a unique way of interaction with their computing device. While a modern computing device easily fits a human hand, our world of interaction is not limited by the size of the device. With this motivation, we have developed a multi-sensor based framework called Pingu [1] that helps a user perform gestural signatures to access his computing device (e.g. smartphone). Pingu is calibrated in the form of a miniature, wearable finger ring that can perform sharp and tiny gestures. When the user performs his signature as a general gesture, sensor readings specific to each sensor are recorded, even if the device is not in the vicinity of the user. These sensor readings define the 3D trajectory of the ring and, therefore, are unique to each individual.

With wireless connectivity, feedback mechanism, and an advanced set of sensors, Pingu offers a wide range of applications. In addition, unlike previously proposed wearable devices (such as gloves, wristwatch [2, 6]), Pingu is a standalone device that does not require any extra hardware for interaction with a computing device and is also socially wearable. In this work, we explore the usability of Pingu in providing a secure authentication method for users to access their computing devices. To illustrate further, we conducted a user study with 24 participants, where each participant performs his signature in the form of a gesture and the sensor readings specific to the gesture are recorded. We show that the recorded sensor readings provide rich information specific to each gesture made by a user and with simple classification algorithms, the users can be authenticated based on their signatures with very high accuracy.

The rest of this paper is organized as follows. In Section 2, we review the related potential solutions for generating 3D signatures. Then in Section 3, we explain the architecture of the Pingu's hardware. In Experiments Section, the data collection and feature extraction via Pingu are explained. Section 5 presents our results of signature classification via different machine learning algorithms. In Section 6, further classification based on correlation and frequency features is illustrated. Finally, we conclude the paper in the Section 7.

2 Related Works

In recent years, different gestural recognition approaches are developed which are either used to generate 3D signatures such as MagiSign [5, 16], or can potentially be used to generate a 3D signature [2, 3, 4, 6, 7, 15].

In our previous work, MagiSign [5, 16], a 3D signature is created via influencing the magnetic field of a magnetic (compass) sensor embedded in mobile devices. However, the space of interaction is limited to the immediate 3D space around the device. Moreover, while Pingu can work with any computing device, MagiSign works only with smartphones (e.g., an iPhone). Finally, using multiple sensors in Pingu leads to a more precise gesture recognition in comparison to only one magnetic sensor in MagiSign.

In other approaches, which can potentially be used for 3D signature such as Acceleration Sensing Glove [2], a user has to wear additional gloves to interact with the computing device. The disadvantage of this approach is that they can be socially unacceptable or obtrusive. Other frameworks, such as Gesture Pendant [3] and SixthSense [4], require users to wear pendant and additional hat, respectively, which suffer from the same problems. Moreover, in approaches like SixthSense and Gesture Pendant, there is a need for an optical sensor (e.g., camera) which causes problem when performed gestures are not in the direct line of sight of the sensor.

Finger rings or wristwatches can be used to solve the problem of social awkwardness. Pinchwatch [6] uses a wristwatch for finger gesture recognition with the help of a camera. By performing sliding and dialing motions, some functions are invoked. However it still has the occlusion problem. More recently, Nenya [7], a magnetically-tracked finger ring, is developed which includes a permanent magnet in the form of a finger ring and worn-watch wireless tracking bracelet. The magnetometer is used to track the ring's position and a Bluetooth radio allows the bracelet to send ring input to the user's device. However, Nenya only supports 1D input in comparison to 3D inputs supported by Pingu. Furthermore, the IR Ring provides an innovative method which can be used for Authenticating users' touches on a multi touch display [13].

3 Design

Figure 1 shows the prototype for Pingu. Specifically, Pingu has four sensors: a tri-axis accelerometer, a tri-axis gyroscope, a tri-axis magnetometer, and proximity sensing plates with two channels. The accelerometer is used to detect the orientation and motion of the device along x, y, and z axes. A tri-axel gyroscope detects the angular rate of movement of the ring along the three axes x, y, and z. The deformation of magnetic fields is useful in recognizing coarse gestures made around the device. In addition, the proximity-sensing plates allow sensing the proximity of other fingers. The feature set obtained from one or more sensors can then be combined to form a feature vector specific to each gesture. Based on the movement of the ring, each of these sensors provides a feature set. Table 1 lists the details specific to three sensors and radio technology used in the design of Pingu.

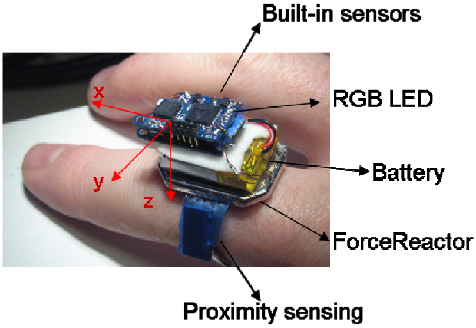


Fig. 1. Prototype for our multi-sensor based framework called Pingu

Table 1. Sensors used in the design of Pingu with their specifications.

Sensor	Description
Accelerometer	[-8g, 8g]
Magnetometer	[-2gauss, 2gauss]
Gyroscope	[-2000deg/s, 2000deg/s]
Bluetooth	Up to 2m

4 Experiment

To evaluate the usability of Pingu in secure authentication, we perform gestures defined as a signature. Since Pingu is worn on a finger, even sharp and tiny gestures can be used for the purpose of authentication. When the user performs a gesture, the associated sensor data is collected. The sensor readings define the temporal pattern of the signature and, thus, can be used in matching the signature for authentication. Our experiments were split into two categories:

1. Signature in the air and
2. Signature on the table

Setting the two medium of air and desk provides a variety of surfaces for gesturing. In this way, the methodology can be tested under more variable yet practical scenarios. The desk medium is a surface which is commonly available for users during the gesturing process. The air medium also provides the fantasy of writing in air for the user, when the two other mediums are not available.

Signatures for each user are recorded on two different mediums to evaluate Pingu for its dynamic usability. In other words, these two experiments ensure that the usability of Pingu in secure authentication is irrespective of the surface (or medium) of interaction. Each signature is first performed in the air and then on the table. Multiple templates per signature are collected. Specifically, when a signature is performed, the 3D trajectory of the ring is recorded in the form of sensor readings. For example, as the ring moves, the accelerometer, embedded in Pingu, measures the linear acceleration along three axes: x, y, and z.

Since Pingu performs sharp and tiny gestures, any general gesture can be used as a signature pattern. When a user performs a gesture, the sensor readings specific to the gesture are compared to the previously recorded signature pattern (template) for the user. The two patterns can be compared via Dynamic Time Warping (DTW) technique and if the difference between the two patterns is less than a pre-defined threshold, the signature is accepted. Next, we provide details on the datasets and the classifiers used to analyze signatures made by the users.

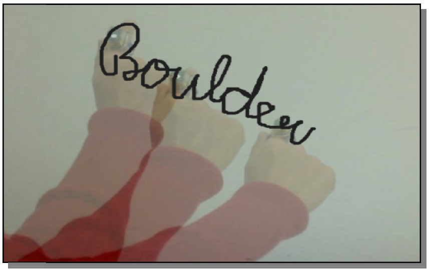


Fig. 2. An example of a 3D signature made in the air

4.1 Data Collection

Our dataset consists of six signatures, obtained from 24 users. Every user performs each of these six signatures 15 times. The sensor readings specific to each signature are captured via a Java desktop application. To classify the signatures based on the sensor readings captured, we extract an extended set of features, specific to every sensor reading captured for each signature performed by a user. To extract feature vector from the sensor readings, we use the following approach.

4.2 Feature Extraction

We mixed the data collected from all the 24 users and cross-validated. For this purpose, we formed a feature vector containing the data specific to each sensor. For example, the feature vector specific to accelerometer contains the following:

1. Mean and variance of the linear acceleration along x, y, and z axes (6 features),
2. Mean and variance of the Euclidian norm of the linear acceleration along x, y, and z axes (2 feature),

The feature vector from gyroscope is obtained in a similar manner. Feature vector for each sensor, therefore, contains 8 elements. Since multiple windows provide more detailed information in gesture classification, our results are based on 4 windows. Feature vectors obtained from each window are concatenated to form a new feature vector of 32 (=8x4) features.

5 Signature Classification

The feature vectors obtained for each sensor are then concatenated to form a large feature set that represents the features defining each signature. To classify users based on their signatures, we use a set of four classifiers: (a) Decision Tree (DT), a decision tool that uses graphs and model of decisions to derive the outcomes and consequences, (b) Multi-Layer Perceptron (MLP), a feedforward artificial neural network that models the relationship of inputs and outputs to find the patterns, (c) Naïve Bayes (NB), a probabilistic classifier that uses Bayes' theorem with strong independence assumptions, and (d) Support Vector Machines (SVM), which set hyperplanes in high dimensional space for using classification and regression. The current implementations available for these classifiers in Weka machine learning toolkit version 3.7.0 [11, 12] on Mac OS X are used. Tables 2-3 list the classification accuracy obtained for both sets of experiments. As shown, MLP and SVM outperform the other two classifiers (i.e., DT and NB). In addition, we note that using simple features (i.e., mean and variance of sensor readings) can enable us to classify users (based on their signature patterns) with an accuracy of about 99% in both experiment categories.

Table 2. Signature Classification for 24 Users in Signature in the air

Classifier	Accuracy
MLP	98.8889%
DT	82.2222%
NB	97.5%
SVM	99.1667%

Table 3. Signature Classification for 24 Users in Signature on the table

Classifier	Accuracy
MLP	99.1549%
DT	87.0423%
NB	97.4648%
SVM	99.4366%

6 Correlation and Energy Features

To illustrate the effect of a feature set on the accuracy of classification techniques, we extract piecewise correlation and frequency features of sensor readings. Frequency features measure the intensity in the movement of ring and are calculated as the sum of squared discrete FFT magnitudes. The correlation features, on the other hand, help differentiate between sharp and tiny gestures made by users. Together, these features help capture the periodicity in sensor readings. Thus, we performed another study of classifyingsignatures with a feature set that contains frequency and correlation features in addition to mean and variance extracted from each of the three sensors. Specifically,

1. Piecewise correlation between linear acceleration along x, y, and z axes (3 features), and
2. Frequency domain features along x, y, and z axes (3 features).

Feature vector for each sensor, therefore, contains 14 elements. With a window size of 4, the size of the feature set is 56 (=14x4). To classify, we again use the four classifiers listed earlier. Tables 4-5 present our results obtained for the experiments performed in air and on the table. The results indicate that with correlation and frequency features, the accuracy can be excelled to 100%.

Table 4. Signature Classification for 24 Users in Signature in the air (with Correlation and Frequency features)

Classifier	Accuracy
MLP	100%
DT	86.6667%
NB	98.3333%
SVM	100%

Table 5. Signature Classification for 24 Users in Signature on the table (with Correlation and Frequency features)

Classifier	Accuracy
MLP	100%
DT	86.6667%
NB	99.1549%
SVM	99.7183%

7 Conclusions

In this work, we have proposed a unique secure authentication solution and presented our results for this system using a standalone, miniature, and wearable finger ring called Pingu. Pingu is a socially wearable, small finger ring that is equipped with multiple sensors to provide rich information about the signatures made by a user. Our analysis for signature recognition is based on a large dataset of 24 users and we have shown that with simple classification algorithms, the signature performed by a user can be recognized with a very high accuracy. Therefore it can be a trustworthy authentication solution for many applications.

References

1. Ketabdar, H., Moghadam, P., Roshandel, M.: Pingu: A new miniature wearable device for ubiquitous computing environments. In: 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), IEEE (2012)

2. Perng, J.K., Fisher, B., Hollar, S., Pister, K.S.J.: Acceleration sensing glove (ASG). In: The Third International Symposium on Wearable Computers (ISWC 1999), pp. 178–180 (1999)
3. Starner, T., et al.: The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In: The Fourth International Symposium on Wearable Computers. IEEE (2000)
4. Mistry, P., Maes, P.: SixthSense: a wearable gestural interface. In: ACM SIGGRAPH ASIA 2009 Sketches. ACM (2009)
5. Ketabdar, H., Moghadam, P., Naderi, B., Roshandel, M.: Magnetic signatures in air for mobile devices. In: Mobile HCI 2012, pp. 185–188 (2012)
6. Loclair, C., Gustafson, S., Baudisch, P.: PinchWatch: a wearable device for one-handed microinteractions. In: Proc. MobileHCI (2010)
7. Ashbrook, D., Baudisch, P., White, S.: Nenya: subtle and eyes-free mobile input with a magnetically-tracked finger ring. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems. ACM (2011)
8. Kratz, S., Rohs, M.: HoverFlow: expanding the design space of around-device interaction. In: Proc. of the 11th International Conference on Human Interaction with Mobile Devices and Services, Bonn, Germany, pp. 1–8 (2009)
9. Butler, A., Izadi, S., Hodges, S.: SideSight: multi- "touch" interaction around small devices. In: Proc. UIST, pp. 201–204 (2008)
10. Kim, J., He, J., Lyons, K., Starner, T.: The Gesture Watch: a wireless contact-free gesture based wrist interface. In: Proc. ISWC, pp. 15–22 (2007)
11. Witten, H.I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999)
12. http://www.cs.waikato.ac.nz/ml/weka/
13. Ring, T.I., Roth, V., Schmidt, P., Güldenring, B.: Authenticating users' touches on a multitouch display. In: Proc. UIST (2010)
14. Roth, V., Richter, K., Freidinger, R.: A PIN entry method resilient against shoulder surfing. In: Proc. 11th ACM Conference on Computer and Communications Security, Washington, DC, USA (2004)
15. Ketabdar, H., Abolhassani, A.H., Roshandel, M.: MagiThings: Gestural Interaction with Mobile Devices Based on Using Embedded Compass (Magnetic Field) Sensor. IJMHCI 5(3), 23–41 (2013)
16. Ketabdar, H., Moghadam, P., Naderi, B., Roshandel, M.: Magnetic signatures in air for mobile devices. In: Mobile HCI 2012, pp. 185–188 (2012)

What You Draw Is What You Search: The Analog Gesture

Benoit Rouxel[1], Franck Poirier[2], Jean-Yves Antoine[3],
and Gilles Coppin[1]

[1] Lab-STICC, Telecom Bretagne CS 83818, 29238 Brest, France
{benoit.rouxel,gilles.coppin}@telecom-bretagne.eu
[2] Lab-STICC, Université de Bretagne-Sud 56000 Vannes, France
franck.poirier@univ-ubs.fr
[3] Université François Rabelais de Tours, LI, 3 place Jean Jaurès, 41000 Blois, France
jean-yves.antoine@univ-tours.fr

Abstract. This paper presents a new type of gesture for identifying spatio-temporal patterns: the analog gesture. Analog gestures can be characterized by some features (speed, acceleration, direction, and angle) which describe the dynamic morphology of the gesture. At first, we detail interactive tasks that should benefit for the use of analog gestures. Then we give a state of the art concerning gesture recognition and investigate the specificity and the main properties of the analog gesture. Then, we propose a review of the surveillance maritime system called Hyperion which uses analog gestures. Finally, we give an example of the use of this type of gesture by the operator. It concerns the interactive detection of ship abnormal trajectories in the context of maritime surveillance.

Keywords: Gesture recognition, time-space pattern search, tabletop computing.

1 Introduction

Gestural interfaces are increasingly present in our daily lives. Nowadays, many different gestures have been investigated to enrich interaction. It is possible to use 3D-gestures to control characters in video games or 2D-gestures to make a call with a smartphone. Gestures may be associated with different commands; for instance, symbol drawing can be used as a shortcut for calling software functions [1].

The recognized gestures refer to an action, a symbol or an idea, which refer themselves to software functions. However, to the best of our knowledge, there is no attempt in the literature [6,7,8,12] to consider gestures which can directly refer to a time-space pattern of reference. We call time-space pattern (TSP) a series of positions that takes into account space and time simultaneously. We consider positions composed of a location and a timestamp. TSPs are useful in many domains, such as the behavioral analysis of pedestrians in a crowd, optical character recognition, and more generally any study that aims at spatio-temporal clustering and classification. In this paper, we considered the application domain of maritime surveillance, where TSPs

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 139–147, 2014.

correspond to boat trajectories that take into consideration the boats speed. Schematic examples of such trajectories are given in figure 1. The TSP on the left corresponds to a uniform speed while the right corresponds to a deceleration.

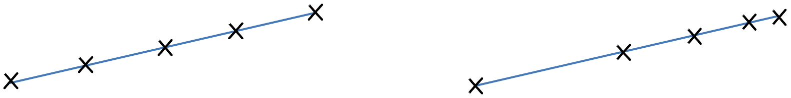


Fig. 1. TSP with uniform speed (left) and deceleration (right). All positions are sampled with the same frequency.

Our proposal consists in investigating the potential uses a new gesture type called analog gesture (AG) which is based on the explicit specification of a TSP. An analog gesture is a gesture dedicated to the expression of spatial and temporal features of a trajectory (or by extension a shape). The gesture is called "analog", because we expect lengths, orientations and speeds to be proportional or representative of the real trajectory, which is expressed through a TSP.

The paper is organized as follows. Section 2 presents some works related to our problematic. To the best of our knowledge, gestures have not been used so far to express of all the features describing a TSP. This is the aim of the analog gesture, which is defined in section 3. We then described the Hyperion platform, a maritime surveillance application where analog gesture is useful. Finally, we focus on the integration of the analog gesture in Hyperion before a final conclusion.

2 Related Work

Previous works in gesture recognition mainly focus on path recognition. *$1 Gesture Recognizer* [14] is a 4-step algorithm that recognize a predefinite gesture extracted from a finite alphabet of unistroke gestures. The algorithm was later extended ($-family) to enable multistroke gesture recognition [13]. Other algorithms as like in *Octopocus* [2] or the turning angle algorithm [5] use template alphabets to recognize gestures. Though originally applied on images, the turning angle algorithm can be applied on gesture recognition.

Contrary to the aforementioned algorithms, *PaleoSketch* [10] does not use any alphabet. This application improves free hand draws by replacing parts of the sketch with ideal shapes. For example, it replaces a round sketch with a circle and a line sketch with a straight line. When a new shape is drawn, a corner finding algorithm produces a polyline interpretation that closely fits the original shape. After that, each subpart of the computed polyline is analyzed and replaced with the closest simple shape. The simple shape library is composed of several shapes, like line, arc, circle and ellipse. This approach does not use templates to recognize a big shape; the shape is seen as an addition of simple shapes. A very large number of shapes can thus be represented with gestures as long as those shapes can be decomposed in simple shapes. Nevertheless, all those works only focus on the shape. They don't take into account the speed of the gesture which was used to draw the shape.

Holz and al. [4] takes into account the speed of the gesture in a selection technique they proposed for querying time-series graphs. To select a part of the graph, users sketch over part of the graph, establishing the level of similarity through the speed at which they sketch. Whereas this technique uses a temporal parameter (speed of sketch), it does not allow the expression of free shapes. To search a specific shape, a similar one has to exist.

Rubine's algorithm [11] classifies gestures according to 13 criteria like the length and the angle of the bounding box diagonal, the maximum speed and the duration of the gesture. This algorithm requires an initial training by drawing sample gestures. Even though time is taken into accounts in gesture classification with maximum speed and duration criteria, those two time parameters are too few to express acceleration in a TSP for example. In addition, the template alphabet required by this technique does not allow the expression of all spatial characteristics of a TSP.

While some works focus on the spatial dimension, others use template alphabets which reduce the number of recognized path to the size of the alphabet.

3 The Analog Gesture

The analog gesture is a gesture dedicated to the expression of spatial and temporal characteristics of a trajectory. This gesture taken as a whole is devoted to be decomposed into a series of segments. Spatial characteristics will mostly correspond to the lengths of the segments as well as their orientations; while temporal the features are expressed via the speed or acceleration within the segment.

Table 1. Parameters used to characterize a trajectory

<table>
<tr><th>Dimension</th><th colspan="2">Parameter</th><th>Recognized feature</th></tr>
<tr><td rowspan="6">Space</td><td colspan="2">Number of dimension</td><td>2D, 3D</td></tr>
<tr><td rowspan="6">Shape of the drawing</td><td>Angle</td><td>Each 15°-interval</td></tr>
<tr><td>Direction</td><td>Continuous</td></tr>
<tr><td>Spatial inking</td><td>Previously defined area</td></tr>
<tr><td>Path</td><td>Sequence of remarkable object in the environment crossed by the drawing</td></tr>
<tr><td rowspan="2">Orientation</td><td rowspan="2">A direction or the opposite direction</td></tr>
<tr><td rowspan="3">Time</td></tr>
<tr><td colspan="2">Speed</td><td>Zero, slow, medium, fast</td></tr>
<tr><td colspan="2">Acceleration</td><td>Strong deceleration, deceleration, acceleration, strong acceleration</td></tr>
<tr><td>Force</td><td colspan="2">Variation of pressure</td><td>Same pressure, increasing pressure, decreasing pressure</td></tr>
</table>

The speed and variation of pressure are important when we produce the gesture. Table 1 presents the three dimensions of the gesture realization were used to characterize the gesture and therefore the intended TSP. These three dimensions are space, time and force. For each dimension, one or more parameters are recognized.

For each parameter, the assigned value is either continuous or discrete according to human abilities. Since human can draw a direction with a fairly good accuracy, therefore this is a continuous parameter. On the contrary, people are unable to draw a TSP with an accurate speed, so this parameter can take only few values (zero, slow, normal and fast), and therefore is a discrete parameter.

Analog gestures are multi-touch gestures whenever the objective is to express relative evolutions of multiples trajectories. For instance, in the maritime surveillance domain, if we want to indicate that two vessels are sailing close together on near parallel courses (this situation occurs in boarding situation), we have to use two fingers.

Analog gestures allow in one hit, to express simultaneously various parameters of a segment (length, orientation, speed) as well as complex objects composed of chained segments.

To prove the utility of this type of gesture and how it works in real situation, we will show how we integrated AG in an effective application of maritime surveillance called Hyperion. In the two next parts, we will present the Hyperion platform and after that, we will expose how we use the gesture in this platform.

4 Hyperion Platform

VTS are control centers from which the maritime traffic is monitored. They aim at improving the safety and the organization of the traffic and at protecting the environment. The VTS controllers deal with many different types of information at the same time (AIS, radar, weather …). This considerable amount of raw information involves a heavy cognitive load which reduces the efficiency of the operator.

We propose to develop a domain-specific maritime surveillance system (Hyperion) to reduce cognitive load through a process of computer-aided decision-making. Hyperion is an application dedicated to help vessel traffic service (VTS) controllers. It is developed in Java on Diamond Touch DT107 a touch table.

The main aim of Hyperion platform is to highlight abnormal behaviors of moving vessels. The abnormal behaviors are defined by rules. These rules can contain static properties, a behavior (trajectory) and an anchorage (restriction of a rule to a specific area). Since there rules are strongly related to TSP patterns of sailing behaviour, such behaviours are defined by the controller (expert) using AG.

In order to detect abnormal behaviors of moving vessels in a maritime area, we propose to combine a bottom-up and a top-down approach. The analysis of how surveillance operators work [9] has shown that they were more looking for abnormal trajectories than checking normal ones. This is why we propose in a rule based expert system devoted to the maritime traffic analysis, to focus on the detection of abnormal behaviors.

The top-down approach allows the operator to define a rule characterizing what he/she considers an abnormal situation in a given area. These rules work like a filtering function. Any vessel matching the rules is highlighted (fig. 2).

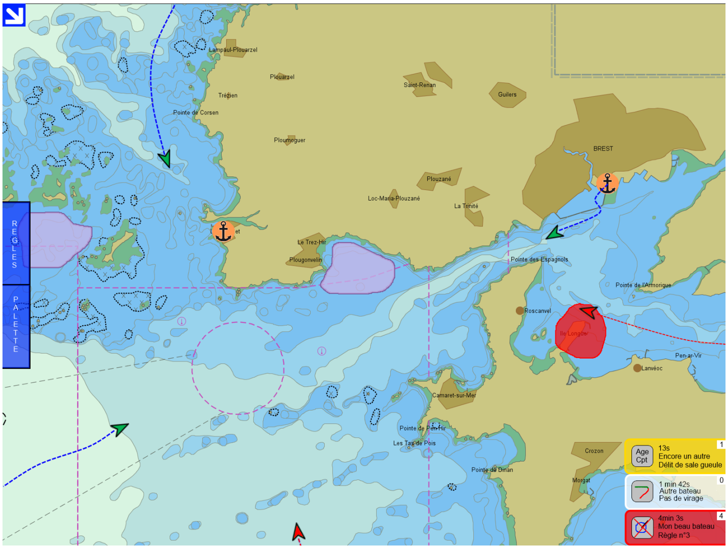
The bottom-up approach restricts the identification of abnormal behaviors to predefined but well established rules. For example, if a vessel breaks a rule of navigation, it must be reported to the operator. This type of detection is robust in trivial situations. Without these predefined rules, the operator would have to define a larger number of rules for simple situations. Subsequently, he would not be focusing on the detection of abnormal behaviors.

Hyperion works on 2 main modes: an operational mode and a rule edition mode. When the application starts, the operational mode is on. In this mode, a ship breaking a rule is highlighted, and the user has to check the alert report and possibly report a false positive recognition. In rule edition mode, the user can create, search and delete rules, and apply them on map elements (vessels, harbors and areas).

4.1 Operational Mode

Figure 2 shows the interface in operational mode: a map represents the current situation. Vessels, harbors and areas appear on the map while alert reports are displayed on right of the map. Rules can be applied on every map elements. Moreover, the areas can be created, modified and deleted by the operator.

When an alert occurs, an alert box which give an overview of the alert goes down from top to bottom right of the screen while, an animation catches the operator attention on the boat triggering the alert. Until alert is treated, the alert box stays in the stack of alerts. Alert box presents the icon of the rule, the time since the alert, the boat name, the rule name, and a number and a color that corresponds to the rule priority. If a vessel breaks a rule, it takes the color of the rule priority or a color corresponding to the sum of broken rules priorities, if more than one rule is broken.



1
13s
Encore un autre
Délit de sale gueule

0
1 min 42s
Autre bateau
Pas de virage

4
4min 3s
Mon beau bateau
Règle n°3

Fig. 2. Hyperion GUI in operational mode

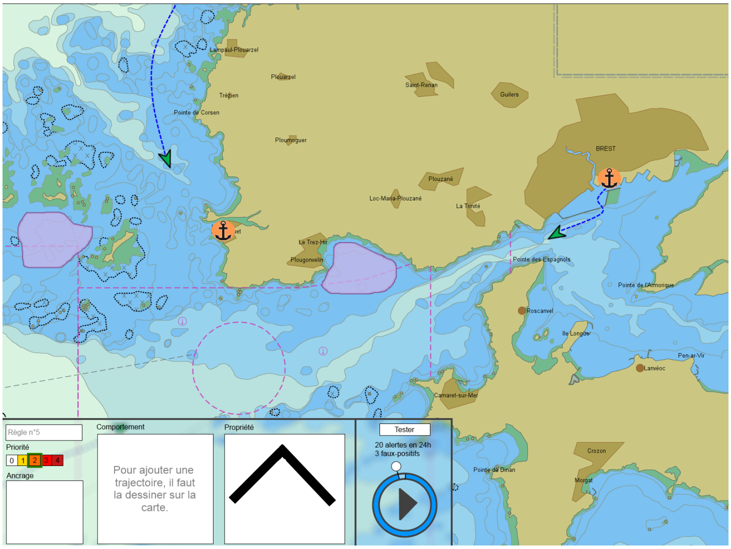
4.2 Rule Edition Mode

To add a new rule on map elements, the user has to tap and hold on an element. A circular menu appears and the operator can choose the create rule item. If the tap and hold gesture is performed directly on the map (not on a map element) the created rule will apply on the entire area monitored by the VTS. The user can also apply an existing rule on a map element. In this case, the user has to select an existing rule and apply it on a map element.

The set of applied rules can be modified in rule edition mode. In Hyperion platform, a rule is composed of 5 properties: a rule name, a priority, an anchorage, a dynamic behavior (trajectory) and a set of static properties like the boat name or its size.

Figure 3 shows Hyperion GUI when the user is creating a new rule. In this situation, the map and its elements are displayed in the background and two new boxes appear in the bottom left corner. The first one is a detailed view of the currently rule edited and the second one allows the testing of the rule.

In the rule box, each part of the rule is visible. First, the text area in the top left corner allows the showing and editing of the rule name. First, a default name "rule n°--" is given to the rule. Below, the slider corresponds to the rule priority that allows the user to order alert treatment by setting the colors and the numbers to the Vigipirate code (a French alert state). The anchorage area shows the "anchorage" property of the rule. The box in the middle ("behavior") shows the abnormal trajectory expressed by the rule. These two properties are expressed using analog gestures performed directly on the map. Finally, the last box shows the static rule properties. For example, if we want to express a rule forbidding a military ship to enter an area, we have to add the military property. The static property set can be accessed by performing an up swipe on the property box.



Règle n°5

Priorité

0 1 2 3 4

Ancrage

Comportement

Pour ajouter une trajectoire, il faut la dessiner sur la carte.

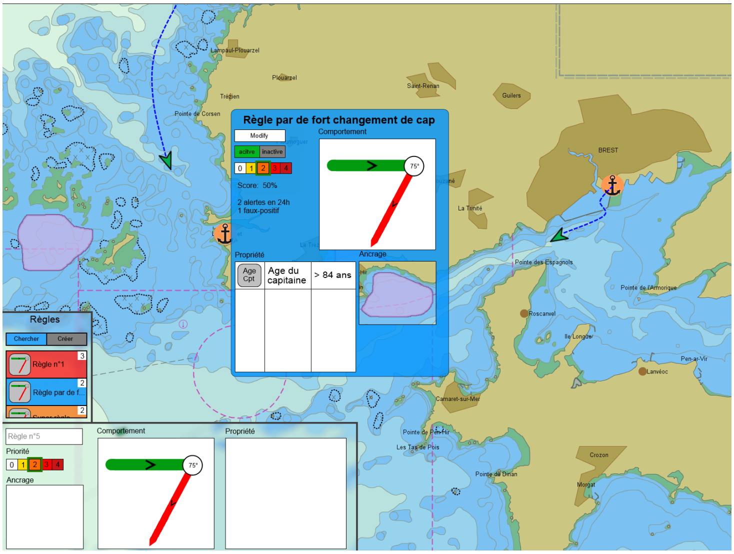
Propriété

20 alertes en 24h
3 faux-positifs

Fig. 3. Create a new rule in Hyperion platform

The second box concerns rule testing. At the top, the button "test" is used to test rule on the last 24 hours or less according to the recorded history. When a rule is tested a text area under test button displays the number of alerts that would have been raised over the recorded period, as well as the number of false alarms (boats triggering an alarm but considered as normal by the operator). Finally, the circle allows user to replay the recorded history to understand better the alerts triggered by his rule. To play history, it is possible to tap on the play symbol or to move the slider around the symbol to go backwards or forwards.

Figure 4 shows the research of rules, where the user in looking for rules expressing a given behaviour: a list appears above the rule view. It contains every rule in the system, which corresponds to the research. To apply an existing rule on a map element, user just has to drag and drop rule from the list, to the element. In this case, the rule is cloned and the "anchorage" property of the cloned rule is replaced with the new map element.



Règle par de fort changement de cap

Score: 50%

2 alertes en 24h
1 faux-positif

Propriété

Age Cpt	Age du capitaine	> 84 ans

Règle n°5

Fig. 4. Rule selection in Hyperion GUI

This section aimed at the description of the Hyperion system, and how it uses rules: next section focusses more on the use of AG in the platform.

5 Use of Analog Gesture for Vessel Trajectory Appointment

Let us consider a simplified example to describe how AGs are used in Hyperion: suppose the controller wants to describe a vessel trajectory starting from a harbor, suddenly turning to the southeast while accelerating.

Without AG, operator should to select the harbor to indicate the anchorage property of the rule. To express the abnormal trajectory, the operator has to learn the rule syntax and write it in a text editor like in [3].

With AG, the surveillance controller has just to perform a gesture from the point on the map indicating the harbor, going away from the port at normal speed and turning abruptly and rapidly to the bottom right corner of the touch table screen (figure 5a). In a same gesture and without learning process from the operator, “anchorage” and “behavior” properties are added to the edited rule.

When the gesture is completed, a feedback appears on rule view (figure 5b). This feedback allows the user to see what is understood by the system. If the operator does not agree with the recognized TSP, he/she can cancel his/her action by performing another gesture.

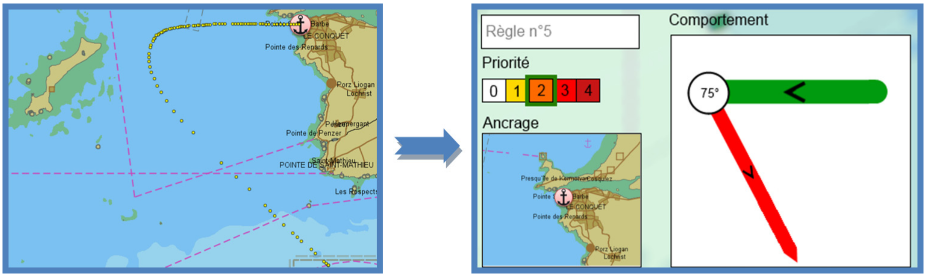


Fig. 5. a- Example of real gesture (left) and b- its corresponding feedback, with the same sampling frequency.

Analog gesture is limited to maritime surveillance domain. In aerial monitoring domain, it would be possible to use this type of gesture to watch the air traffic. It would be possible to add AG to easily plan the itinerary of an unmanned vehicle like UAV or unmanned car.

6 Conclusion

We saw that in some situation, we need to refer to TSP global features. Therefore, we propose the concept of analogical gesture which allows people to directly match to the TSP of reference in the system. Finally, we present the Hyperion platform, our maritime surveillance application which uses the AG.

The analog gesture recognizer is developed. Now, we have to make some experiments to know if it is a real benefit for an operator to express the main characteristics of a trajectory via a gesture, in terms of precision and time to realize this task.

References

1. Appert, C., Zhai, S.: Using strokes as command shortcuts: cognitive benefits and toolkit support. In: CHI 2009, pp. 2289–2298 (2009)
2. Bau, O., Mackay, W.E.: OctoPocus: a dynamic guide for learning gesture-based command sets. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, pp. 37–46. ACM (2008)

3. Cetin, F.T., Yilmaz, B., Kabak, Y., et al.: Increasing Maritime Situational Awareness with Interoperating Distributed Information Sources. In: 18th Interantional Command and Control Research and Technology Symposium, pp. 9–22 (2013)
4. Holz, C., Feiner, S.: Relaxed selection techniques for querying time-series graphs. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology - UIST 2009, vol. 213. ACM Press (2009)
5. Iannizzotto, G., Vita, L.: A multiscale turning angle representation of object shapes for image retrieval. Visual Information and Information Systems, 609–616 (1999)
6. Karam, M.: A taxonomy of gestures in human computer interactions, pp. 1–45 (2005)
7. McNeill, D.: Gesture & Thought. University of Chicago Press (2005)
8. McNeill, D.: Gesture: A Psycholinguistic Approach. The Encyclopedia of Language and Linguistics, pp. 1–15 (2006)
9. Nilsson, M., van Laere, J., Ziemke, T.: Extracting rules from expert operators to support situation awareness in maritime surveillance. In: 2008 11th Fusion, pp. 908–915 (2008)
10. Paulson, B., Hammond, T.: Paleo Sketch: accurate primitive sketch recognition and beautification. ...of the 13th International Conference on ..., pp. 1–10 (2008)
11. Rubine, D.: Specifying gestures by example. ACM SIGGRAPH Computer Graphics 25(4), 329–337 (1991)
12. Scoditti, A.: Gestural interaction techniques for handheld devices combining accelerometers and multipoint touch screens. Sciences-New York (2012)
13. Vatavu, R., Anthony, L., Wobbrock, J.: Gestures as point clouds: a $ P recognizer for user interface prototypes. In: Proceedings of the 14th ACM ..., pp. 273–280 (2012)
14. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 159–168. ACM (2007)

Remote Collaboration with Spatial AR Support

Nobuchika Sakata, Yuuki Takano, and Shogo Nishida

Osaka University, Machikaneyama 1-3, Toyonaka City, Osaka, Japan
{sakata,takano,nishida}@nishilab.sys.es.osaka-u.ac.jp

Abstract. Typical view sharing system has same camera alignment that camera take images from back of remote instructor. We change this alignment to camera take images from front of remote instructor for preventing occlusions caused by a body of remote instructor self. Also as visual feedbacks, a mirror image of remote instructor is indicated in display of remote instructor side. Eventually remote instructor can confirm own instruction in the display. Therefore due to displaying the mirror image of remote instructor and changing camera alignment, we proposed and implement a novel remote collaboration system which prevents occlusion problems caused by instructor body self when he/she sends clear instructions by whole body gesture and allows instructor to use direct manipulation.

Keywords: Remote collaboration, Occlusion, Augmented Reality, View sharing system, Spatial AR.

1 Introduction

Work conducted by a local worker under the instructions of a remote instructor is called remote collaboration [1-3]. 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. To conduct such communication smoothly, a support system sends the remote instructions including the place of 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 indicate the gesture image to the real-world directly [9-12]. 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. This study focuses on remote collaboration in which a local worker works with real objects using a remote instructor. The goal of this study is to achieve an interaction that allows a remote instructor to provide a local worker with clear and accurate instructions by means of various gestures. A view sharing system between remote instructor and local worker are often used in this type of remote collaboration. In particular, we focus to occlusion problems when making clear

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 148–157, 2014.

instructions by gestures. To solve this problem, we apply spatial augmented reality technique to view sharing system for remote collaboration..

2 Related Work

Some researches study 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 teleoperated laser pointer is adopted in some research as a pointing tool for remote collaboration [4-6]. Cterm [4] and GestureLaser [5] are device placed in a work space, and WACL [6] is a wearable device. Each of these is compact size and consists of a camera, a microphone, a speaker and a laser pointer which can be controlled remotely. The instructor can pan and tilt the laser pointer on the camera to point at real-world objects. GestureMan [7] is a system equipped with not only a teleoperated 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 [8] 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 researches display the image or the shadow of the instructor on the work environment [9-12]. These systems allow to transmit the embodiment and awareness to the remote worker by sharing their arms and gestures each other on the displayed image. These research show the remote communication becomes smooth by considering embodiment and transmitting the awareness information or gestures. Therefore, the instruction via work field images including target object 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 little researches which proposes the instructor system, the remote interaction for the instructor and deploying spatial AR techniques.

3 View Sharing System Using Instructor Mirror Image

In typical view sharing system for remote collaboration, as conveying remote instruction to local worker, researchers studied indicating only arm image of instructor and line drawing [14-15], whole body or upper of instructor [13][16] and some instructions in VR space rebuilt for remote instructor. Especially, we focus to Kuzuoka works [17] that upper body image can help to understand instructor's gestures and measure intelligibility of worker and instructor. Along to the principle, we use images of instructor's upper body to support nonverbal communication.

In typical view sharing system for remote collaboration, HMD and table top display are used for displaying situation of remote instructor and local worker sites each other. Remote instructor and local worker cannot take different field of view because

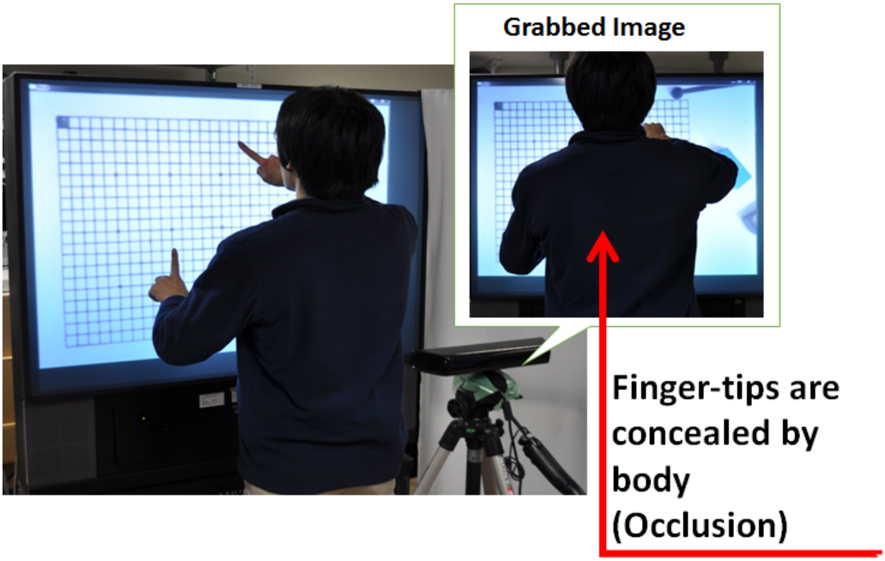


Fig. 1. Occlusion problems in typical view sharing system

view of remote instructor and local worker are perfectly matched in remote collaboration systems such as both remote instructor and local worker wears HMD. Finally, remote instructor cannot observe worker sites freely, and then performance of remote collaboration is decreased in tasks such as searching and picking. Using table top display in such remote collaboration are often used to compensate the previous problem that remote instructor cannot observe working site. In such remote collaboration system, instructions for real objects are conducted by gestures of only finger and arm. However, gesture of whole body, face and other body parts cannot be used in such kind of system due to deployment of the display placed horizontally. In our proposed system we use old fusion wall type LCD panel as output device for instructor to use gestures of whole body.

Typical view sharing system for remote collaboration has almost same camera alignment that camera take images from back of remote instructor as shown in fig 1. We change this alignment to take camera images from front of remote instructor for preventing occlusions caused by a body of remote instructor self as shown in bottom of fig 2. Also as visual feedbacks, a mirror image of remote instructor is indicated in display of remote instructor side as shown in fig 3. Eventually remote instructor can confirm own instruction in the display with this Spatial AR technique. Therefore due to displaying the mirror image of remote instructor and changing camera alignment, we proposed and implement a novel remote collaboration system which prevents occlusion problems caused by instructor body self when they sends clear instructions by whole body gesture and allows instructor to use direct manipulation. We suppose to unveil relationship among tasks efficiently, whole body gesture and even face expressions in remote collaboration regarding objects and places in real work spaces.

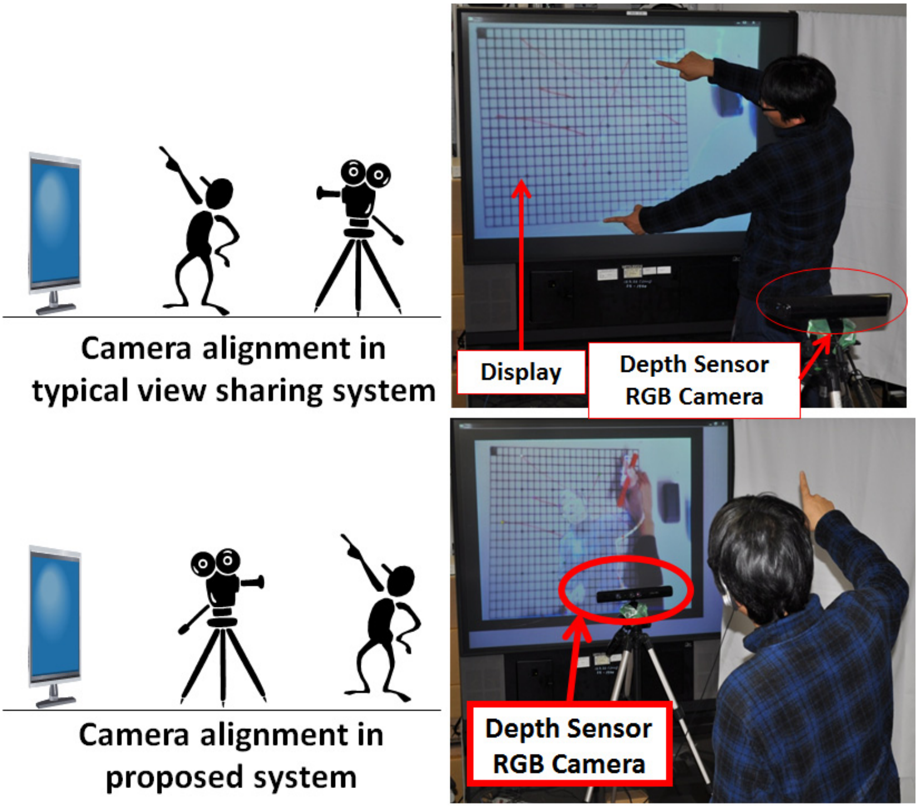

Camera alignment in typical view sharing system
Display
Depth Sensor RGB Camera
Camera alignment in proposed system
Depth Sensor RGB Camera

Fig. 2. Camera alignment in typical view sharing system and proposed system

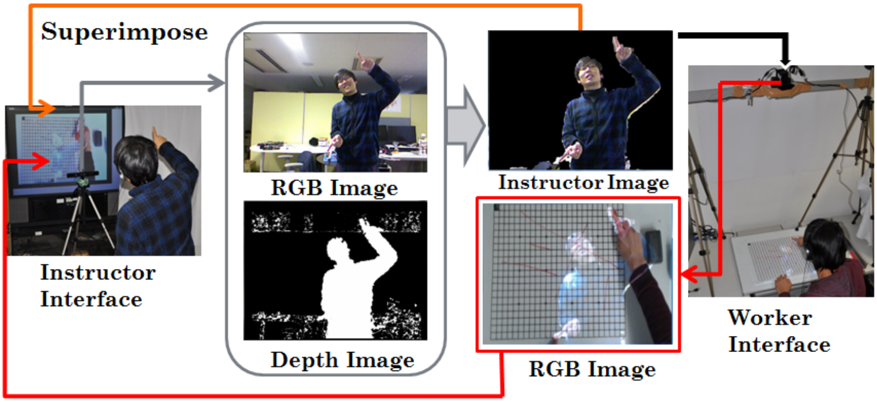


Fig. 3. System diagram

3.1 System Overview

Our proposed system is composed of instructor and worker interfaces. Fig 4 shows appearance of instructor interface. It composed of Flat panel display (Mitsubishi, 55P-FD100) to indicate video image of worker side, and Microsoft Kinect to capture instructor's body movement. The Kinect is deployed between the Flat panel display and standing position of instructor. Kinect can capture depth and RGB image of instructor at the same time. Image of only instructor can be extracted from RGB image according to the depth image. After that extracted instructor image is sent to worker interface as shown in Fig 3. Also instructor cannot touch the display directly with deploying the Kinect between the display and standing position of instructor. Instructor cannot recognize where instructor is pointing to exact. It causes a lack of direct touch and decreasing usability. To compensate those issues, transparent image of instructor body is superimposed to image on the display as shown in upper right of Fig 4.

Fig 5 shows worker interface. The worker interface is composed of a projector (Mitsubishi, LVP-DX95) and RGB camera (Logicool, HD Pro Webcam C920). The RGB camera capture circumstance of worker side. Image of instructor upper body is overlaid on working place with the projector as shown in Fig 5. This interface is not special and this style is typical Procams (Projector and Camera systems). Because of that we do not focus to improvement of worker interface in this paper.

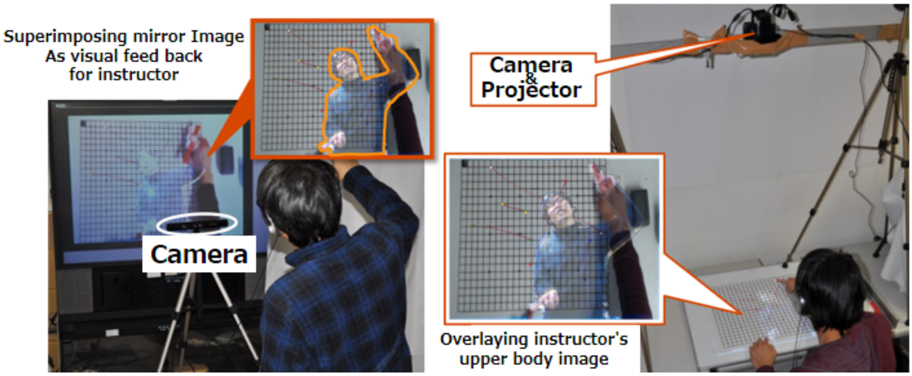


Fig. 4. Appearance of instructor interface (Left) and worker interface (Right)

Fig 3 shows process of instructions. Worker interface capture working place including worker's hand and objective parts of this task. The captured image stream is sent to instructor interface via TCP/IP network. At this time, we apply keystone effect to configure and compensate distortion. And then the corrected image indicate on the large display located in front of the instructor. Observing the image stream of worker side, the instructor make instructions by finger pointing and whole body gesture.

Also, image of instructor upper body can be extracted according to depth image of Kinect. Those image is sent to worker side, and then the instructor upper body is overlaying on work place. Finally, instructor's finger pointing and whole body gesture are duplicated in worker side. Referring duplicated nonverbal channel, the worker conduct tasks. Simultaneously, the image is superimposed to the display in instructor side as visual feedback for the instructor. Also aural communication can be used each other as full duplex via Skype.

3.2 User Study and Result

We conduct 2 type user study. After that we called them Experiment 1 and 2. Task 1 can be accomplished by pointing out only 1 place. Also we set that occlusion problems do not occur so much as the experiment condition of Experiment 1. Aim of Experiment 1 is to confirm that the proposed system can mark almost same performance as well as typical view sharing interface as shown in upper of fig 2.

In Experiment 2, subject should be pointing out 2 places to accomplish. We set Experiment 2 condition that occlusion problems occur so much. Aim of Experiment 2 is to confirm that the proposed system can perform much more than typical view sharing interface.

In Experiment 1, we set a task that subjects place some blocks on gridded paper as shown in left of Fig 5. The blocks are painted by random color pattern not to distinguish at once. Also the gridded paper is painted by random color pattern. It means that we let instructor use finger pointing rather than aural instruction. In experiment 2, we set a task that subject draw a line from two point indicated by remote instructor on a gridded board (Right of Fig 5). Some base points are painted as large black circle on the gridded board. The base points avoid that remote instructor spend time for searching two points which convey to a local worker.

12 subjects (ages 21-25, 11 male and 1 female) conduct Experiment 1 and 2 to consider order effects. As a result, typical view sharing is faster than the proposed interface in Experiment 1. Significant difference is found in task completion time of Experiment 1. In Experiment 2, the proposed interface is faster than the typical view shared interface. Significant difference is found in task completion time of Experiment 2.

Also we conduct questionnaire after Experiment 1 and 2. In experiment 1, we cannot find significant difference among almost all evaluation items excepting "Which condition do you transmit the instruction easier?". In terms of "Which condition do you transmit the instruction easier?", the proposed interface is more easier than typical interface. In Experiment 2, the proposed interface obtain good impressions and obtain significant differences in "Which condition do you transmit the instruction easier?", "Which condition do you conduct the instruction precisely?" and "Which condition do you feel burdens during instructions?". However, we cannot find significant difference in "Which condition do you conduct the instructions faster?"

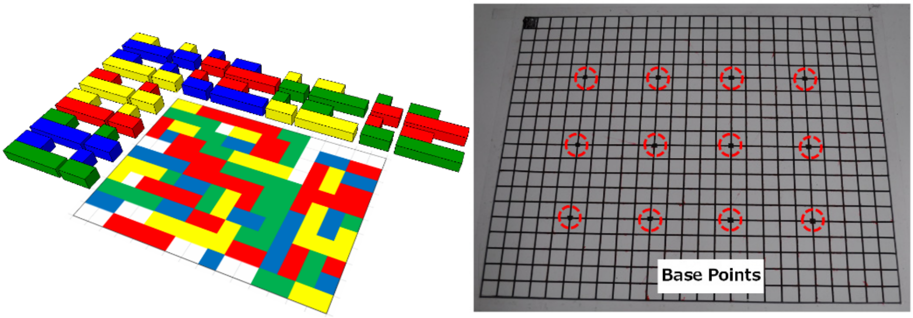


Fig. 5. Gridded board (Left). Blocks and gridded paper (Right).

3.3 Discussion

In Experiment 1 which needs one pointing instructions, the typical existed interface mark shorter completion time than the proposed interface. As quantities evaluation, we cannot find significant difference among almost all evaluation items excepting “Which condition do you transmit the instruction easier?”.

In Experiment 2 which needs two pointing instructions, the proposed interface marks shorter completion time than typical interface. As quantities evaluation, the proposed interface provides better impressions than the typical interface. Especially, quite large difference is obtained in evaluation item of “Do you feel burden when instructions?” because instructors should take unnaturally posture during instruction in the typical interface. It can say the proposed interface using mirror image can compensate those occlusion problems. Also the proposed interface can provide same impression when conducting one pointing instruction.

4 Sharing Face Expression among Worker and Instructor Using Mirror Image

In previous chapter, we proposed, implemented and evaluated view sharing system using instructor mirror image as application of remote collaboration with installing spatial AR techniques. As a result, we can compensate occlusions problems during two pointing instructions. Also we propose a method of sharing face expression among worker and multiple instructors using mirror image as other application of remote collaboration with installing spatial AR technique. We assume that application can mark good performance in remote collaboration between two or three instructors and one field worker.

As shown in fig 6, two or three instructor are considering how to instruct. Then, local worker can watch the conversation among instructors with observing face expression, gestures and body. The local worker might obtain much nonverbal communication comparing to a method of left and middle of Fig 6. Finally, understanding the conversation deeply, the local worker can conduct tasks smoothly.

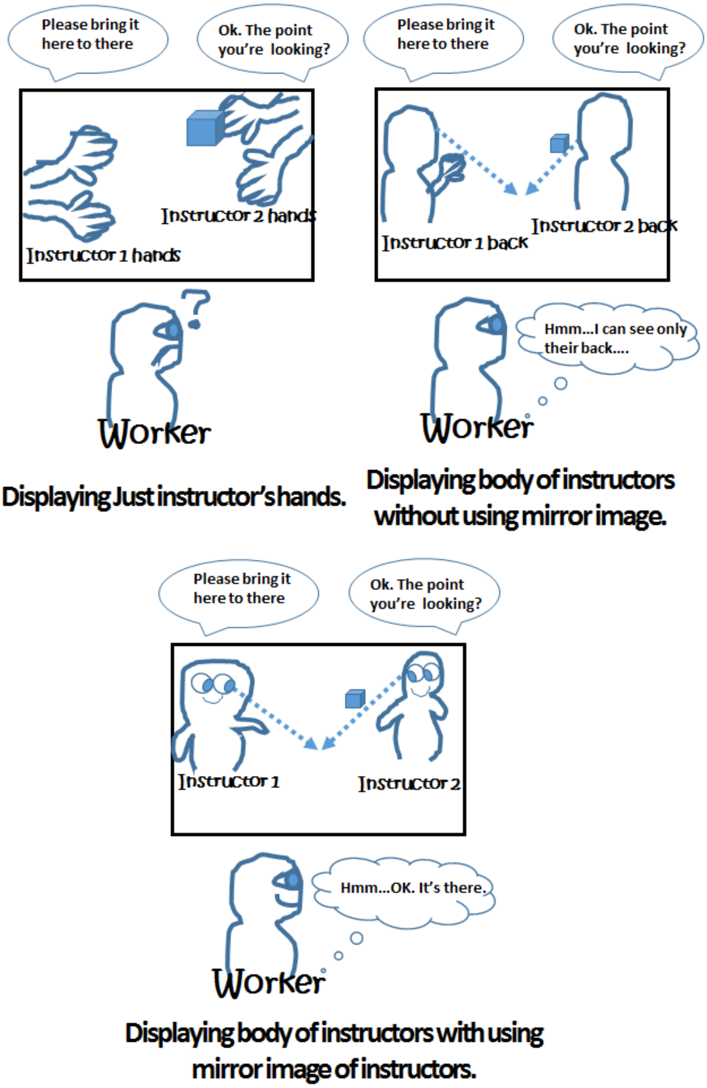

Please bring it here to there
Ok. The point you're looking?
Instructor 2 hands
Instructor 1 hands
Worker
Displaying Just instructor's hands.
Please bring it here to there
Ok. The point you're looking?
Instructor 1 back
Instructor 2 back
Hmm...I can see only their back....
Worker
Displaying body of instructors without using mirror image.
Please bring it here to there
Ok. The point you're looking?
Instructor 1
Instructor 2
Hmm...OK. It's there.
Worker
Displaying body of instructors with using mirror image of instructors.

Fig. 6. Advantage of sharing face expression with using mirror image of instructors

5 Conclusion

We propose, implement and evaluate new view sharing system for remote collaboration. We change this alignment to camera take images from front of remote instructor for preventing occlusions caused by a body of remote instructor self. Also as visual feedbacks, a mirror image of remote instructor is indicated in display of remote instructor side. Eventually remote instructor can confirm own instruction in the display. Therefore due to displaying the mirror image of remote instructor and changing camera alignment, we proposed and implement a novel remote collaboration system which prevents occlusion problems caused by instructor body self when he/she sends clear instructions by whole body gesture and allows instructor to use direct manipulation. Also we propose a method of sharing face expression among worker and multiple instructors using mirror image as other application of remote collaboration with installing spatial AR technique.

References

1. Kuzuoka, H.: Spatial workpace collaboration: Asharedview video support system for remote collaboration capavility. In: Proc. CHI 1992, pp. 533–540 (1992)
2. Fussell, S.R., 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. Mikawa, M., Matsumoto, M.: Smooth and easy telecommunication using CTerm. In: Proceedings of IEEE SMC 1999, pp. 732–737 (1999)
5. Yamazaki, K., Yamazaki, A., Kuzuoka, H., Oyama, S., Kato, H., Suzuki, H., Miki, H.: Proceedings of the Sixth Conference on European Conference on Computer Supported Cooperative Work, pp. 239–258 (1999)
6. Sakata, N., Kurata, T., Kato, T., Kourogi, M., Kuzuoka, H.: WACL: Supporting telecommunications using wearable active camera with laser pointer. In: Proceedings of the Seventh IEEE International Symposium on Wearable Computers, pp. 53–56 (2003)
7. 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)
8. Kondo, D., 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)
9. 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)
10. 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)
11. Yamashita, N., 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)
12. Izadi, S., Agarwal, A., Criminisi, A., Winn, J., Blake, A., Fitzgibbon, A.: C-Slate: A Multi-Touch and Object Recognitio System for Remote Collaboration using Horizontal Surfaces. In: IEEE Workshop on Horizontal Interactive Human Computer Systems, Rhode Island, USA, pp. 3–10 (October 2007)
13. 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)
14. Grevich, P., Lanir, J., Cohen, B., Stone, R.: TeleAdvisor: A Versatile Augmented Reality Tool for Remote Assistance. In: CHI 2012 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 619–622 (2012)

15. Hiura, S., Tojo, K., Inokuchi, S.: 3-D Tele-direction Interface using Video Projector. In: 30th Int'l Conf. on Computer Graphics and Interactive Techniques (ACM SIGGRAPH 2003) Sketches & Applications, San Diego, California, July 27-31. ACM, New York (2003)
16. 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)
17. Kuzuoka, H., Yamashita, J., Yamazaki, K., Yamazaki, A.: Agora: A Remote Collaboration System that Enables Mutual Monitoring. In: Proc. CHI 1999, pp. 190–191. ACM Press (1999)

Prediction of Multi-touch Gestures during Input

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. In the work at hand, a method is presented that can predict gestures during input. The scheme is based on the specification of prominent points defining subgestures within templates. Classification of a partial input is only against a small set of subgestures pre-selected by nearest neighbor searches regarding these prominent points. The gesture prediction is invariant against variations in scale, rotation, translation and speed of an input and handles single-touch, single-stroke and (sequential) multi-touch gestures. We provide thorough investigations of the classifiers performance on tests with two medium sized gesture sets. Results are promising and feasible for a wide range of applications. Even common direct manipulation operations can be reliably detected.

Keywords: gestures, multi-touch, prediction, classification, template-based.

1 Introduction and Motivation

In this work, the task of a gesture's prediction during input is investigated. In the best case, input is accompanied by continuously adapting interpretations in terms of the most probably intended gestures. In this way, users can finish their input as soon as enough of it is seen for proper recognition. Besides the incorporation of multi-touch to encode more information in time, such shortening can drastically reduce the time of gestural interaction, too. In addition, a predictive recognition can connect gestural interaction and direct manipulation or support tools for dynamic training of gestures. Freeman et al. [7] see the barrier for users in the necessity of learning complex physical input methods as the main cause why developers of commercial systems avoid implementations of multi-touch interaction beyond basic direct manipulations as defined by Shneiderman [18]. Consequently, literature mainly focuses on such training purposes or support of input by feedforward mechanisms that show current predictions [2,3].

2 Known Methods and Applications

In [14], purpose of a gesture's 'eager recognition' is the fluent transition from gesturing to direct manipulation, targeting to convey additional parameters of

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 158–169, 2014.

the operation. This is realized by forming two sets for each class, one for the sufficient complete ones (unambiguous) and one selection of ambiguous subgestures of all possible prefixes each gesture can contain. Under modification of weighting parameters by cross-validation, a binary classifier is trained that decides whether enough of an actual input is seen (input falls into an unambiguous set) to pass it to the standard classification method presented in the same work.

The classification method of [14] and its eager recognition routine is used in [9] for the interpretation of hand drawn sketches of ER-diagrams (4 simple geometric symbols for entities, relationships, and attributes). A similar method is used in [19] for the recognition of sketches. Partial (separated by strokes) sketches are added to the training data which is clustered (supported by a supervised SVM) by similarity of visually represented features (as in [12]). Assignment of partial inputs to a class is by a Bayesian approach.

The feedback system 'Octopocus' [2] supports single-touch input of gestures normalized regarding their size by presenting suggestions of possible progress. Beginnings of each classes' templates that correspond in length are replaced by the actual input. This modified gesture is then classified based on the method in [14] by Mahalanobis distances against the original template set. An alternative classification by distances between the shape signatures of angular traversal of the trajectory is proposed, but not investigated. The error measurement in terms of the distance of a template to the input indicates the probability of performing a gesture equivalent to this template. This probability is visually presented by the stroke's thickness in the depiction of each possible outcome. An approach similar to 'Octopocus' that only displays the most probably intended gesture to not stress the user is presented in [3]. In contrast to the work in [2], the prediction scheme scales templates to the bounding box of the current single-touch input.

More sophisticated estimation of the size of a partially entered gesture is done in [1] on the basis of a scale independent gesture representation by sequences of quantized absolute angles. Subsequences of similar angles are collapsed to achieve independence of an input's duration, which otherwise needs an equal distance resampling under knowledge of the complete gesture. Two thus computed shape signatures are compared - up to the length of the shorter one - by the ratio of their pair-wise angles that exceed a threshold of $(\pi/4)$. Is this ratio below 10%, a scaling factor is determined by the mean lengths of all examined segments[1] in both trajectories that are represented by those concordant angle-pairs. Tests showed an average over-estimation of a partial gesture's real size by 1/3.

In [11], a DTW approach is used to recognize partial input of planar gestures of the hands. Classification of a partial input is done by comparisons with subgestures of templates in length of the input's duration. A gesture network is used to model common subgestures and the formal prediction capacity. It provides points in time where the classification of a partial input is possible. Further progress is predicted by averaging trajectories following the input part in the graph for predefined gestures. However, common partial sequences of gestures require a manual analysis of the gesture set. Kawashima et al. [10] extend the

[1] Ignoring the last one, as it can not be determined, if it is already completed.

concept of [11] to handle input with strong variations in its duration compared to the specified and quantized (per Self Organizing Maps) templates. The method, however, requires the computation of Euclidean distances between input and all possible subsequences of a template to choose the most similar segment as soon as it differs by a threshold to the second most similar one.

In general, classification by DTW can also be done by relaxing constraints, so that partial matchings are possible or even preferred [8]. Further approaches can be found in other application areas. Classification by HMM, for instance, can be extended by combination of the models [21] or modification of the Viterbi procedure [6] to detect gestures in continuous input streams. Such 'gesture spotting' could be transferred to find partially entered gestures within templates.

The Main Problem of a gesture's early recognition is the estimation of the amount of input already done. Obviously, absolute criteria as time or the length of trajectories can provide sufficient indications of this amount, but presume certain restrictions in input (i.e. fixed size, orientation, speed or scale). Additionally, nearest neighbor searches within all possible subgestures of all templates are too expensive if several classifications are to be done during a gesture's input.

In the methods available so far, single-touch is the common form of input [14,2,1,3] and tools that provide sophisticated multi-touch gesture input at all are rare.[2]. Multi-stroke is supported, for instance, by [19], but prediction is restricted to partial sketches containing fully drawn strokes. Due to the selection of absolute (i.e. angular) features [14,1,19] normalization by size [2,3] or usage of time as a criterion to find subgestures of equal length [11], our required invariances are currently not fully supported, too.

We require the prediction of gestures to be invariant against variations in scale, rotation, translation, and speed as having such constraints limits the versatility of the classifier. On the other hand, if such natural variations in input are required, they can easily be integrated by parameter checks or enhancements by absolute features. To prevent restrictions to the diversity of gestures, we require our classification approach to handle single-touch gestures as well as multi-stroke or (sequential) multi-touch ones as defined in [17].

3 The Proposed Method

The proposed method is based on the definition of prominent points within templates. The recognition routine of [17] is applied at each new input sample point (time-outs are possible), but classification is only against templates with a prominent point similar to this last point in input. This way, the classification's workload can be scaled regarding real-time requirements. Each prominent/landmark point, is represented by a feature vector which is independent of a gesture's position, scale or orientation. If an input is compared to a template, the best fitting

[2] Direct manipulation operations (for instance, pinch-gestures) as defined by Shneiderman [18], though possibly applied by multi-touch, are not regarded as gestures.

prominent point allows to estimate the common portion within the template, which in turn allows the prediction of further progress and training schemes as in [2,16,3]. Our detailed procedure includes the following steps:

- Find prominent points in each template during training phase.
- Define a representation of landmarks in respect to features supporting all required invariances and incremental (with gesture length) computation.
- Store landmark points (together with references to the corresponding subgesture) in a data structure that provides fast searches for nearest neighbors.
- During input, find landmarks that are most similar to its currently terminating point and classify it against their corresponding subgestures.

We find prominent points within a gesture's token (trajectory) by a modified Ramer-Douglas-Peucker-algorithm [13,5] (abbr.: RDP). It approximates curves by omitting points that do not provide much information to its contour.

Algorithm 1. ModifiedRamerDouglasPeucker(T,n,f,l)

```
Require: INPUT: T - single trajectory of an gesture input
Require: INPUT: n - maximum number of landmark points
Require: INPUT: f - index of first point
Require: INPUT: l - index of last point        ▷ defining relevant sequence in trajectory
  ▷ by ignoring duplicates/first point, point set contains landmarks on end of recursion
  StoreToPointSet(T(f))
  StoreToPointSet(T(l))
  ▷ if further landmark points are to be included, find the one with maximum
  ▷ perpendicular distance to the line segment defined by index l and f
  if $n > 0$ & $l - f > 0$ then
      for all $i = f + 1$ to $l - 1$ do
          distance ← PerpendicularDistance(T(i),T(l),T(f))
          if $distance > maxdistance$ then
              landmark ← i
              maxdistance ← distance
          end if
      end for
      StoreToPointSet(T(landmark))
      ▷ distribute next landmarks approximately equally on left and right side of the
      ▷ currently found one by the number of samples within both parts
      left ← $n \cdot (landmark - f)/(l - f)$
      right ← n-left
      ModifiedRamerDouglasPeucker((T,left,f,landmark))
      ModifiedRamerDouglasPeucker((T,right,landmark,right))
  end if
```

In its original version, the recursive procedure successively selects points that contain the most relevant contour information. It terminates as soon as a point would be added whose distance to the polyline formed by already chosen points falls below a threshold. We adapted the procedure to collect a maximum number of points. If one such point is found, the reduced maximum number is distributed in the ratio of sample points within the trajectory that precede or follow this currently found prominent point. The first point in the token of a template is excluded from the list of landmark points after the algorithm's termination.

All landmark points found by the RDP-algorithm are represented by a feature vector. The features are depicted in figure 1. They contain the angle between the first point, landmark and its preceding point (1), the angle enclosed by the landmark's preceding point, landmark itself and an incremental center of gravity (2), the angle between first point, the point half way to the landmark (median) and the landmark (3) as well as the distance of incremental center of gravity to the landmark in relation to the length of the trajectory up to the landmark (4).

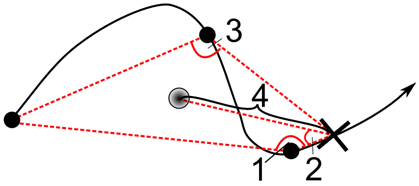


Fig. 1. The feature set of a landmark (cross) is used for retrieving similar points within templates. It contains measurements of angles and distances incorporating interesting points on the trajectory (black dots) and an incremental center of gravity (gray).

An additional feature not depicted in figure 1 is the cosine (self-) distance between two segments of the partial trajectory up to the landmark that are bisected by the median. It indicates the trajectory's continuity. If more than one (partial) trajectory is included in a subgesture up to the landmark, the structural and temporal features of [17] are added. All features can be computed in maximum time relative to the length of the trajectory and in this case incrementally.

For each landmark point within a trajectory of a gesture, the feature vectors of simultaneous points in possibly existing concurrent[3] trajectories are retrieved, too. The combined feature vector for all trajectories at a given point in time is seen as a point in multi-dimensional space. Every possible combination of the tokens' feature vectors (and a reference to the corresponding subgesture) is added in a kd-tree [4], appropriate for this number of tokens. This data structure supports efficient searches for nearest neighbors within radii of fixed number or range. Figure 2 illustrates the complete process.

For each new sample of an input, the kd-tree for its current number of tokens is chosen. A set of nearest neighbor landmarks in respect to the combined last points of the input's trajectories is requested and used for classification. The result of this classification is our best guess of the intended input.

[3] In case of terminated trajectories, their last point is chosen.

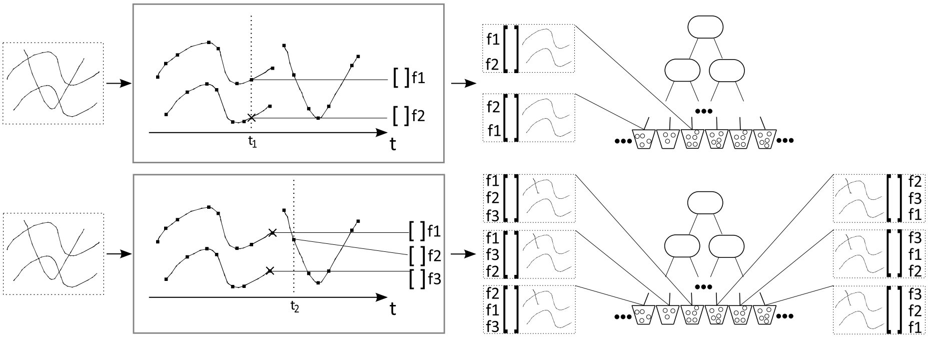


Fig. 2. Signal processing of a gesture (left). Landmark points for a token are detected and features retrieved and combined with features at simultaneous time or (if at the current time already completed) past endpoints of other tokens (second picture). For each ordering of tokens a feature vector is generated and the order corresponding subgesture is referenced by this point which is included in a kd-tree structure (right).

4 Evaluation

We evaluated our procedure by classification tests of partial gestures for two sets of templates. The first set (see figure 3 left) contains only multi-touch gestures and is introduced in [17]. The set was not developed for this purpose and due to inherent identical prefixes of the gestures[4], it is impractical for real world applications of gesture prediction. However, for analysing purposes and first impressions of our approach's performance, this systematical 'construction flaw' may be useful. For a more realistic scenario, a second set (see figure 3 right) was constructed which contains single- as well as multi-touch gestures. One member of each group of identical prefixes of set 1 was included. In addition to the three-finger pinch gestures, author-defined two-finger versions were added to investigate the potential for recognizing direct manipulations by our approach. Due to the lack of other known multi-touch gesture sets, a selection of letters from the gesture alphabet presented in [15] and four single-stroke gestures of [20] (in 'medium' speed) are included. Each gesture in the second set contains all available user-independent templates.

From the first set, six user-dependent test cases were generated and results averaged. One user-independent test case is used for the second set. For each test case of each set the following procedure was executed five times: For each of the 20 gesture classes in a set five[5] specifications were randomly selected as templates and five were randomly selected as test instances. In each template, ten landmark points were detected by the modified RDP method and their feature

[4] Considering delays in input between strokes, more than half of the gestures within the sets $\{1, 2\}$, $\{5, 6, 7\}$, $\{4, 9, 10, 11, 13, 14\}$, $\{15, 19\}$ and, at recognition invariant against rotation, $\{8,16\}$ are completely equal.

[5] In one exception only four templates were chosen as in the first gesture set, one user specified only nine templates for class 2.

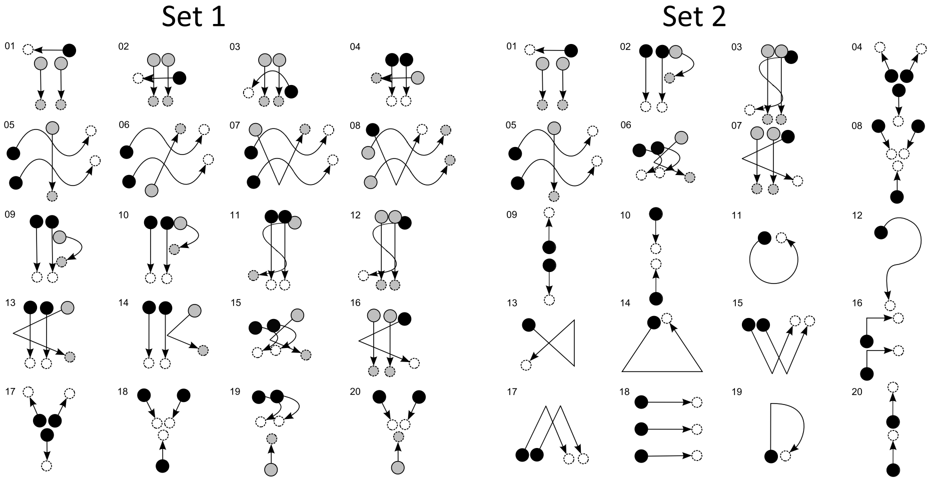


Fig. 3. The two gesture sets used in the tests. Depicted on the left is the original multi-touch gesture set. A modified and more realistic set combined with gestures of [20], multi-touch letters of a gesture alphabet [15] and two finger pinching gestures is seen on the right. Larger dots depict the start of a trajectory, arrows their movement and dashed smaller dots symbolize their end. Black colored starting points belong to the first stroke, gray ones to the second.

vectors together with the corresponding subgestures included in kd-trees.[6] The test instances were splitted into subgestures of lengths between 10% and 100% (in steps of 10%) by their temporal progress to simulate continuous gesture input (see figure 4). At classification of a subgesture, the search within a kd-tree (the one storing instances of the current input's number of tokens) was restricted to ten nearest neighbors (approximately 0.14% of all possible subgestures). The input is classified against these candidates and the best one returned as result.

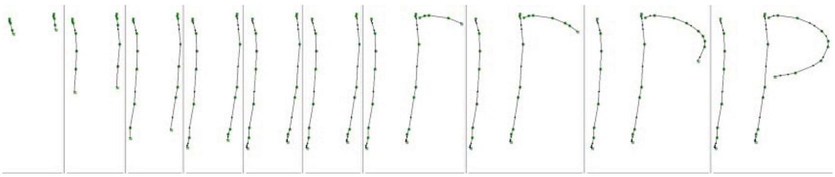


Fig. 4. Segmentation of a gesture (class 10 of first set) into 10%-steps of its duration

[6] For the first set, this resulted in approx. 7300 generated subgestures per test run.

5 Results and Discussion

In the following section, results regarding both gesture sets are presented. Table 1 lists accuracy values for classifications of partial gestures of set 1 sorted by average results per gesture class. The left side of figure 5 shows the ratio at which correct results are already included in the set of one to three or ten candidates chosen by the nearest neighbor searches for a partial input. On the right side, the results of the classification against the ten nearest candidates are presented. Accuracy values are explicitly given for 13 gestures with identical prefixes, the remaining seven as well as the best and the worst gesture class.

Table 1. Prediction Results for Gesture Set 1

Gesture	10	20	30	40	50	60	70	80	90	100	⊘
20	0.8	0.91	0.89	0.89	0.92	0.92	0.92	0.97	0.98	0.99	0.92
17	0.51	0.79	0.94	0.97	0.99	1	1	1	1	1	0.92
18	0.41	0.77	0.97	1	1	1	1	1	0.99	0.98	0.91
12	0.64	0.68	1	0.96	0.95	0.94	0.73	1	1	1	0.89
8	0.95	0.64	0.8	0.84	0.84	0.82	0.99	1	1	0.99	0.89
3	0.62	0.8	0.85	0.82	0.83	0.91	0.79	0.93	0.94	0.99	0.85
15	0.48	0.73	0.74	0.78	0.83	0.81	0.95	1	1	1	0.83
19	0.51	0.66	0.85	0.84	0.87	0.85	0.85	0.87	1	1	0.83
16	0.59	0.53	0.78	0.85	0.85	0.87	0.73	1	1	1	0.82
1	0.54	0.52	0.55	0.57	0.57	0.44	0.71	0.92	1	1	0.68
2	0.36	0.49	0.44	0.42	0.42	0.52	0.78	0.96	0.99	0.99	0.64
11	0.17	0.23	0.35	0.35	0.59	0.87	0.87	0.87	0.98	0.98	0.63
7	0.27	0.29	0.46	0.48	0.49	0.53	0.8	0.88	0.97	0.97	0.61
13	0.34	0.33	0.39	0.39	0.39	0.65	0.84	0.86	0.89	0.98	0.61
6	0.37	0.3	0.45	0.47	0.52	0.55	0.57	0.88	0.99	0.95	0.61
14	0.18	0.32	0.33	0.26	0.27	0.6	0.97	1	1	1	0.59
9	0.27	0.28	0.23	0.37	0.37	0.53	0.95	0.96	0.96	0.97	0.59
10	0.22	0.2	0.29	0.23	0.23	0.39	0.93	0.96	0.98	0.99	0.54
5	0.31	0.37	0.36	0.4	0.41	0.42	0.45	0.78	0.93	0.96	0.54
4	0.24	0.22	0.29	0.37	0.4	0.42	0.63	0.85	0.97	0.98	0.54
⊘	0.44	0.5	0.6	0.61	0.64	0.7	0.82	0.93	0.98	0.99	0.72

As soon as more than 10% of a gesture is entered, in over 50% of the cases, nearest neighbor templates already represent correct results. For restricted searches to ten candidates, an upper bound is given in figure 5 (left). Recognition rates of 90% are possible if at least 20% of an input is seen. At 80% of input, 98% accuracy can be achieved by correct choices from nearest neighbor sets. In comparison (figure 5 right), average prediction accuracy is above 50% with at least 20% of a gesture entered. A correct selection within the two nearest neighbors would give better results (overall in average 94% against 72% actual prediction rate). With progression of input, the classifier's selection from nearest neighbor sets becomes more reliable.[7] Actual prediction accuracy is best for gesture 17 which is classified correctly at more than 60% of input and predicted with at least 94% rate if more than 30% of it is seen. The seven gestures without common prefixes are recognized with no less than 88% if input passes 30%.

[7] Prediction rates for finished gestures (100%) are only slightly worse than results of the classification approach alone against all full templates.

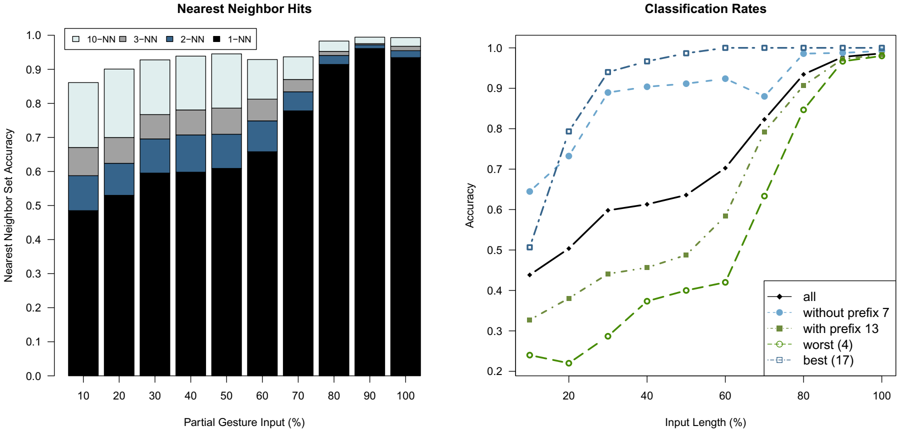


Fig. 5. The ratio at which one to three or ten nearest neighbor templates of partial gestures already include correct results is shown left. Next to it, classification accuracies in relation to a partial input's progress are presented. Results for best and worst gestures and subsets that are or are not afflicted by identical prefixes are given separately.

Prediction accuracies regarding gesture set 2 are given in table 2. Figure 6 (left) illustrates these results in comparison to prediction hits by nearest neighbor searches.

Table 2. Prediction Results for Gesture Set 2

Gesture	10	20	30	40	50	60	70	80	90	100	⊘
9	0.76	1	1	1	1	1	1	1	1	1	0.98
2	0.92	0.88	0.92	1	1	1	1	1	1	1	0.97
10	0.8	0.96	0.96	1	1	1	1	1	1	1	0.97
8	0.56	0.8	1	1	1	1	1	1	1	1	0.94
6	0.64	0.84	0.96	0.92	0.92	0.92	1	1	1	1	0.92
19	0.6	1	0.92	0.56	1	1	1	1	1	1	0.91
4	0.44	0.92	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.9
18	0.36	0.8	0.84	0.92	0.96	1	1	1	1	1	0.89
16	0.64	0.96	0.92	0.96	0.96	1	1	0.88	0.68	0.88	0.89
5	0.44	0.64	0.84	1	0.92	0.92	0.92	0.96	1	1	0.86
15	0.64	0.56	0.56	0.72	0.84	1	1	1	1	1	0.83
11	0	0.4	0.96	1	1	1	1	1	0.96	0.88	0.82
3	0.76	0.48	1	0.76	0.76	0.72	0.68	1	1	1	0.82
17	0.68	0.88	0.84	0.64	0.48	0.68	1	0.92	0.96	0.96	0.8
7	0.48	0.56	0.88	0.8	0.76	0.76	0.76	1	1	1	0.8
20	0.56	0.64	0.68	0.8	0.84	0.96	0.92	0.88	0.88	0.6	0.78
12	0.2	0.4	0.76	0.8	0.84	0.8	0.8	0.88	0.96	1	0.74
1	0.72	0.52	0.44	0.44	0.44	0.36	0.88	0.96	1	1	0.68
13	0.24	0.24	0.16	0.28	0.44	0.48	0.64	0.88	1	0.92	0.53
14	0.12	0.24	0.36	0.4	0.4	0.4	0.36	0.48	0.96	0.88	0.46
⊘	0.53	0.69	0.8	0.8	0.83	0.85	0.9	0.94	0.97	0.95	0.82

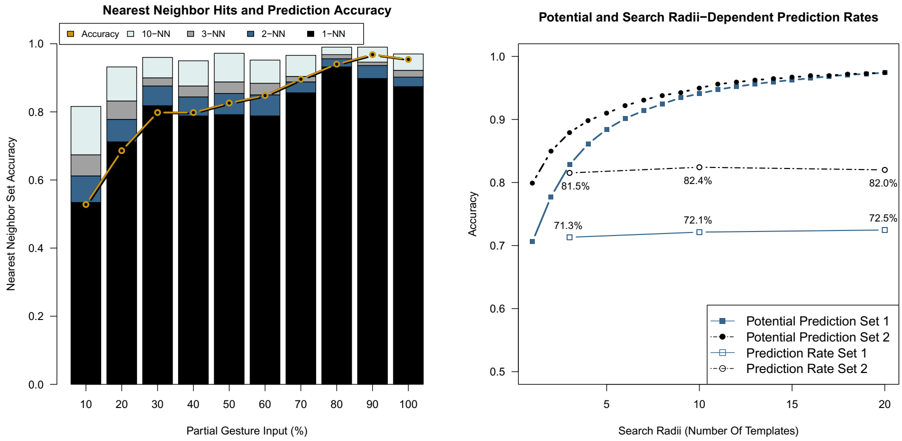


Fig. 6. Left: Achieved prediction rates in comparison to the rate at which the set of one to three or ten nearest neighbor templates of a partial input already include the correct result for classifications regarding gesture set 2. Right: Averaged prediction rates at three different restrictions (3, 10, 20) of the nearest neighbor search in comparison to potential prediction rates on optimal choices, i.e., the average rate on which a correct result would be within the nearest neighbor sets of different sizes.

Accuracies for gesture set 2 are no less than 80% at at least 30% progression in a gesture's input and no less than 90% if at least 70% of a gesture is entered. In average, an overall prediction rate of 82% is achieved.[8] Again, if less than 30% of a gesture is seen, the nearest neighbor selection alone would give best results and with input's progress the subsequent classification gains benefit.

Trying to get more insight into how our approach is influenced by parameter choices, figure 6 (right) shows averaged (over all lengths) rates of the correct result being within the set of one to 20 nearest neighbors (prediction potential). Besides, actual prediction rates by our procedure at three different sizes of nearest neighbor sets are given. The results show that the set of 20 nearest neighbors already includes a good pre-selection of gesture templates and searches beyond that size promise no significant improvements. On the other hand, prediction rates for our two gesture sets do not improve with a double sized set (20) at all. The limiting factor probably is a suboptimal selection of good candidates at the first phases of an input.

The tokens of gestures in set 1 contained 20-27, in average 24, sample points. Choosing all sample points as landmarks, recognition accuracy improves marginally by 0.59%. The same modification, however, increases 10-nearest neighbor hits regarding gesture set 2 from 80% to 83% and average prediction rates from 82% to 85%. Classification of a partial gesture of set 1 required with

[8] Completed gestures of gesture set 2 were classified with an accuracy of 99.80% by the classifier alone using all full templates.

our test setting[9] on average 97.73ms whereas the mean input time for our randomly chosen test gestures was 1.41s with minimum of 1.29s. Practically, more than ten classifications per input would therefore be possible.

6 Outlook

We presented a method allowing to predict gestures during input that were specified by templates. This approach can support gesture designers and application developers in quick prototyping or investigating manifold gesture interaction techniques. Utilizing a realistic gesture set, even common (for zooming or rotation) gestural direct manipulation operations can be handled when specified by templates only. If some invariances to input variations are not required, the feature set can be enhanced by a small set of absolute measurements for further accuracy improvements. Improvements are conceivable by incorporating passed observations or predictions by nearest neighbors only depending on an input's progression. Still we are keen to see sophisticated applications for a gesture prediction scheme that are beyond every day multi-touch interaction. We imagine tools to provide dynamical feedforward mechanisms for versatile sketching, text input systems, and training.

References

1. Appert, C., Bau, O.: Scale detection for a priori gesture recognition. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 879–882. ACM, New York (2010)
2. Bau, O., Mackay, W.E.: Octopocus: A dynamic guide for learning gesture-based command sets. In: UIST 2008: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, pp. 37–46. ACM, New York (2008)
3. Bennett, M., McCarthy, K., O'Modhrain, S., Smyth, B.: Simpleflow: Enhancing gestural interaction with gesture prediction, abbreviation and autocompletion. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part I. LNCS, vol. 6946, pp. 591–608. Springer, Heidelberg (2011)
4. Bentley, J.L.: K-d trees for semidynamic point sets. In: Proceedings of the Sixth Annual Symposium on Computational Geometry, SCG 1990, pp. 187–197. ACM, New York (1990)
5. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica: The International Journal for Geographic Information and Geovisualization 10(2), 112–122 (1973)
6. Deng, J.W., Tsui, H.T.: An hmm-based approach for gesture segmentation and recognition. In: Proceedings of the 15th International Conference on Pattern Recognition, vol. 3, pp. 679–682 (2000)

[9] AMD Phenom II X4 945 processor (3.01 Ghz).

7. Freeman, D., Benko, H., Morris, M.R., Wigdor, D.: Shadowguides: Visualizations for in-situ learning of multi-touch and whole-hand gestures. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2009, pp. 165–172. ACM, New York (2009)
8. Giorgino, T.: Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. Journal of Statistical Software 31(7), 1–24 (2009)
9. Henry, T.R., Hudson, S.E., Newell, G.L.: Integrating gesture and snapping into a user interface toolkit. In: Proceedings of the 3rd Annual ACM SIGGRAPH Symposium on User Interface Software and Technology, UIST 1990, pp. 112–122. ACM, New York (1990)
10. Kawashima, M., Shimada, A., Nagahara, H., Taniguchi, R.-I.: Adaptive template method for early recognition of gestures. In: 2011 17th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), pp. 1–6 (2011)
11. Mori, A., Uchida, S., Kurazume, R., Taniguchi, R., Hasegawa, T., Sakoe, H.: Early recognition and prediction of gestures. In: Proceedings of the 18th International Conference on Pattern Recognition, ICPR 2006, vol. 3, pp. 560–563. IEEE Computer Society, Washington, DC (2006)
12. Ouyang, T.Y., Davis, R.: A visual approach to sketched symbol recognition. In: Proceedings of the 21st International Jont Conference on Artifical intelligence, IJCAI 2009, pp. 1463–1468. Morgan Kaufmann Publishers Inc., San Francisco (2009)
13. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing 1(3), 244–256 (1972)
14. Rubine, D.: The Automatic Recognition of Gestures. PhD thesis, Carnegie Mellon University (1991)
15. Schmidt, M., Fibich, A., Weber, G.: Mtis: A multi-touch text input system. In: Streitz, N., Stephanidis, C. (eds.) DAPI 2013. LNCS, vol. 8028, pp. 62–71. Springer, Heidelberg (2013)
16. Schmidt, M., Weber, G.: Multitouch Haptic Interaction. In: Stephanidis, C. (ed.) UAHCI 2009, Part II. LNCS, vol. 5615, pp. 574–582. Springer, Heidelberg (2009)
17. Schmidt, M., Weber, G.: Template based classification of multi-touch gestures. Pattern Recognition 46(9), 2487–2496 (2013)
18. Shneiderman, B.: Direct manipulation: A step beyond programming languages. Computer 16(8), 57–69 (1983)
19. Tirkaz, C., Yanikoglu, B., Sezgin, T.M.: Sketched symbol recognition with autocompletion. Pattern Recognition 45(11), 3926–3937 (2012)
20. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In: UIST 2007: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 159–168. ACM, New York (2007)
21. Yang, J., Xu, Y.: Hidden markov model for gesture recognition. Technical Report CMU-RI-TR-94-10, Robotics Institute, Pittsburgh, PA (May 1994)

"Will Use It, Because I Want to Look Cool" A Comparative Study of Simple Computer Interactions Using Touchscreen and In-Air Hand Gestures

Vidya Vaidyanathan and Daniel Rosenberg

San Jose State University
vidya.vn@gmail.com, dan@rcdoux.com

Abstract. The Xbox Kinect and now the Leap Motion Controller have brought about a paradigm shift in the way we interact with computers by making the recognition of 3D gestures affordable. Interfaces now understand natural user interfaces, integrating gestures, voice and various other kinds of multimodal input simultaneously. In this paper we attempted to understand in-air gesturing better. The purpose of the study was to understand differences between touchscreen and in-air gesturing for simple human computer interactions. The comparison of the gestures was done in terms of Muscle effort/fatigue and Frustration, Satisfaction and Enjoyment We have also tried to study the learnability of in-air gesturing. In our research we found that in-air gesturing was significantly superior with respect to muscle effort and fatigue when compared with touchscreens. We also found that in-air gesturing was found to be more fun and preferred because of its "coolness factor". Lastly, in-air gesturing had a rapid learning curve.

Keywords: HCI, Touch Screens, in-air gestures, ergonomics, EMG, learnability, social acceptability, natural user interfaces (NUI).

1 Introduction

Gone are the days when user interfaces were based entirely on buttons, joysticks, keyboards and mice. Today the world has advanced into direct manipulation devices such as touchscreens and smart phones. An external device that maps onto the x-y co- ordinate system of a computer control is no longer required. The future of the computing world lies in interfaces described in the press as gesture controlled, motion-controlled, direct, controller- less and natural. The most popular gesture controlled devices that exist in the market today are gaming devices such as the Nintendo Wii, the Microsoft Kinect, the Sony Eye Toy and the Leap Motion. Smart phones and tablets are joining the trend of using gestures.

1.1 What Is Gesture Recognition?

Gesture recognition mainly concerns with identifying, recognizing and making meaning of human movements. The human body parts involved can be the hands, arms,

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 170–181, 2014.

face, head or the body [1]. Kendon states that amongst all human body parts conveying gestures, the hand gestures are the most natural and universal [7]. They form a direct and instant form of communication. Hand gestures are therefore the most used method for interaction with technological systems [7]. According to Dr. Harrison at Carnegie Melon University, the human hands alone are capable of tens of thousands of gestures, individually and in combination. Some tasks hinder the use of hands to interact with devices, such as checking email while driving a car [2]. Atia et al showed that in such cases certain applications use face and body related gestures. They also showed that using the leg to gesture was limited due to the spatial constraints [2].

1.2 Background and Related Work

In a survey, where Americans were polled for the top two inventions that improved their quality of life, "television remote" and "microwave oven", emerged as the winners [17]. Freeman and Weissman explored the control of a television using gesture recognition [17]. They compared voice and gesture as two candidates for equipment control. Voice had the advantage of having an established vocabulary, but was deemed not appropriate for the context. Gestural control was more appropriate for the context, but lacked a natural vocabulary [17].

Perzanowski et al explored the possibility of building a multi- modal interface based on voice and gestures [13]. Their interface used natural gestures especially those made using the arms and hands. They made use of meaning-bearing gestures that were associated with locational cues for a human-robot interaction. The meaning-bearing gestures mainly included indication of distances (by holding the hands apart) or directions (tracing a line in the air). When a user says "go there", the accompanying gesture signaling the direction was essential to make sense of the verbal command. Perzanowski et al observed that in noisy environments, gestures was largely used to compensate for lack of comprehensible auditory input [13].

The idea of using "free hand" gestures as an input medium is based out on the famous "put that there" experiment conducted in 1979. This experiment used primitive gestural input in the form of gestural languages: Task control primarily used gestures. Sign language interpretation was one of them. Some other examples were those where Sturman [16] presented a gestural command system to orient construction cranes, while Morita et al showed the use of gestural commands in an orchestra [12].

1.3 Naturalness of Gestures

The more natural a gesture is to its context and the more coherent in its mapping to human performance, the higher its interaction fidelity will be [4]. Bowman et al conducted a series of experiments to answer the questions they posed. They found that increased "interaction fidelity" has an increasingly positive experience on the user performance and efficiency of user tasks. Natural gestures were especially beneficial when the tasks were more complex. Users perceived that interactions with a higher degree of interaction fidelity were more fun, engaging and had higher immersive value.

Considering learnability of a NUI, Wigdor and Wixon claim that a NUI is one that provides a quick and enjoyable learning experience from novices to skilled users [18]. This rapid learnability occurs due to practice. They also define an NUI to be extremely enjoyable.

1.4 Social Acceptability of Gestures

Beyond recognizability, the acceptability of gestures is also critical. Certain cultures have politeness conventions for gestural use [8]. For example, pointing with the left hand is considered impolite in the country of Ghana. Here, receiving and giving with the left hand is also considered taboo [9, 10]. Hand gestures might have some drawbacks, such as acceptance or rejection in a public space [2]. Atia et al found that public found it threatening when a user performed the gesture of a large circle in a public place [2]. Studies have examined the usability of hand gestures in different generic environments, especially public places [15]. Ronkienen et al conducted "tap gesture" based experiments where they presented participants with gesture-based scenarios and quizzed on their willingness to use the gesture in various situations [15]. It was observed that the social acceptability of performing a gesture was dependent on where it is performed and the audience it was performed for. Further, certain gestures could be viewed as threatening in public spaces. Rico and Brewster expanded Ronkienen's experiment and examined the social acceptability of eight common gestures, example wrist rotation, foot tapping, nose tapping, shoulder tapping, etc. [14]. They showed that acceptability depends on the combination of audience and workplace. For example in the US, nose tapping was acceptable when the performer was alone at home, or in a pub among strangers, but not when alone in a workplace or in front of friends and family.

1.5 Drawback of Gestures

Baudel and Beaudouin-Lafon extensively explored the limitations of any gesture recognition system [3]. Fatigue was found to be one the key limitations. Gestural communication used more muscular activity than simple keyboard interaction, mouse interaction or speech. The wrist, fingers, hands and arms all contributed to the commands. In order for the gestures to be of minimal effort, they had to be concise and fast. Over time they may induce fatigue in the user [5]. Among the more recognized tools to measure muscle fatigue is the Electromyographic (EMG) analysis [5]. The surface EMG has limitations related to electrode placement, skin impedance and cross-talk [11]. In spite of the limitations, the surface EMG has been shown to be a valid and reliable tool to identify muscle fatigue [5].

2 Experimental Design

The purpose of our study was to understand differences between touchscreen and in-air gesturing for simple computer interactions. Gestures were used to select from a series of tiles displayed on a computer screen. The comparison of the gestures was done in terms of measuring

- Muscle fatigue/effort
- Frustration, satisfaction and enjoyment
- Learnability of in-air gesturing, as a measure of the time component was also measured

2.1 Hypothesis

We hypothesized that in-air gesturing would be preferred to a touchscreen for interacting with a computer and that users would easily learn to use in-air gesturing during the experimental period

2.2 Participants

Thirty-two participants (SJSU) students taking the course Psych 1 and a few volunteers) were recruited to perform the tasks for this study. The participant pool was coordinated with the SJSU Psychology Department. The mean age of the participants was 20 years old and ranged from 18-29 years old. Fourteen participants were male and eighteen were female. Recruitment of the participants was entirely voluntary and scheduling was done online using SONA (human-subject pool management software).

Participants with active musculoskeletal disorders were excluded. This information was elicited by asking the participant about any disorders. All participants, except two were right handed. These two were ambidextrous and conducted the experiment using their right hand. All participants had used a smart phone or tablet with a touchscreen for at least one month.

The study was approved by the SJSU Institutional Research Board (IRB). A consent form, a photo consent form and an NDA was signed by each participant before beginning the testing session.

2.3 Apparatus and Instrumentation

A Dell AIO with an 3rd generation Intel Core i7-3770S processor 3.10 GHz with Turbo Boost 2.0 up to 3.90 GHz configured with 8GB Dual Channel DDR3 SDRAM at 1600MHz was used to conduct the study. Its' display was a 27.0" diagonal widescreen native resolution (FHD) with tilt base and a Touchscreen with HD support.

The system ran software that emulated the Windows 8 64 bit (Metro) home screen in English. Surface EMG sensors and software provided by Biometrics Ltd. (http://www.biometricsltd.com) was utilized. A range of surface EMG pre amplifiers was used with either the Biometrics DataLink DLK900 or DataLOG P3X8 for monitoring, storing and analyzing muscle electrical activity.

2.4 Procedure

Learnability Section: Task 1. Learnability section: Task 1 All participants completed a learnability task. This task helped familiarize the participant with the equipment (interfaces) and the gestures used to perform the task. This task helped determine if the gestures were easy to learn and remember.

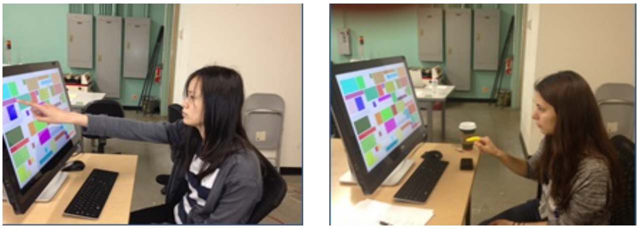


Fig. 1. Tile selection using touch screen and in-air gesturing

The participant was seated upright (back firmly against backrest) on a comfortable chair with armrests. Armrest height, seat height and distance from the screen were adjusted so to be consistent relative to each participant's body size and reach. In preparation for the in-air gesturing, the participant wore a yellow tag on his right index finger. This helped the recognition algorithm detect the finger for in-air gesturing more robustly.

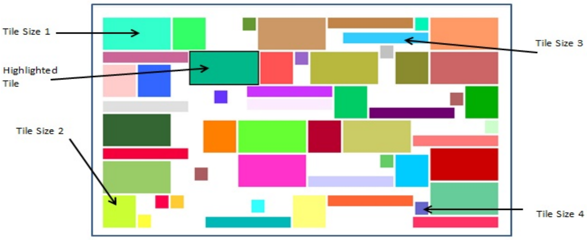


Fig. 2. Input screen with varying tile sizes

In this task, tiles lighted up in a pre-determined order every three seconds and participants selected the highlighted tile. The screen consisted of a collage of "Metropolitan" like tiles of four different sizes. The sizes were 310 x 150 pixels - rectangle shaped tile, 150 x 150 pixels - square shaped tile, 390 x 150 pixels - rectangle shaped tile and 60 x 60 pixels - square shaped tile. The first two sizes were the native Windows 8 desktop icons. The third size was from an email client, a highly used application. The last size was one of the smaller sized tiles prevalently used in Windows 8.

Tiles were separated uniformly by a 10 pixel gutter. Tiles were highlighted in a predetermined fashion every three seconds. The colors were randomized.

Tiles of any size would highlight by showing a black blinking border around the tile. This task was repeated for both the touchscreen and in-air interface. Participants tapped the screen in a touchscreen interface and moved a finger in free space for in-air gesturing to perform a "selection gesture". The selection gesture was a "Hold to Click" gesture, where the pointer controlled by the finger was held motionless for about 1.5 seconds on a tile to indicate selection. Less than the 1.5 second hold would result in unsuccessful selection of the tile. On selection of a tile, a graphic was displayed on the software to provide selection feedback. The tile remained highlighted until successful selection of the tile was complete, after which the next tile was highlighted. The hand moved from the resting position (which is the position where the participant is comfortably seated, with no hand lifted up) to the relevant point of selection. The participants selected a total of 20 tiles during the task. The software running the task measured the following factor:

Duration: Response time from the point of tile highlight to selection. Questionnaires were administered to elicit subjective data about the experience for the touchscreen and in-air gesturing interface.

EMG Setup. The surface EMG transducers were placed parallel to the muscle fiber at three locations on the dominant side of the body as in Figure 3. They are the Upper trapezius, Anterior deltoid and Extensor Digitorum (the center of the dominant posterior forearm at approximately 30% of the distance from the elbow to the wrist). The muscle was palpated to detect the exact point of muscle activity when the participant extended his fingers. The ground electrode was connected to the left ankle.

Maximal Voluntary Electrical (MVE) activation measurements were performed against manual resistance to normalize the EMG signals from each location.

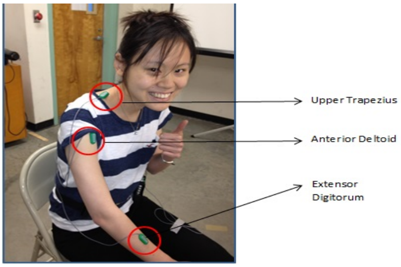


Fig. 3. EMG Transducers fixed to the 3 positions in the body

Task 2. Next, the participant performed Task 2. The task and setting was similar to task 1 but with a different tile layout. Here tiles of a particular size alone were highlighted each time. The task was performed four times, once for each size. Each task took approximately 1-2 minutes. EMG data was recorded for each task. The order of the tasks was randomized to reduce order effects.

Task 3. Next, Task 1 of the experiment was repeated. The duration was measured and compared with the initial session. The comparison helped us understand to what extent learning happened. The same questionnaires were administered once again and later analyzed for any change in subjective measures. Subjective ratings on discomfort and ease of use of the touchscreen and in-air gesturing interface was elicited by means of self-report questionnaires. This was done at the end of task 1 and 3. The questionnaires consisted of check boxes, semantic differential scales and open ended questions. The semantic differential scales used 7 points ranging from very high to very low with a center point of neutral stance.

3 Analysis of Data

3.1 Learnability Task

The time taken to perform the Learnability Task 1 and Learnability Task 2 for in-air gesturing alone were compared. Of 32 participants, 26 showed an improvement in speed in the second session. That made up about 81.25% of the participants. A paired samples T-test was conducted to compare the mean differences between the times taken for the two learnability sessions for in-air gesturing alone. The mean time for Learnability 1 was 142.40 seconds while the mean time for Learnability 2 was only 128.36 seconds. The mean difference was found to be statistically significant, M=14.04, SD= 16.82, t (31) =4.722, p<0.05, Cohen's d=0.83. This shows a large effect in the mean difference. When 1.5 seconds of hold to click time and three seconds between highlights (82.5 seconds) was reduced from each participant's time, we found a statistically significant result with the same t and p values.

3.2 EMG Setup

The participants were subjected to four trials with each one of the four tile sizes. There were eight tiles in each category in all of the permutations. For each participant, six values were obtained which were the average value for Upper Trapezius, Anterior Deltoid and Extensor Digitorium for touchscreen and in-air gesturing.

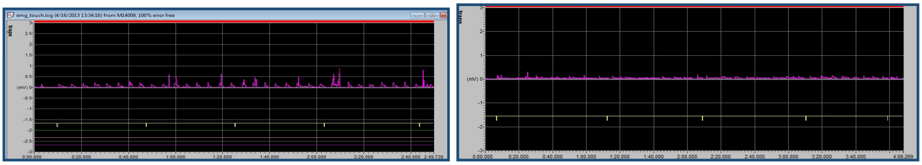


Fig. 4. Filtered Signals for Upper Trapezius for touch screen and in-air gesturing

Figure 4 shows the values for participant 18's upper trapezius values for a touchscreen and in-air gesturing after application of filters. The spikes show activity in the upper trapezius as it moved to select a tile. We barely see any activity in the second graph, showing that in-air gesturing requires less effort when it comes to the upper trapezius. Similarly we saw barely see any activity showing that in-air gesturing requires less effort for the anterior deltoid. The spikes in the touchscreen were attributed to every time the participant stretches out his arm to touch the screen. For the Extensor Digitorum, some activity was seen with the in-air gesturing when compared to touch screen. The spikes in the touchscreen were attributed to every time the participant closes the wrist to point to the screen.

Three Repeated measure ANOVAs were conducted to compare the muscle effort of the Upper Trapezius, Anterior deltoid and Extensor Digitorum immaterial of the tile size. Significant results were obtained for Upper Trapezius and Anterior deltoid. Very small significance was obtained for Extensor Digitorum.

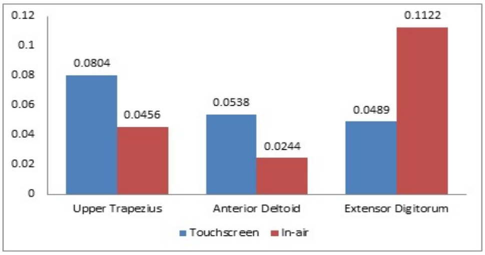


Fig. 5. Comparing means values for the 3 muscle points for touchscreen and in-air gesturing

Further ANOVA analysis showed no statistically significant difference between the four tile sizes for touchscreens. There was statistical significance between the four tile sizes for in-air gesturing. It was found that tile size 4 was the most difficult to manipulate in in-air gesturing.

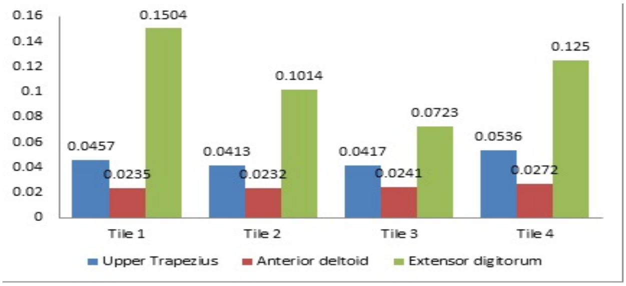


Fig. 6. Comparison of mean values for Tile sizes versus Muscle point for In-air gesturing

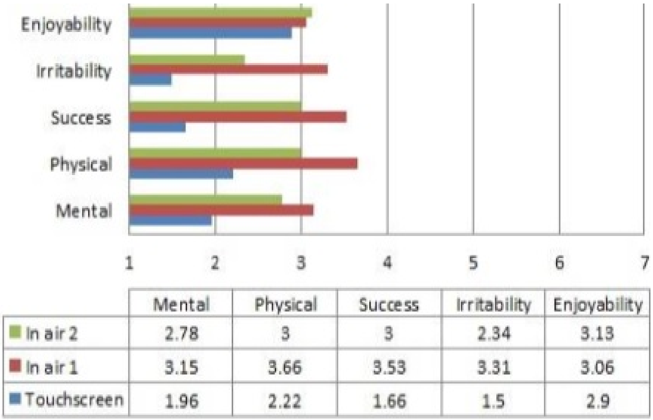
3.3 Subjective Questionnaire Analysis

Familiarity with In-Air Gesturing: To begin with, 10 out of 32 participants were familiar with in-air gesturing either through Xbox Kinect etc. while 22 were unfamiliar. About 69% of the participants were unfamiliar with in- air gesturing

Interface Preference: Of the 30 participants, 27 preferred touchscreen at the end of both learnability sessions. Three participants preferred in-air gesturing to begin with. Two participants changed their preference from touchscreen to in-air gesturing. One participant changed preference from in air to touchscreen. One participant's preference with respect to in-air gesturing remained the same.

Analysis of Individual Questions: Touchscreen data reported represents data recorded after the first learnability session. In-air data represents data from both learnability sessions. All data reported is an average of the individual ratings given by the 32 participants.

Figure 7 compares the values for 5 factors. The scale defined 1=low and 7=high. Touchscreen generally reported the best value. In-air gesturing after the second learnability session reported better values than the first.



	Mental	Physical	Success	Irritability	Enjoyability
🟩 In air 2	2.78	3	3	2.34	3.13
🟥 In air 1	3.15	3.66	3.53	3.31	3.06
🟦 Touchscreen	1.96	2.22	1.66	1.5	2.9

Fig. 7. Comparing above 5 questions for touchscreen, in- air gesturing session 1 and in-air gesturing session 2

No participant felt silly or embarrassed to use the touchscreen. Nine participants felt so using in-air gesturing after the first session. The number fell to four after the second session. For in-air gesturing, the degree of embarrassment was 4 (around mean) after the first session, but fell to a low 2.75 after the second session.

Participants found in-air gesturing initially easier in the first session at a value of 2.84 than in the second session of in-air gesturing, with a value of 3.13

All 32 participants said they would use the touchscreen in a public place. 9 said no to in-air gesturing after the first session, which came down to 7 after the second session. 2 participants changed their preference to yes. Among the 9 participants in the first session, 6 found it silly to use in-air gesturing. 3 were ready to use it in a public place even though they found it silly. After the second session, only 2 out of the 7 found it silly to use in-air gesturing. The number increased to 5 for those people who found it silly but still would use it in a public place.

After the first session, 5 participants didn't want to own a device, while after the second session, the number increased to 2. 5 participants who found in-air gesturing silly, wanted to own a device after the first session. After the second session, 3 who found it silly wanted to own a device. Further, 2 people who found it silly and didn't want to use in-air gesturing in a public place, still wanted to own a device. The various reasons people wanted and didn't want to own a device were multifold. The number of reasons to own a device outnumbered the ones that were against owning a device as seen in Table 1.

Table 1. Reasons quoted verbatim

I want a device	I don't want a device
It's Cool	It will make me self-conscious
Fun and exciting	Too unreliable
For development purposes	Touchscreen is more efficient
Support new technology	Hard to use
Enjoyable for gaming	Hard to control
Easy to use	In accurate
Makes life more efficient	Feel no need for gesturing
Fun and less work for the shoulder	
People get more exercise while using the Kinect kind of things	
For curiosity	
I will eventually get used to it	
Use when hands are not free	
I will use it to create my own gestures	
Fun to do something at a distance than up close	
Don't want to do the extra work in touch-screens	
Only in situations where physical touch is not possible	

4 Discussion

The primary goal of the study was to elicit preference between two interfaces, the touchscreen and in-air gesturing. A secondary goal of the study was to understand the learnability of in-air gesturing as it is a new and upcoming technology, especially given its limitations.

EMG recordings very clearly showed that in-air gesturing was a more ergonomic methodology of interacting with the computer when compared to touchscreen. Statistically significant results were found for two of the critical muscle points, the Upper Trapezius and Anterior deltoid. EMG recordings also showed that muscle effort increased significantly when the size of the target decreased. Among the 4 tile sizes uses, tile size 4 was found most difficult to select during in-air gesturing and differed statistically significantly from the other 3 sizes.

The experiment session lasted for about 1 hour 15 minutes, approximating about 1 hour of time between the first learnability and second learnability sessions. That accounts for a total of 8 minutes maximum of in-air gesturing. It was found that there was a statistically significant improvement in the time taken to perform the two similar sessions. Mean value of the time taken decreased by 14.04 seconds. This was found to be a large effect. Note that 69% of the participants were using in-air gesturing for the first time in their lives during the experiment. This shows that significant learnability happened over a period in in-air gesturing within an hour of time.

Subjective questionnaire analysis showed a similar trend throughout. Touchscreen always rated better for almost all the questions over in-air gesturing. But between the two sessions of in-air gesturing, ratings after the second session were always found better than the first. It is evident that with time and practice, in-air gesturing is comparable to touchscreen eventually for almost all the factors. The only factor that saw a higher rating in the second session was the "ease of use" of in-air gesturing.

Majority of participants did not find it silly to use in-air gesturing and were ready to use it in a public place. There were some conflicting answers such as, some participants who found it silly, were ready to use it in a public place. Some participants who found it silly and were not ready to use it in a public place still wanted to own a device. Majority of participants wanted to own a device capable of in-air gesturing. The most popular reason was because they found it cool, among various other relevant reasons.

According to Harrison's definition, the gesture does not directly indicate its intent because in real life, "hold to click" does not indicate selection [6]. It is interesting to note that though this gesture is not very intuitive for selection purposes, the learning curve was found to be very easy and showed statistical significance. This goes against the literature that claims that it is the naturalness and intuitiveness of a gesture that defines the learning curve.

We learnt in this experiment that "social acceptability" does play a role. But this experiment has also shown that this self-consciousness of people actually fades away with time and people would want to use a device because in-air gesturing is considered more technologically advanced. It overrides the social taboo of in-air gesturing.

5 Conclusion

In-air gesturing definitely emerged as a winner during the period of the experiment. Participants clearly showed that it was easy to learn the technique of interacting with the interface and the subjective attributes were comparable to that of the current reigning touchscreen with the passage of time. The reasons why participants preferred in-air gesturing outnumbered the reasons why participants preferred the touchscreen and the reasons were variant and spread across a large spectrum. In-air gesturing emerged as the winner ergonomically when compared to touchscreen. This experiment has shown what Steve Jobs once quoted that touchscreen computers are "ergonomically terrible". It has also been shown that participants prefer the in-air gesturing to touchscreen because of the coolness factor associated with new technology.

Acknowledgements. My special thanks to Dr. Louis Freund and Dan Rosenberg of San Jose State University whose guidance remained from start to end of the experiment. My thanks also go to the employees of the startup who made possible the experiment. I would like to acknowledge Joyce Kii and all the participants who made valuable contributions to the research. I would also like to thank my family who have been my stronghold through this journey.

References

1. Acharya, T., Mitra, S.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(3), 311–324 (2007)
2. Atia, A., Takahashi, S., Tanaka, J.: Smart gesture sticker: Smart hand gestures profiles for daily objects interaction. In: Proceedings of: 9th IEEE/ACIS International Conference on Computer and Information Science, IEEE/ACIS ICIS 2010, Yamagata, Japan, August 18-20 (2010)
3. Baudel, T., Beaudouin-Lafon, M.: Charade: Remote control of objects using free-hand gestures. Communications of the ACM – Special Issue on Computer Augmented Environments, Back to the Real World 36(7) (1993)
4. Bowman, D., McMahan, R., Ragan, E.: Questioning naturalism in 3D user interfaces. Communications of the ACM 55(9), 78–88 (2012)
5. Christova, P., Kossev, A., Kristev, I., Chichov, V.: Surface EMG recorded bybranched electrodes during sustained muscle activity. J. Electromyogr Kinesiol 9, 263–276 (1999)
6. Harrison, C.: Meaningful gestures, http://www.economist.com/node/21548486 (retrieved March 3, 2012)
7. Kendon, A.: Gesture: Visible action as utterance, pp. 326–355. Cambridge University Press, Cambridge (2004)
8. Kita, S.: Theoretical issues in nonverbal behaviors. Presentation Slides (2007), retrieved from http://ling75.arts.ubc.ca/cogs//cogs401
9. Kita, S., Essegbey, J.: Pointing left in Ghana: How a taboo on the use of left hand influences gestural practice. Gesture 1(1), 73–95 (2001)
10. Kita, S., Danzinger, E., Stolz, C.: Cultural Specificity of Spatial Schemas manifested in spontaneous gestures. MIT Press, Cambridge (2001)
11. McQuade, K., Dawson, J., Smidt, G.: Scapulothoracic muscle fatigue associated with alterations in scapulohumeral rhythm kinematicsduring maximum resistive shoulder elevation. JOSPT 28(2), 74–80 (1998)
12. Morita, H., Hashimoto, S., Ohteru, S.: A Computer Music System that Follows a Human Conductor. IEEE Computer, 44–53 (July 1991)
13. Perzanowski., D., Schultz, A., Adams, W., Marsh, E., Bugajska, M.: Building a multimodal Human-Robot interface (2001)
14. Rico, J., Brewster, S.: Usable gestures for mobile interfaces: Evaluating social acceptability. In: Proceedings of CHI 2010, pp. 887–896 (2010)
15. Ronkainen, S., Hakkila, J., Kaleva, S., Colley, A., Linjama, J.: Tap input as an embedded interaction method for mobile devices. In: Proceedings of TEI 2007, pp. 263–270. ACM Press (2007)
16. Sturman, D.: Whole-hand input, Ph.D thesis, Media Arts & Sciences. MIT Press (1992)
17. Weissman, C., Freeman, W.: Television control by hand gestures. In: IEEE Intl Workshop on Automatic Face and Gesture Recognition (June 1994)
18. Wigdor, D., Wixon, D.: Brave NUI World, 1st edn. Morgan Kaufmann, Burlington (2011)

Beyond Presentation - Employing Proactive Intelligent Agents as Social Catalysts

Madlen Wuttke[1] and Michael Heidt[2]

[1] Research Training Group crossWorlds, Faculty of Humanities, Chemnitz University of Technology, Thüringer Weg 5, 09126 Chemnitz, Germany
madlen.wuttke@phil.tu-chemnitz.de

[2] Research Training Group crossWorlds, Faculty of Informatics, Chemnitz University of Technology, Thüringer Weg 5, 09126 Chemnitz, Germany
michael.heidt@informatik.tu-chemnitz.de

Abstract. Despite long standing attention from research communities, the technology of intelligent agents still harbours a large amount of unrealised potential. In this text, we argue that agent technology can benefit from a shift in focus from presentation to possible functionalities. In doing this, our focus is on the provision of pro-activity: The ability of agents not to merely react but to predictively shape their environments. In order to illustrate our arguments, we present an instance of interactive technology, showing how pro-active intelligent agents can be employed in exhibition contexts.

1 Introduction

Scientific research regarding pedagogical agents has mainly been focused on analysing different forms of their depiction, rather than possible features. For example, numerous studies have analysed whether an agent should be designed as either male or female [13] or whether or not an agent should be displayed as realistic as possible, including facial animations [1]. In addition to this, Lusk et al. [18] tested if there was a positive effect on learning with an animated or a static agent and Baylor et al. [2] as well as Huang et al. [10] hypothesise a beneficial effect on learning as long as the agent is depicting one's own peer-group and ethnicity.

Regarding the features of an agent-system however, the focus is largely about establishing behavioural strategies. These focus on questions such as if agents are to be depicted as either polite or rude [25], whether the implementation of gestures and mimic behaviour changes the acquisition of learning material [6] or if social conversations, not touching on the topic itself, help to create a positive learning environment [23]. What all these research projects have in common is the tendency to analyse passive features of an agent. But what appears to be missing from empirical discourse is research regarding active components of pedagogical agent designs such as active listening, observation of real-world surroundings and just-in-time information aggregation. Those active components would allow for an agent to analyse the environment and to react pro-actively to changes in it [26] as well as acting on behalf of the user.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 182–190, 2014.

In order to demonstrate the role pro-active agents are able to play with respect to design of interactive technologies, we discuss a series of design prototypes developed. These have been implemented in various degrees of fidelity, ranging from paper-prototypes to mid-fidelity digital artefacts. The devices conceived are targeted at the museum domain. Their goal is to strike up verbal interaction between previously un-acquainted museum visitors. Embedded in the wider scope of a design ecology [7], the system comprises mobile components as well as a stationary wall mounted installation. The stationary setup is equipped with depth cameras used for monitoring of users. Museum visitors are provided with tablets which replace traditional printed museum documentation. On these tablets a personalised instance of an intelligent agent is presented. This agent acts in the capacity of a museum docent, providing both additional information as well as helpful incentives regarding the possibilities of the exhibition visited.

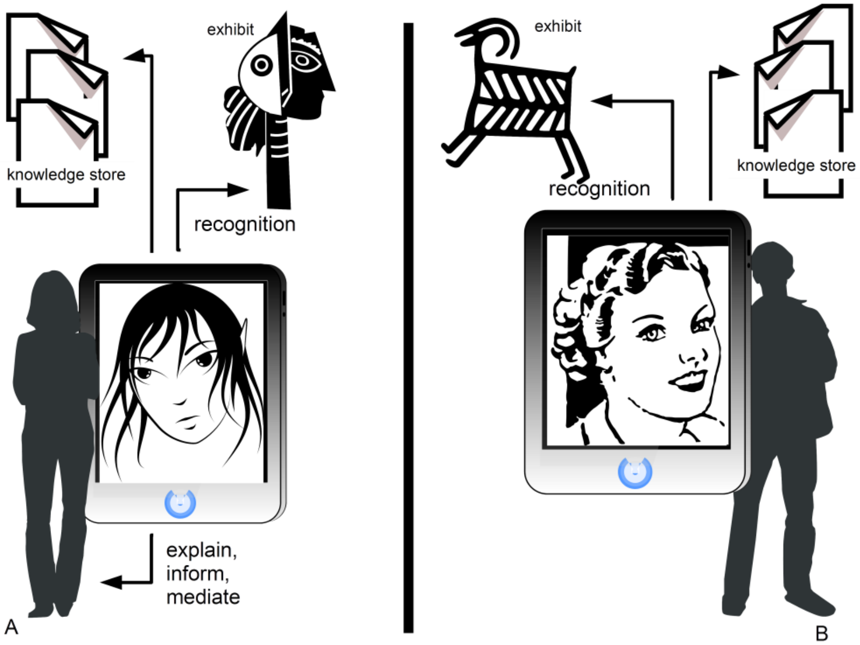


Fig. 1. Exploration phase: Agents on mobile devices

The system's main functionality is localised at the wall mounted installation. It serves in analogy to information plaques, displaying personalized multimedia content. Designed to accommodate two visitors at the same time, its screen setup realizes a split-screen configuration. When users approach the display, respective individual agents migrate from the tablet into the stationary screen space, taking up position at the left and right periphery of the screen.

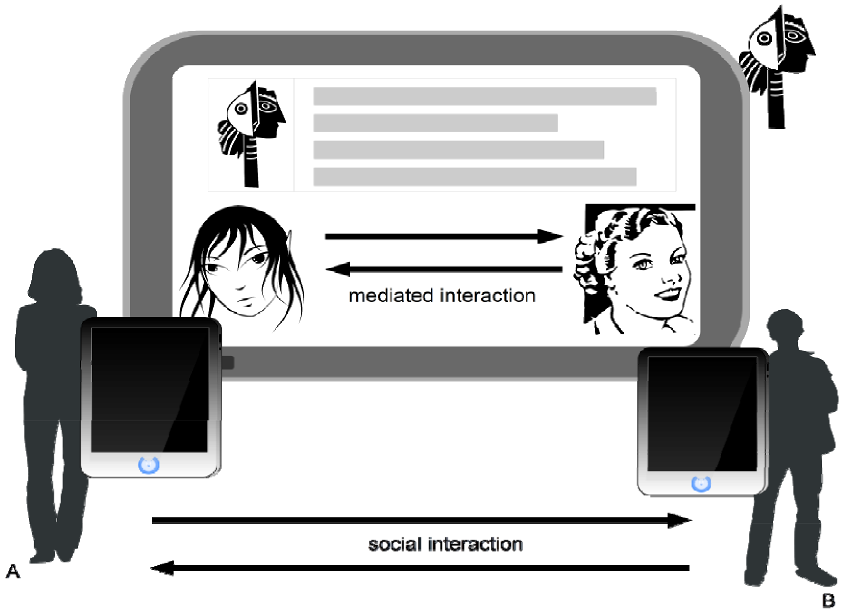


Fig. 2. Fig. 3. Wall-mounted display: Agents as facilitators of social interaction

Whenever two visitors use the station concurrently, this marks the critical part of system operation. User monitoring is employed in order to assess if users are orientated towards each other communicatively or not. Should the system infer communicative interest, an attempt is made to connect both visitors by supplying a communicative incentive. This is provided as follows: Individual agents situated within screen space leave their position at the periphery of the screen and meet at the lower centre. Here, they engage in pseudo-social interaction with one another. Hereby an element of surprise is provided, acting as a helpful catalyst for interaction in exhibition spaces [14]. The intended effect is for the agents' owners to react to the surprising behaviour of their "virtual pets", ideally by engaging in direct discourse with one another. The situation is constructed in analogy to phenomena such as dog-owners striking up conversations posterior to a meeting of the animals they were walking.

A crucial part of system operation lies in judging if respective users are communicatively inclined during the critical phase. To this end, we intend to employ Hidden Markov Models trained with manually annotated data sets. Among the markers to be analysed are eye-movement behaviour, body posture as well as complementing proxemics features [19].

2 Museum Scenario

The depicted scenario consists of a two level system. One is installed on a device which is handed out to the visitors at the beginning of their tour while the other, the main program, is located on a server which is administrating the informational database and which is able to initiate crossovers between the interests of individual visitors. In addition, the exhibits have explanatory screens which present additional information like the basic description, usage, relevance up to videos of seeing the object in context of, for example, everyday life or whatever purpose.

Following the extensive examples of Lieberman and Selker [16] regarding an agent's depiction and usage as a helpful tool inside a virtual environment, the primary directionality of the aspired social catalyst would be that of an 'advisor' instead of being an 'assistant'.

Although in later steps it might be necessary to not keep this explicit distinction, we employ it here for the sake of conceptual clarity. Once stepping into the museum the aforementioned tablets are handed out together with the admittance ticket. The tablet would ideally be a small one in the range of current 7" display sizes in order to be able to keep it in one hand or to easily store it in a pocket.

The screen would be populated by applications like an in-door positioning system, providing for an accurate 'you-are-here'-button at all times. Additionally, there would be different routes presented, available by pushing a button on the side of the screen.

This would allow to find:

- the nearest exit and other points of interest (coffee spots, sanitary installations, phone spots, souvenirs etc.)
- an information officer, a real human 'agent' to talk to and help in case of any problems with the device or the exhibition
- a personalised route through the exhibition, perhaps even planned ahead from home

In addition to the device being able to plot routes, the system would be represented by an embodied agent being able to react and offer conversational topics regarding the museum and presented objects. A conversational database in the background would continually track the user's interaction with the agent and compare it to other visitor's inquiries. Due to this, the administrating program can check for similar requests to the database and proximity of visitors based on their location inside the museum. It could provide access to personal information about the visitors via their social network connections.

Based on those two cornerstones of information, the system would engage visitors in a conversation by pre-structuring conversations towards similar interests. Once two devices and their associated visitors converge to a distance which would allow for a regular volume speaking voice, the agents initiate their social catalyst routine. This would happen in three steps.

1. The agents individually acknowledge each other and their respective individual or group to the person using the device. This happens by virtue of a visible turn of the agent towards the other one.
2. A user then has the option to either confirm the upcoming interaction or deny it, resulting in a courteous discontinuation of the initiated process. As soon as one party

denies, the respective agents would suggest politely continuing the tour at another point of the exhibition – further away from the other group.

3. If the interaction is confirmed by both parties, the agents visually leave the tablet space and appear on the screen in front of the exhibition.

At first the agents start to interact with each other, which enables them to get the respective parties up to speed about their individual pathways through the exhibition. Afterwards, the significance of the current exhibit would be explained in the context of the whole exhibition and conversational pointers engage the humans in front of the screen to interact with each other.

If the catalyst worked and the visitors continue their journey through the museum together, the agents continue to provide conversational incentives by engaging the humans and each other to keep the discussion going.

If it did not work as intended, then the agents return to their previous state as an informational advisor about the exhibit.

3 Research Directions

The aforementioned scenario and requirements provide numerous research opportunities. A conversational agent is already very well researched as well as regarding agents working in groups as companions [12]. But due to the necessity of implementing new ways of interconnected databases and agent's behaviour, the need for an interdisciplinary approach is obvious. Social sciences for analyzing and categorizing human conversational behaviour and information sciences for implementing the software infrastructure for the agent's behaviour and the administrating instance governing the database.

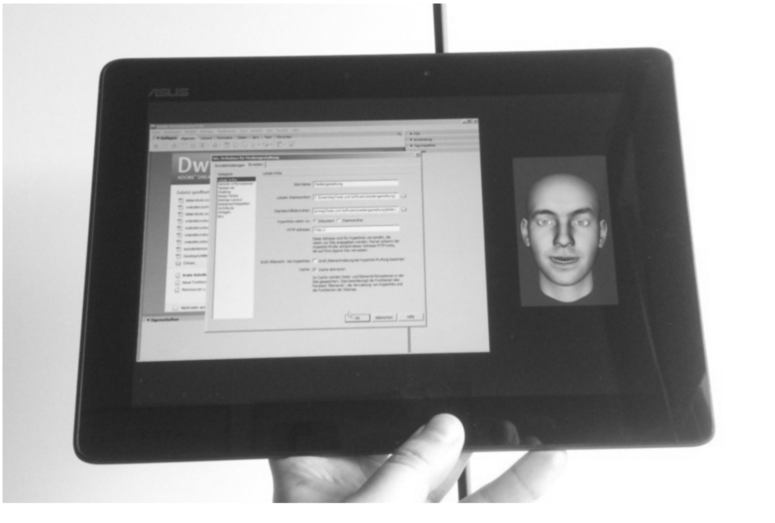


Fig. 4. Tablet with webcam and microphone and pedagogical agent capable of pro-actively reacting to environmental disturbances while transferring knowledge

The goal has to be the implementation of a pro-active conversational agent which is capable of gathering various forms of input. As mentioned before, numerical statistics are able to elicit certain behaviour, like in the case of proximity to another agent system. Environmental information like shambles in the vicinity or auditory superimpositions which would hinder a conversation, can be used to either get away from such incidents or avoid plotting through such areas beforehand. These behaviours would be in accordance with the postulations of Lieberman and Selker [17] as they urge to enable computer systems to be able to grasp the context of a situation.

Other pro-active components are facial interpretations by camera systems, fingertip temperature through sensors on the surface of the tablet, gait information, body posture as well as gaze and eye tracking. These person centred information can be used to indicate a user's current emotional, vigilance and inquisitiveness state. The cognition of emotional states via facial action recognition, as shown by Kapoor, Qi and Picard [11], provide a reliable prognosis of human reactions to certain events. While Breazeal [5] developed a humanoid robot's emotional model which is able to register affective intents based on a user's voice.

Regarding the interaction between two users, the system should be able to 'read' users reactions to the initiation process. Even for humans this is not an easy task since nonverbal cues are often polysemantic. To adequately register the emotion, context is once more of relevance to the process. As stated by Olsson and Ochsner [21] the prior experiences with the persons become relevant which ideally have been tracked by the agent system along the way up to the point of becoming acquainted with the other visitor.

Regarding the depiction of the agent, empirical research postulates an agent to be able to act socially intelligent. Meaning it knows about cultural peculiarities and possesses the ability to detect and act on them. It has to be perceived as being polite [15, 22, 25] which also extends to the choice of clothing and grooming. The user should be offered a choice of agent representations since learning from a representative of one's own peer group seems to be beneficial [12, 13, 20]. If the agent is equipped with a voice, then the choice of words and tone of the voice should be polite as well but also it must not be identifiable as a text-to-speech software. Although immense progress has been visible over the last decade, it still is not comparable to a human voice, which might even be more important than appearances [23, 24].

Facial animations of the agent seem to be an issue as well. Static agents still transfer information but some studies [1, 4, 18] postulate a positive effect on motivation, retention and transference of learning material. Gestures and body postures of an agent however apparently do not have a positive effect. Once added to an already facially acting agent, the gestures either showed no [6] or even hindering effects [3].

4 Experiential Structure

Design efforts are guided by a three-partite construal [9] of the museum experience:

Prior to their visit, users become aware of the museum and the possibilities contained within it. During this phase, users utilise information offered in order to decide which institution to visit.

During their visit, users interact with installations, and with each other. After the visit, users possibly relive experiences made and reflect on knowledge gained.

Following this structure, interactive artefacts as well are grouped into three interaction ecologies [8]:

- A web and app-based ecology, allowing for information about the museum to be gathered.
- A spatially structured ecology within the museum, allowing for incentives to be generated in the context of interactive installations and mobile devices.
- A web and app-based ecology, allowing for additional information on exhibits to be obtained and for furthering of social contacts made.

Likewise, distinct content presentation strategies are adopted in order to address different requirements during the phases. I.e. consumption of time-based media potentially creates problems within a museum setting, running the danger of distracting visitors from the experientially rich environment around them.

However, watching videos or listening to historical recordings can be a useful activity during a train ride antecedent to the actual visit. They refer to experiences already made while prolonging the possibility to exist within the historical space encountered.

Agents provide an experiential tie between all three phases. They can be gently introduced in the first phase, provide helpful incentives during the second and act as gentle reminders in the last one.

5 Conclusion

We have argued for a shift in focus from presentation to functionality within interactive agent research. Numerous scenarios exist where proactively behaving agents could be beneficial. We detailed one such scenario within the domain of interactive installations in museums.

The discussion points to a broader issue. Refocusing agent research onto the level of functionality forces us to reopen the design space. Many of the tacit assumptions present within existing discourse surrounding pedagogical agents have to be reexamined. This will provide both for new possibilities while creating new challenges for the agent research community.

References

1. Agada, R., Yan, J.: Research to Improve Communication by Animated Pedagogical Agents. J. Gener. Inf. Technol. 3, 1 (2012)
2. Baylor, A.L., et al.: Interface Agents As Social Models: The Impact of Appearance on Females' Attitude Toward Engineering. In: CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 526–531. ACM, New York (2006)
3. Baylor, A.L., Kim, S.: Designing nonverbal communication for pedagogical agents: When less is more. Comput. Hum. Behav. 25(2), 450–457 (2009)

4. Baylor, A.L., Ryu, J.: The Effects of Image and Animation in Enhancing Pedagogical Agent Persona. J. Educ. Comput. Res. 28(4), 373–394 (2003)
5. Breazeal, C.: Emotion and sociable humanoid robots. Int. J. Hum.-Comput. Stud. 59(1-2), 119–155 (2003)
6. Craig, S.D., et al.: Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features and redundancy. J. Educ. Psychol. 94(2), 428–434 (2002)
7. Heath, C., et al.: Crafting participation: Designing ecologies, configuring experience. Vis. Commun. 1(1), 9–33 (2002)
8. Heidt, M., Kanellopoulos, K., Pfeiffer, L., Rosenthal, P.: Diverse Ecologies – Interdisciplinary Development for Cultural Education. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part IV. LNCS, vol. 8120, pp. 539–546. Springer, Heidelberg (2013)
9. Heidt, M.: Examining Interdisciplinary Prototyping in the Context of Cultural Communication. In: Marcus, A. (ed.) DUXU 2013, Part II. LNCS, vol. 8013, pp. 54–61. Springer, Heidelberg (2013)
10. Huang, H.-H., et al.: Toward a multi-culture adaptive virtual tour guide agent with a modular approach. AI Soc. 24(3), 225–235 (2009)
11. Kapoor, A., et al.: Fully automatic upper facial action recognition. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, AMFG 2003, pp. 195–202 (2003)
12. Kim, Y., et al.: Pedagogical agents as learning companions: The impact of agent emotion and gender. J. Comput. Assist. Learn. 23(3), 220–234 (2007)
13. Kim, Y., Wei, Q.: The impact of learner attributes and learner choice in an agent-based environment. Comput. Educ. 56(2), 505–514 (2011)
14. Vom Lehn, D., et al.: Exhibiting Interaction: Conduct and Collaboration in Museums and Galleries. Symb. Interact. 24(2), 189–216 (2001)
15. Lepper, M.R., et al.: Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. Comput. Cogn. Tools. 1993, 75–105 (1993)
16. Lieberman, H., Selker, T.: Agents for the user interface. Handb. Agent Technol., 1–21 (2003)
17. Lieberman, H., Selker, T.: Out of context: Computer systems that adapt to, and learn from, context. IBM Syst. J. 39(3.4), 617–632 (2000)
18. Lusk, M.M., Atkinson, R.K.: Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? Appl. Cogn. Psychol. 21(6), 747–764 (2007)
19. Mead, R., et al.: Recognition of spatial dynamics for predicting social interaction. In: Proceedings of the 6th International Conference on Human-Robot Interaction, pp. 201–202. ACM, New York (2011)
20. Moreno, R., Flowerday, T.: Students' choice of animated pedagogical agents in science learning: A test of the similarity-attraction hypothesis on gender and ethnicity. Contemp. Educ. Psychol. 31(2), 186–207 (2006)
21. Olsson, A., Ochsner, K.N.: The role of social cognition in emotion. Trends Cogn. Sci. 12(2), 65–71 (2008)
22. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, New York (1998)

23. Veletsianos, G.: How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences. Comput. Hum. Behav. 28(1), 275–283 (2012)
24. Veletsianos, G.: The impact and implications of virtual character expressiveness on learning and agent–learner interactions. J. Comput. Assist. Learn. 25(4), 345–357 (2009)
25. Wang, N., et al.: The politeness effect: Pedagogical agents and learning outcomes. Int. J. Hum.-Comput. Stud. 66(2), 98–112 (2008)
26. Wuttke, M.: Pro-Active Pedagogical Agents. Int. Summerworkshop Comput. Sci. 2013(17), 59 (2013)

A Method for Lifelong Gesture Learning Based on Growing Neural Gas

Paul M. Yanik[1], Anthony L. Threatt[2], Jessica Merino[2], Joe Manganelli[3], Johnell O. Brooks[4], Keith E. Green[3], and Ian D. Walker[2]

[1] Department of Engineering and Technology,
Western Carolina University, Cullowhee, NC 28723, USA
[2] Department of Electrical and Computer Engineering
[3] School of Architecture
[4] Department of Automotive Engineering
Clemson University, Clemson, SC 29634, USA
pyanik@wcu.edu,
{anthont,jmerino,manganelli,jobrook,kegreen,iwalker}@clemson.edu

Abstract. Gesture-based interfaces offer the possibility of an intuitive command language for assistive robotics and ubiquitous computing. As an individual's health changes with age, their ability to consistently perform standard gestures may decrease, particularly towards the end of life. Thus, such interfaces will need to be capable of learning commands which are not choreographed ahead of time by the system designers. This circumstance illustrates the need for a system which engages in *lifelong learning* and is capable of discerning new gestures and the user's desired response to them. This paper describes an innovative approach to lifelong learning based on clustered gesture representations identified through the Growing Neural Gas algorithm. The simulated approach utilizes a user-generated reward signal to progressively refine the response of an assistive robot toward a preferred goal configuration.

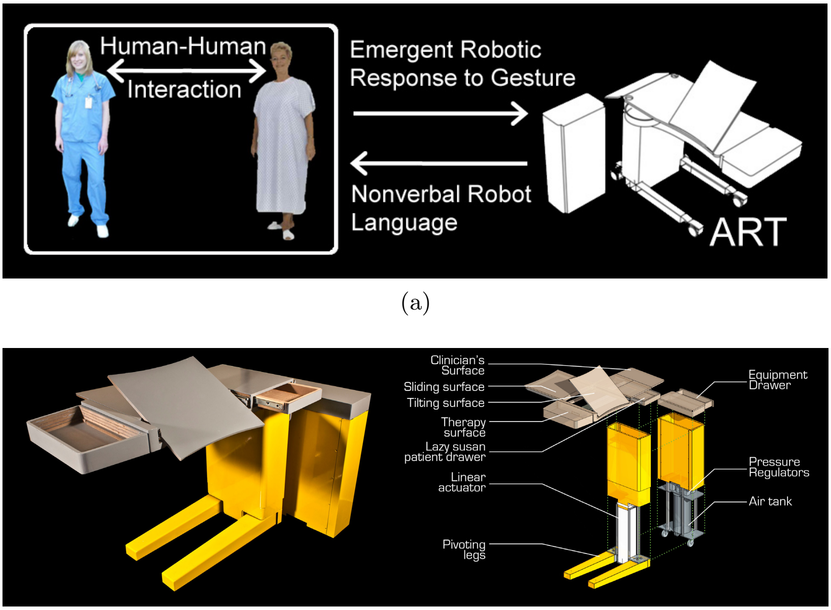
Keywords: machine learning, gesture recognition, human-robot interaction, assistive robotics.

1 Introduction

As the population ages, their desire to retain a level of independence in the face of diminished mobility and health will increasingly draw upon assistive technologies to facilitate essential Activities of Daily Living (ADLs). The work described in this paper is motivated by a dearth of technologies that might provide adequate support of these essential ADLs. Effective design, deployment, and use of such technologies are seen as critical to promoting an improved quality of life and prolonged independence for the user. The Assistive Robotic Table (ART) project begun at Clemson University seeks to develop an intelligent class of assistive devices and services which are highly integrated into the built environment. In so doing, the environment becomes an adaptive partner to facilitate *aging in place* for users whose ability levels are changing.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 191–202, 2014.

Non-verbal communication interfaces, and in particular, gesture-based interfaces offer the possibility of an intuitive command language for assistive robotics and ubiquitous computing. However, as an individual's health evolves with age, their ability to perform standard gestures consistently may decrease, particularly towards the end of life. The envisioned non-verbal communication loop between a user and the ART appliance (a robotic version of the standard over-the-bed table) is depicted in Fig. 1.



(a)

(b)

Fig. 1. (a) The non-verbal communication loop of the *Assistive Robotic Table*. The focus of this work is on the emergent (learned) response of this device to the user. (b) A recent project artifact.

In addition, for impaired or unskilled users, such interfaces will need to be capable of learning commands whose choreography is not strictly prescribed by the system designers. These circumstances illustrate the need for a system which engages in *lifelong learning* [1] and is capable of discerning new gestures and the user's desired response to them. The reported research targets the ART appliance and presents an approach which learns a user's preferred three-dimensional configuration of the appliance for tasks performed in a healthcare or home setting. Results are based on arm-scale gesture motions collected from human participants and interactions using a simulated human user which controls the application of a success indicator (reward) signal.

Extending past work by the authors [2], a system based on the Growing Neural Gas (GNG) algorithm [3] is used in this research to create an active mapping between performed gestures and robotic actuations. The proposed method takes advantage of the user's broad view of the problem space to selectively apply positive rewards where robot actions are tending toward the user's preferred goal configuration for a given gesture. Corrections in the form of negative rewards are similarly applied when the agent is diverging from the intended configuration.

Toward practical application with a live human user, a use/training model for the system is proposed which aims at reducing both the number of observations of a new gesture required to train ART to desired responses and the effort borne by the user in doing so. Thus, the success of the proposed approach is measured in terms of its speed of convergence to the user's preferred response in terms of decreasing numbers of cycles of observation and reward. Also, the ability of the approach to learn new information while retaining past knowledge is investigated.

This paper is structured as follows. Section 2 presents past research efforts in lifelong learning and describes their respective advances and shortcomings. Section 3 discusses specifics of the system design including data representations, algorithms and the simulation environment. Section 4 discusses the data collection fixture, and experimentation scenarios. Section 5 presents and interprets the experimental results. Finally, conclusions and future work are given in section 6.

2 Related Work

Often, the operational life of a learning system is divided into the distinct phases of learning versus recognition. This paradigm neglects the possibility that the system may need to acquire new recognition capabilities in the face of a changing input distribution from its environment. Conventionally, systems forced to consider new forms of input must reiterate the training phase. In so doing, they may suffer degradation in their ability to preserve knowledge acquired in the past. Thus, by extending their recognition capability, the stability of the system is compromised [1]. This problem is termed the *Stability-Plasticity Dilemma* [4]. Toward the development of a system which can acquire new gestures as the user requires, the need for lifelong learning is considered.

A variant of Kohonen's self-organizing feature map (SOFM) [5], GNG is capable of tracking a moving distribution, of adding new reference nodes, and of operating from static input parameters [6]. Given these qualities, GNG is well suited to the task of gesture recognition where no labelled data is available. Indeed, since the acquisition of gesture data is often expensive in terms of the effort and time required of both the user and the researcher, such a technique which learns online is particularly desirable. Further, the capabilities of GNG to add nodes and to alter its topology over time suggest that it may be effective in learning new gestures as they are observed. For these reasons, GNG is the clustering method employed in this paper.

The plasticity of the GNG network lies in its ability to add and delete nodes during normal operation. The feature vectors of new nodes represent input patterns which differ from those seen in the past and the topology of the network is altered accordingly. Indeed, this feature of GNG is one of the primary motivations for its selection in this research. Fritzke [3] proposed the incremental augmentation of GNG based on the periodic assessment of local error at each node. The node with the largest accumulated local error is the node whose receptive field (or *cell*) is too large to adequately represent the distribution of inputs within the region and which is most in need of a new node to reduce the global error of the network. However, in this simple form, incremental learning may result in the addition of a large number of nodes over time. In such a case, both overfitting at overlapping cluster boundaries and excessive computing time may ensue. Alternatively, a maximum node count may be set which potentially limits network plasticity [1].

Fritzke [7] also proposed a utility-based approach (GNG-U) for the resource-conserving deletion of nodes in order to allow GNG to track non-stationary input distributions. However, in terms of life-long learning, this approach may remove nodes which represent past learning and thus leading to instability. Hamker [1] proposed a method for stategic insertion of nodes using local error thresholds developed from quality measures based on both long-term and short-term local error. The method was effective but focused on supervised learning scenarios. Furao and Hasegawa [8] extend this work to focus on the insertion of nodes in unsupervised tasks. This method attempts to assign unlabeled data to clusters autonomously before applying an adaptive similarity threshold based on cluster size. Input to an existing node is compared to the threshold to determine if it represents a new pattern class and is thus a candidate site for node insertion. The method also performs assessment to determine whether a particular insertion effectively reduced the network error in the long-term. Nodes which do not reduce the error are deemed ineffective and removed. This method, however, presupposes separable input distributions in order to place nodes in distinct clusters.

In each of the approaches mentioned above, however, the possibility of online learning and the need to accommodate a human user/trainer is neglected. The presence of a human user poses significant challenges in terms of input data separability and learning rate. As noted in [2], gesture motion data collected from human participants may be poorly separated and thus, may adversely affect the speed of convergence for an algorithm dealing with unlabeled input. This issue becomes especially important when considering the physical and congitive burden to the user as they perform and apply feedback to potentially large numbers of gesture samples. Key differentiating features of our research include the proposal of a use model which reduces the physical burden on the user, and a method for making gesture classifications for lifelong learning with unlabeled data. These features are detailed in section 3.

3 Method

This section describes the gesture set used for experimentation and its relationship to the ART device. Toward the goals of reducing user effort and size requirements of the input data set, a use model and training paradigm are detailed. Also, a novel method for node insertion which preserves network stability while promoting the rapid learning of new gestures is described. Essential components of this method including data representation, simulation, reward generation and action learning based on GNG were first developed in [2].

3.1 Gesture Types

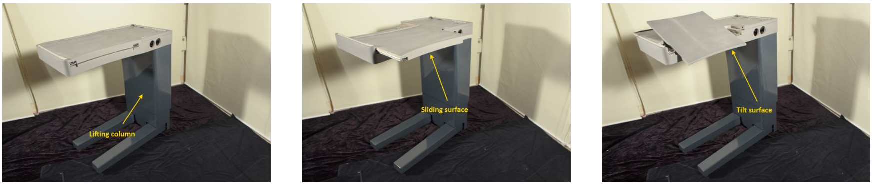
For the experimentation discussed in this paper, six gesture types are considered. These are selected with the user's intention in mind and are broadly indicative of activities in which the user wishes to engage or to have ART support. These include *eat*, *read*, *rest*, *take* (take an item away), *give* (bring an item closer) and *therapy* (use the specially designed therapy surface - see Fig. 1b). Although no particular choreography is required, performance models for these gestures were taken from the American Sign Language Dictionary [9] for repeatability among participants. Envisioned goal configurations for these these gestures are understood to exercise the three degrees of freedom within ART shown in Fig. 2. Their numerical values are mappings to distinct points (x, y, θ) for simulation purposes. The qualitative labels and their mappings are given in Table 1.

Table 1. 3D goal configurations for ART

Gesture Type	Lift	Slide	Tilt	Mapping in (x, y, θ)
eat	low	center	down	$(-3.95, 3.95, 135^o)$
read	high	center	up	$(3.95, 0, 0^o)$
rest	high	center	down	$(0, 3.95, 90^o)$
take	high	away from user	down	$(-3.95, 0, 180^o)$
give	high	toward user	down	$(0, 3.95, 270^o)$
therapy	middle	center	down	$(3.95, 1.98, 22.5^o)$

3.2 Data Collection and Gesture Representation

A representation of gesture motion based on the concept of Dynamic Instants (DIs) [10] is employed. DIs are defined as the extrema of acceleration in the motion of an actor. Using the Microsoft Kinect RGB-D sensor [11] to capture 3D depth data for the motion of an actor's left hand, the five DIs of greatest magnitude during an isolated five second performance interval are concatenated to form a gesture motion descriptor (Fig. 3). The sensor was placed at a height of 75 *cm*. Participants stood at a distance of 1.3 *m* in front of the sensor to perform gesture samples.



(a) (b) (c)

Fig. 2. The three DOFs of ART: (a) the vertical lifting column, (b) the horizontal sliding table top and (c) the tilting work surface

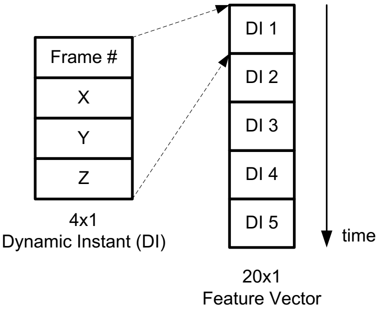


Fig. 3. Feature vector format for a depth-sampled gesture. DIs are concatentated in chronological order by frame number.

3.3 The Growing Neural Gas Algorithm

The Growing Neural Gas (GNG) algorithm [3] is a vector quantization technique in which neurons (nodes) represent codebook vectors that encode a submanifold of input data space. GNG forms connections between nodes and thus preserves a topological representation of input space in a manner functionally similar to the Self-Organizing Feature Map (SOFM). It is further capable of adding new nodes so as to allow for a changing input data distribution. The reader is referred to [2] for details of the algorithm and its implementation in this research.

3.4 Use Model

A use model is proposed which aims at reducing the physical and cognitive burden to the user in terms of the number of training iterations required for the system to fully learn the desired actuation. In this model, the user demonstrates a single sample of a new gesture to a system which has been pretrained to respond to a baseline set of gestures. The user then observes the robotic agent's incremental attempts to assume a desired configuration. As they do so, the user provides a series of consecutive rewards until the system is fully trained for that sample.

Training (or, *path shaping* [12],[13]) consists of simple binary rewards $r \in \{-1, 0, 1\}$ (*cold*, *hot*, *warm*, respectively) assigned to incremental movements of the robot agent in response to the gesture. Movements toward a user-defined goal are assigned rewards of 1. Movements away from the goal are assigned rewards of -1. Gestures which, in the course of training, elicit the full and complete action toward the user's goal are deemed fully *trained* and are given a reward of 0. Upon completion of training for a given gesture, the learning policy for the GNG node (the action associated with that node) is frozen. Thus, any subsequent similar gesture whose feature vector falls into the receptive field for the same node require no further training. For the available data set, this approach is shown to require a human-tolerable number of training iterations.

3.5 Lifelong Learning

As previously stated, the presence of a human trainer represents a key difference between the past methods described in section 2 and that presented in this research. Here, input gesture samples are unlabeled and may not be well separated. However, using the proposed use model, the user-generated reward may be considered a binary *in-cluster/out-of-cluster* indicator. In the case of fully trained nodes, an input pattern which receives negative rewards when executing the action vector associated with that node is interpreted to be of a different class. The cell location indicated by the input feature vector is, then, likely to be a good candidate site for node insertion and the formation of a new cluster.

In the proposed approach, the local accumulated error of the winner in this case (the node nearest the input feature vector) is artificially inflated to the network maximum. At the same time, any nodes in the network whose most recent reward is negative (*cold* nodes) are considered for deletion. The GNG *age* parameter for connections within the network may loosely be thought of as being indicative of a node's nearness to a cluster center. A node with older-aged connections has previously been matched with fewer incoming patterns in those regions where its connections are oldest. When the network has reached a defined maximum node count, the cold node with the highest sum of connection ages is targeted for deletion by the artificial aging of its connections to the maximum age limit. If the network is not at the maximum node count, then a new node may be added without deletion elsewhere in the network. In cases where all nodes in the network are either fully trained or are receiving positive rewards, new nodes may be added above the predefined maximum. This effectively relaxes the predefined maximum to afford plasticity when needed. This scheme for node insertion/deletion is summarized in Algorithm 1. In this manner, new node clusters are allowed to form without catastrophically eliminating existing knowledge gained through training.

3.6 GNG Network Distance Metrics

Of particular interest in determining cluster membership for the purpose of gesture classification are the intra-node distances and connectivity which emerge

Algorithm 1. Node insertion/deletion algorithm

```
Apply a gesture input sample.
Determine the *winner* reference node.
Perform the winner's associated action vector.
Observe the user-generated *reward*.
**if** *winner* is trained and *reward* is *cold* or *warm* **then**
   Inflate local error: $winner.E = max(refNode[i].E) + 1$.
   **if** $numNodes < maxNodeCnt$ **then**
      A node will be inserted near *winner*.
   **else**
      Locate a *cold* node having greatest the sum of connection ages.
      **if** A *cold* node exists **then**
         Target if for delection by inflating connection ages: $C[i].age = ageMax + 1$.
      **else**
         $numNodes$ is allowed to increase beyond $maxNodeCnt$.
      **end if**
      A node will be inserted near *winner*.
   **end if**
**end if**
GNG will perform node insertion and deletion in the next time step.
```

from the GNG *cloud* as it matures during operation. These quantities allow for *neighborhood learning* [14]. By examining the past rewards of neighboring nodes, the system may select action vectors from among those neighbors whose actions have received positive rewards in the past. This has the effect of allowing a cluster of nodes to behave similarly and to learn more rapidly, thereby providing indications of cluster membership to otherwise unlabeled data. For this research, two distance metrics are considered. These metrics include:

1. Euclidean distance - node neighbors within a mean distance of all connected nodes are considered, and
2. Estrada's network clumpiness metric [15] - node neighbors of maximum clumpness are considered. Clumpiness Ξ for a given node is computed as in (1).

$$\Xi_{ij} = \begin{cases} \dfrac{k_i k_j}{(d_{ij})^2} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \tag{1}$$

where k_i is the degree of node x_i and d_{ij} is the network distance between nodes x_i and x_j as computed using Floyd's algorithm [16] with connection age serving as length. It is shown in section 5, that although computationally intensive, clumpiness is highly effective as a means of selecting neighborhood nodes with action vectors likely to yield positive rewards.

4 Experimentation

Five participants each performed fifty repetitions of each of the six candidate gestures. This yielded 250 samples of each gesture for a total of 1500 samples. Participants were encouraged to perform gestures as consistently as possible. Dynamic instants (DIs) were computed for each sample. Feature vectors were constructed from the DIs and presented to the system as described in section 3.4.

The 1500 gesture samples for the six candidate gestures were divided into two data sets. The *training data* set consisted of the gestures *eat*, *read* and *rest*. From these, a set of 450 samples (150 samples of each type) were selected and randomized. The second *test data* set consisted of the 750 samples of the gestures *take*, *give* and *therapy* sequenced randomly. The system was initially pre-trained using the training data set. The network was constrained to include 100 nodes. This step yields the essential GNG data structures A (node list) and C (connection list) which define a mature GNG network for the *eat*, *read* and *rest* gestures contained in the training set.

With the system pretrained, a single *epoch* (one presentation of all gesture samples in the data set) was applied one sample at a time according to the use model described in section 3.4. Upon each presentation of a sample to the system, a simulation sequence was performed which included execution of GNG, simulation of robotic action, and assignment of reward. This sequence was repeated for that sample until one of three terminating conditions was reached:

1. The reference node closest to the input gesture sample became fully trained.
2. The input gesture sample received a negative reward in the receptive field of a fully trained node. In this case, the sample was immediately ignored and a new node was inserted near the trained node according to Algorithm 1.
3. The number of training iterations exceeded 1000 (the *confusion threshold*). This indicates that the formed neighborhood is issuing conflicting action advice and the input sample is near a boundary between clusters. In this case also, the sample was ignored. However, the number of attempted learning iterations was considered in the calculation of outcome metrics.

In this way, a 3-epoch sequence was conducted as described below. Following each epoch, performance metrics were recorded. These metrics included the total number of nodes in the GNG network, the number of fully trained nodes, the percentage of samples ignored, and the average number of training iterations per sample. The sequence was conducted for the two distance metrics methods described above.

1. **Demonstration of Plasticity.** With the system initially trained using the training data set, a single epoch of the test data set was applied. This phase was intended to demonstrate the plasticity of the GNG network to learn the *take*, *give* and *therapy* gestures.
2. **Demonstration of Stability of Past Learning.** A single epoch of the training data was reapplied. This phase was intended to demonstrate the

stability of the system learning implementation. If the implementation is indeed stable, the outcome would be expected to reflect an already-trained network. That is, the performance metrics would show iteration counts which remain tolerably few for a human trainer.

3. **Demonstration of Stability of New Learning.** A final epoch of the test data was executed. This phase reinspects the network for the stability of the newer *take*, *give* and *therapy* gestures introduced by the test data set in the first epoch.

Results for this experimental procedure are given in section 5.

5 Results and Discussion

Typical results for execution of the three epochs are given in Table 2.

Table 2. Results for three epochs

Epoch	Distance Metric	# Nodes	# Trained Nodes	Samples Ignored (%)	Average Iterations
1	Mean	100	85	9.1	7.32
	Clumpiness	100	93	7.3	8.89
2	Mean	99	91	4.9	5.62
	Clumpiness	100	98	5.6	2.89
3	Mean	100	93	3.1	0.79
	Clumpiness	101	100	1.2	0.97

For epoch 1, the GNG network was previously trained to the *eat*, *read* and *rest* gestures. Application of the test data in the first epoch shows the plasticity of the network in learning new gesture types under the proposed use model. Two metrics in particular are seen as key to evaluation of the use model: (1) the percentage of samples ignored and (2) the average number of training iterations. As previously stated, samples may be ignored by taking too long to train (exceeding a confusion threshold of 1000 iterations). They may also be ignored if they fall into the receptive field of a previously trained node and receive negative reward. The rationale to ignore such *problem* samples is based on the assertion that non-action on the part of the robot is preferred to persisting with training and ultimately performing an undesirable action. Further, alteration of a previously trained action would negatively affect the stability of the system. Thus, the priority for alteration of the network is set in favor of stability over the attempt to adapt to a rapidly changing input distribution. It can be seen from Table 2 that the percentage of samples ignored is small (less than 10.0%).

The clumpiness metric ignores the fewest samples. This is coupled to the improved separability of the data set as participants were guided to perform gestures in a uniform manner. With well-defined clusters in the GNG network, the proximity of any given gesture input to the cluster center for its class is likely to have improved, while the distance between cluster centers will have increased. Thus, the clumpiness computation would be more apt to form its neighborhood from members its own class.

The average numbers of iterations (less than nine iterations per sample) are manageable in general, if still somewhat burdensome to the user. It is noted, however, that those gesture samples which are ignored for having exceeded the confusion threshold will negatively impact this metric. The attempted iterations are not deducted from the total iteration count over the epoch and thus contribute to a higher average. After several nodes of each gesture class are fully trained within the network, the overwhelming majority of subsequent samples requires no training at all. Further, the average number of training iterations is seen to decrease further in subsequent epochs. These results demonstrate that the fully trained network which existed before the test data was first applied is capable of learning new gestures in a human-tolerable number of time steps.

For epoch 2, training data was reapplied to the network after it had been newly trained with the test data set. These results reflect the stability of the GNG implementation. It can be seen that both the average number of iterations and the percentage of samples ignored are now smaller for both distance metric schemes. Again, the clumpiness metric yields best results.

Epoch 3 underscores the stability of the system which remains stable through the reapplication of test data. Both the average numbers of iterations and the number of samples ignored have decreased from the first application of this data set under both distance metric schemes. The clumpiness metric is typically (though not always) seen to ignore the fewest samples. Although not reported quantitatively here, subsequent epochs for either the training data or the test data frequently resulted in convergence to zero iterations per sample: the entire network had become fully trained for the available data sets. This result may be problematic in cases where gesture data is poorly separable; the algorithm may have *overfit* the data. A more discriminating method for node insertion may be desirable to temper the generalizing capability of the network in such cases.

6 Conclusions and Future Work

In this paper, we have presented a method for training a gesture-based interface to a robotic agent (ART) with a human user/trainer. We have introduced a use model for the agent which attempts to minimize the physical and cognitive loads on the user in terms of training iterations. It has been shown that the GNG algorithm offers a construct for learning new gesture classes while retaining past information. Strategic addition and deletion of GNG nodes based on their history of user-generated reward within a node neighborhood was shown to facilitate both plasticity and stability of learning.

Future work in this area will include development of a means by which training sequences for a given gesture may be abandoned early if they would fail to converge. Also, alternative network distance metrics (and models for assigning connection *lengths*) will be explored in pursuit of faster neighborhood learning.

Acknowledgments. This research was supported by the U.S. National Science Foundation under award IIS-SHB-116075.

References

1. Hamker, F.H.: Life-long learning Cell Structures-continuously learning without catastrophic interference. Neural Networks 14(4), 551–573 (2001)
2. Yanik, P.M., Merino, J., Threatt, A.L., Manganelli, J., Brooks, J.O., Green, K.E., Walker, I.D.: A Gesture Learning Interface for Simulated Robot Path Shaping with a Human Teacher. IEEE Transactions on Human-Machine Systems 44(1), 41–54 (2014)
3. Fritzke, B.: A Growing Neural Gas Network Learns Topologies. Advances in Neural Information Processing Systems 7(7), 625–632 (1995)
4. Grossberg, S.: Nonlinear neural networks: Principles, mechanisms, and architectures. Neural Networks 1(1), 17–61 (1988)
5. Kohonen, T.: The self-organizing map. Proc. of the IEEE 78(9), 1464–1480 (1990)
6. Holmström, J.: Growing Neural Gas: Experiments with GNG, GNG with Utility and Supervised GNG. Master's thesis, Uppsala University – Department of Information Technology (2002)
7. Fritzke, B.: A self-organizing network that can follow non-stationary distributions. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 613–618. Springer, Heidelberg (1997)
8. Furao, S., Hasegawa, O.: An incremental network for on-line unsupervised classification and topology learning. Neural Networks 19(1), 90–106 (2006)
9. ASL Pro Website, `http://www.aslpro.com/cgi-bin/aslpro/aslpro.cgi`
10. Rao, C., Yilmaz, A., Shah, M.: View-Invariant Representation and Recognition of Actions. International Journal of Computer Vision 50(2), 203–226 (2002)
11. Microsoft Xbox 360 + Kinect Website, `http://www.xbox.com/en-US/kinect`
12. Kaplan, F., Oudeyer, P.Y., Kubinyi, E., Miklósi, A.: Robotic clicker training. Robotics and Autonomous Systems 38(3), 197–206 (2002)
13. Yanik, P.M.: Gesture-Based Robot Path Shaping. PhD thesis, Clemson University (2013)
14. Touzet, C.F.: Neural reinforcement learning for behaviour synthesis. Robotics and Autonomous Systems 22(3), 251–281 (1997)
15. Estrada, E.: The Structure of Complex Networks: Theory and Applications. Oxford (2012)
16. Tucker, A.: Applied Combinatorics, 6th edn. Wiley (2007)

Gesture, Gaze and Activity Recognition

The Issues of 3D Hand Gesture and Posture Recognition Using the Kinect

Mohamed-Ikbel Boulabiar[1], Gilles Coppin[1], and Franck Poirier[2]

[1] Lab-STICC, Telecom Bretagne, France
{mohamed.boulabiargilles.coppin}@telecom-bretagne.eu,
[2] Lab-STICC, University of Bretagne-Sud, France
franck.poirier@univ-ubs.fr

Abstract. Besides the emergence of many input devices and sensors, they are still unable to provide good and simple recognition of human postures and gestures. The recognition using simple algorithms implemented on top of these devices (like the Kinect) enlarges use cases for these gestures and postures to newer domains and systems. Our methods cuts the needed computation and allow the integration of other algorithms to run in parallel. We present a system able to track the hand in 3D, log its position and surface information during the time, and recognize hand postures and gestures. We present our solution based on simple geometric algorithms, other tried algorithms, and we discuss some concepts raised from our tests.

Keywords: Gesture, Posture, 3D, Kinect, Interaction, Hand.

1 Introduction

During the last years, we have seen a big interest in 3D gesture interaction in the research and the industrial field, many input devices and sensors were and are still being released to translate human movements into computer information. Sadly, many sensors either have complex systems for gesture recognition [11,7], takes a lot of computation power or still lack good recognition algorithms. Some studies show that the mouse is still unbeaten in its current use [2] and this motivates us to figure out new scenarios for gestures and postures systems [4].

3D gestures have also the specificity of not having a clear hardware timing of when a gesture starts and ends. In contrast with multi-touch devices that define the beginning and the end of the gesture by fingers touching the surface and leaving it, in 3D we do not touch physical objects and this is what makes the problem harder.

The increasing number of sensors and devices, and the emergence of new 3D visualization techniques like the 3D stereoscopy [14], pushes us to test a new approach in creating hand gestures and postures recognizers. We target a user commanding a system using a table and the space above. We take the object itself into consideration taking a part in the recognition method in a way different from just using physics simulation libraries [6,17]. We detail in this paper a simple and

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 205–214, 2014.

real-time solution for recognizing gestures and postures, which can be embedded into other systems. As we take the manipulated object into consideration, the recognition becomes instantaneous and newer concepts start to emerge. We have chosen to mold geometric recognition algorithms towards our needs.

Our contributions are: 1. The fast system for tracking the hands from the 3D raw points data. 2. The use of the same geometric algorithms to detect both gestures. 3. The use of the same algorithms to detect postures by transforming the hand contour into an algorithm input. 4. The experimentation of other methods and ideas in the same context.

In this paper, we start by describing our base system for recognizing, tracking and logging the hand in 3D; then we describe how we have used and extended simple geometric algorithms for 3D gesture analysis and for hand posture recognition. We describe other tested algorithms and finally we discuss the recognition issues and provide some new ideas coming out from our applied study.

2 Kinect-Based System for Tracking and Recognition

2.1 Installation

In our system as shown in fig. 1 we have used the Kinect as a depth sensing camera mounted in the ceiling above the user. We have decided to process the raw data directly to be able to optimize the pipeline and get the maximum speed. The standard Microsoft Xbox Kinect sensor pipes us a raw input of 640x480 3D points cloud at 30Hz frequency. We limit the captured zone to the size of 100x80cm and a 60cm of depth above the table because of the human reachability concerns [10] and table size. The Kinect is connected to an Intel Xeon computer with ubuntu 64bit installed. We have used the open source libfreenect library for kinect access in addition to OpenCV.

2.2 Tracking of the Blobs

To accelerate the hands detection and tracking, we have decided to do all the tracking on 2D surfaces and using well designed one pass per frame algorithms. So while keeping the 3D data of the Kinect on separate data structures, we have flattened the recorded 3d box of points into a single layered image (in comparison to a three layered RGB image) which can proceeded by OpenCV. We have used cvBlob library to label the blobs present on the scene then we apply the same algorithms described in [3] to track the resulted blobs between frames. We still have access to all the 3D information after the flattening operation since the single layer where blobs will be tracked in 2D has been duplicated in memory.

2.3 Extraction of the Hand

The blob we obtained form the hand and the arm, but as we only want information about the hand, we extract only that part from the bigger blob of the

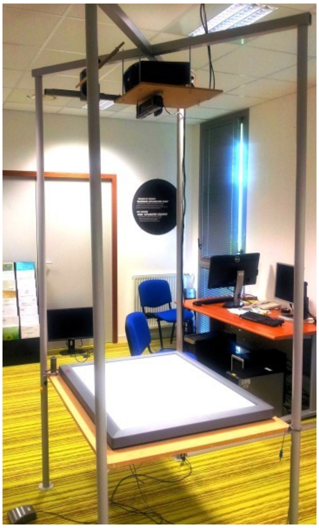


Fig. 1. Kinect Install

full arm and we get the position of the hand center. To extract it, we simplify the blob into a shape just fewer points, we try to mesure and compare distances between points and we select the longer segment. As the segment contains the hand farthest point and the point near the forearm, we select the one which is near the interaction zone center as shown in fig. 2. To select the hand center, we have selected a constant interval from the extreme point towards the other direction. The interval size varies by the hand vertical position.

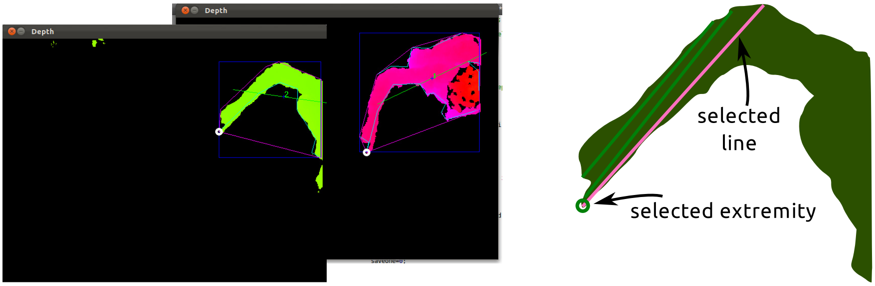


Fig. 2. Hand Extraction

2.4 Logging of the Hands Information

We are able to process data in near real-time while recording the hand surface shape and the 3D coordinates of hand movements through time for further use in posture recognition as shown (fig. 3). Data structures for tracking and identifying the hand are separated from 3D raw data, but we use them to select the zone to be recorded. We select a fixed rectangle around hand center and we verify whether it is inside the recording zone. We have used textual files to facilitate debugging and allow direct visualization using gnuplot. One problem faced with 3D sensors is that they can only provide the surface layer of an object, and not the 3D blob in itself. This means that when we speak about recording the hand, we record only the points of its surface which are between the sensor and the real hand. This limits for example recognizing what happens below the hand surface.

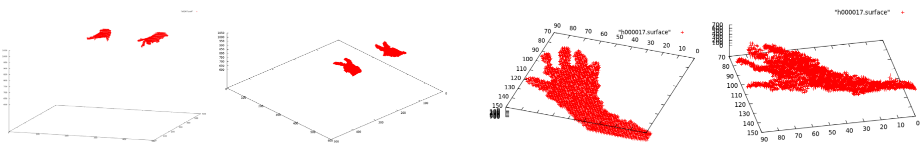


Fig. 3. The recorded scene including hands, and the extracted and recorded hand surface

2.5 Application and Research Context

In the previous section we have described the 3D input handling part, but the general context where our system as described in fig. 4 is used is the maritime surveillance. The input handling and the maritime systems are connected through a software bus where we can pipe recognition results in one direction and the commands for the mode of recognition tuning in the other direction.

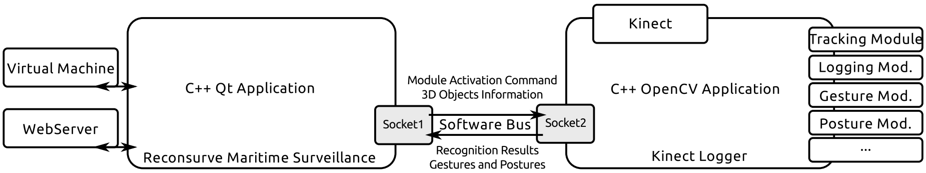


Fig. 4. Maritime suveillance system and Kinect Logger system architecture

3 Recognition Methods

3.1 Use of Geometric Algorithms

When starting the development of our project, the performance was one of our biggest concerns so we focused on using algorithms that take the shortest computing time and we tuned them to our needs. We have studied the Rubine[13] and the "1 dollar" families [18,9] of stroke recognizers. We have chosen to use the 1$ recognizer (or its variation "Protractor") for its simplicity and speed. We define geometric algorithms as those which use simple geometry operations and measurements in order to compute a distance value, in contrast to soft computing algorithms.

3.2 Our Use of 1$ Algorithm in Gesture Recognition

In our system, we track the hand center and the pointing finger. We have extended the 1$ algorithm to work on 3D strokes. Works like [5] or 3$ [8] have only used either a different algorithm or a still 2D recognition of 2D strokes performed in the 3D space, the work of Haubner et al. [5] in particular worked mostly on searching the flat space of a gesture. In our 3D adaptation of 1$, we have tested 3D strokes that can not be reduced to a simple plan.

3.3 Our Use of 1$ Algorithm in Posture Recognition

By posture, we define the current configuration of the hand similarly to Baudel et al. [1]. The 1$ algorithm is supposed to be used with mouse, touch screen or pen strokes. We have got the idea to keep using it but for hand posture recognition based on previous work we have made [3]. We have managed to make the hand contour as the input of the algorithm (fig. 5), then we have recorded a set of template, and the slightly modified algorithm have worked and we are now able to detect our set of hand postures, which are just a subset of the American Sign Language [1].

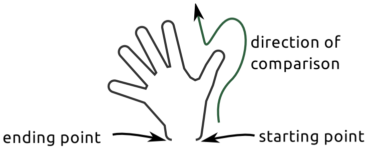


Fig. 5. Using 1$ in posture recognition

3.4 Pointing 3D Objects on the Table

In our system, as we are able to track the hand center and its extreme, which can be the pointing finger, we have developed a mode where we track these two points in 3D and detect the direction pointed by the finger. As a quick and fast application, we computed the fixed position of the table (Z=constant) to simplify equations and detect where the user finger is pointing on the table shown in fig. 6 below.

[1] http://en.wikipedia.org/wiki/American_Sign_Language

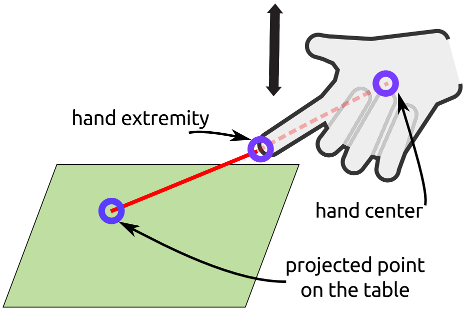


Fig. 6. Pointing on table

4 Other Tested Algorithms

4.1 Extension of Angle Quantization Method in 3D

The angle quantization geometric algorithm [12] works by coding the stroke into a vector of values. These values calculate the parts being in a specified angular zone. The algorithm allows fast and high detection rate of strokes but fails in differencing between repeated stroke patterns like **V**, **W** and **VVV**. These patterns have parts in the same angular zone, so even if they are repeated, the AQ algorithm can't differenciate between them. We have tried extending the AQ to the 3D space [2]

4.2 Application of the ICP Algorithm for Hand Tracking

We have tried using usual point cloud algorithms like Iterative Closest Point (ICP) for aligning the recorded hand on one of the templates and use the angles and positions given by the algorithm to compute the transformation and thus detect the gesture. The prototype code allowed us to get acceptable results but appeared to be very slow. The ICP algorithm was not intended for real-time use and discouraged us from continuing through that research area. We should note that during the tests, we have tried giving the algorithm the fixed part which is the hand back without the fingers. The use of this part makes better rotation recognition. The use of simpler geometric algorithms seems more appropriate.

5 User Experimentation

5.1 Definition of Gestures

In our prototype, and before thinking about how gestures can be natural, we tried recognizing the usual 3D strokes and we have defined 4 arbitrary and simple ones as shown in fig. 7 just to test our recognition algorithm. We have chosen 4 gestures that can not be reduced into a 2D plan by studying their principal components.

[2] `https://github.com/dylandrover/3D-AQ`

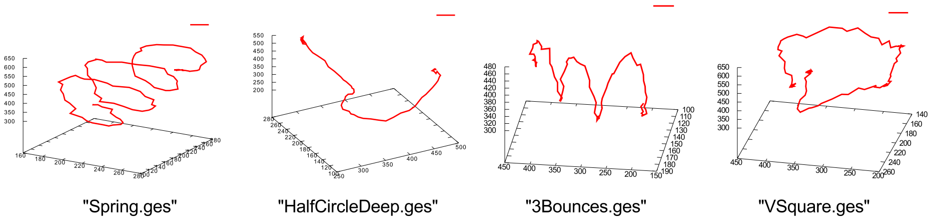


Fig. 7. A set of 4 pure 3D gestures arbitrary selected

5.2 The Naturality of the Performed Gestures

During preliminary tests and recording of 3D command gestures, we have spotted a problem of memorability, which we think it comes from the background of human activities. Humans are very well used to write and draw on a 2D paper but not in space. Only skilled sculptors can interact with a three dimensional element. What we can do is touching an object, moving it, rotating it, and sometimes compressing it, but not commanding an object or a system with indirect gestures. The natural gesture in reference to a hand and an object should be classified into four basic families: (Touch, Move, Rotate, Scale) in reference to how we manipulate objects in nature [15]. We think that a hand interaction with objects need first to be categorized into one of these four classes, then we look further into sub-properties to achieve a fine-grained classification.

5.3 Benchmarks and Recognition Rates

We have tested the time it takes to compute the gesture after we finish recognition. It takes less than 60ms on our machine, and with our set of gestures. For the hand posture recognition, we have made prior tests in the past using the same posture recognition algorithm and an RGB camera, we were able to reach realtime recognition rates [3]. When using the Kinect, we have flattened the hand capture then extracted its contour and we are able to reach a similar but not yet evaluated recognition rates.

6 Recognition Issues and Ideas

6.1 Gesture Parsing (Start and End)

We have been faced though by the problem of real-time gesture parsing. Knowing when a gesture starts and when it finishes pushed us first into using a foot pad to tell the system when it must start considering the recorded 3D points as part of the gesture, and another foot click to stop recording. The system delivers after start and stop the detected gesture. A priror work [16] used a posture to start an interaction. The work could be improved by inheriting ideas from

[3] `http://youtu.be/AbNKPBCw4EU`

speech recognition system as for detecting the commands between two silences. In our case, the detection of a possible gesture will be performed between two stationary positions while posture detection will be performed in them as shown in fig. 8.

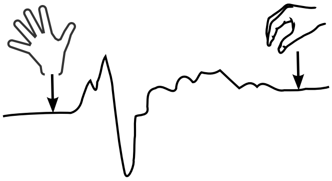


Fig. 8. The gesture parsing by seeking big difference in hand movements

6.2 Recognition Simplification Using Object Position

While searching for methods to easily detect rotation, we have tried first detecting it in the hand itself based only on the point cloud transformation. This appeared a computation hog using the ICP algorithm. Then we have questioned detecting such gesture without the presence of a target object. Having an object interacting with the hand, and willing a rotation, means that the hand-object position will change and the line linking their respective centers will rotate. We can know about the center of an object, and we track the hand, so we know where it is. We are able to detect rotation instantly using this method.

6.3 Spheres of Interaction

The interaction we want to promote is the one with objects because that is where natural interaction goes instead of commands or posture interaction. We have proposed in the previous paragraph that we can simplify algorithms using the hand and object positions. Here, we extend this approach to define interaction zones or spheres of interaction around the object as shown in fig. 9. We dedicate the first sphere around the object, which is bigger than it by two times the hand thickness, to direct manipulation of the object by moving, resizing, rotating and selecting it, and the second layer to accept indirect commands to be applied on it. When the hand is not in these zones, we ignore its posture and movement which is far from the object.

7 Conclusion and Future Work

In this paper, we have shown that we can provide a system capable of tracking hand movements in the space above the table, logging information, and recognizing gestures and postures in real-time. The most relevant feature of our work is that it is able to reach good performances using only very simple algorithms. Use them in a new way. Our approach can help other researchers by giving them

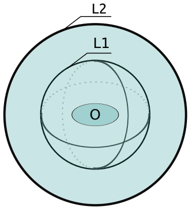


Fig. 9. Spheres or layers of interaction around the object, their size depend on the object and hand ones

the tests and examples that worked and those which ended up with some constraints.We think that the simplification of recognition using information about the object along with the hand, and the position of these two is an idea to consider in further studies. Future studies will target making the system more robust and improve its recognition capabilities. We plan also to target more user testing and perform new benchmarks.

Acknowledgment. We would like to thank Dylan Drover for the internship he has made in our laboratory through the LEIF Exchange program. He worked on exploring the extension of 2D Angle Quantization algorithm to 3D and on the mentioned ICP tests.

References

1. Baudel, T., Beaudouin-Lafon, M.: Charade: Remote control of objects using free-hand gestures. Commun. ACM 36(7), 28–35 (1993)
2. Bérard, F., Ip, J., Benovoy, M., El-Shimy, D., Blum, J.R., Cooperstock, J.R.: Did "minority report" get it wrong? superiority of the mouse over 3d input devices in a 3d placement task. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5727, pp. 400–414. Springer, Heidelberg (2009)
3. 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)
4. 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, pp. 3–12. ACM, New York (2010)
5. Haubner, N., Schwanecke, U., Dörner, R., Lehmann, S., Luderschmidt, J.: Recognition of dynamic hand gestures with time-of-flight cameras. In: Proceedings of ITG/GI Workshop on Self-Integrating Systems for Better Living Environments, vol. 2010, pp. 33–39 (2010)
6. Hilliges, O., Izadi, S., Wilson, A., Hodges, S., Garcia-Mendoza, A., Butz, A.: Interactions in the air: Adding further depth to interactive tabletops. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 139–148. ACM (2009)

7. Ionescu, B., Coquin, D., Lambert, P., Buzuloiu, V.: Dynamic hand gesture recognition using the skeleton of the hand. EURASIP Journal on Applied Signal Processing (2005)
8. Kratz, S., Rohs, M.: A $3 gesture recognizer: Simple gesture recognition for devices equipped with 3D acceleration sensors. In: International Conference on Intelligent User Interfaces, pp. 341–344 (2010)
9. Li, Y.: Protractor: A fast and accurate gesture recognizer. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 2169–2172. ACM, New York (2010)
10. Marquardt, N., Jota, R., Greenberg, S., Jorge, J.A.: The continuous interaction space: Interaction techniques unifying touch and gesture on and above a digital surface. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part III. LNCS, vol. 6948, pp. 461–476. Springer, Heidelberg (2011)
11. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Markerless and efficient 26-dof hand pose recovery. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 744–757. Springer, Heidelberg (2011)
12. Olsen, L., Samavati, F.F., Sousa, M.C.: Fast Stroke Matching by Angle Quantization. In: Proceedings of the ImmersCom (2007)
13. Rubine, D.: Specifying gestures by example. ACM SIGGRAPH Computer Graphics 25(4), 329–337 (1991)
14. Valkov, D.: Interscopic multi-touch environments. In: ACM International Conference on Interactive Tabletops and Surfaces, ITS 2010, pp. 339–342. ACM, New York (2010)
15. Victor, B.: A Brief Rant on the Future of Interaction Design (2011), `http://worrydream.com/ABriefRantOnTheFutureOfInteractionDesign/`
16. Walter, R., Bailly, G., Müller, J.: Strikeapose: Revealing mid-air gestures on public displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 841–850. ACM, New York (2013)
17. Wilson, A., Izadi, S., Hilliges, O., Garcia-Mendoza, A., Kirk, D.: Bringing physics to the surface. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, pp. 67–76. ACM (2008)
18. Wobbrock, J., Wilson, A., Li, Y.: Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 159–168. ACM (2007)

Frontal-Standing Pose Based Person Identification Using Kinect

Kingshuk Chakravarty and Tanushyam Chattopadhyay

Innovation Lab, Tata Consultancy Services Ltd., Kolkata, India
{kingshuk.chakravarty,t.chattopadhyay}@tcs.com

Abstract. In this paper we propose a person identification methodology from frontal standing posture using only skeleton information obtained from Kinect. In the first stage, features related to the physical characteristic of a person are calculated for every frame and then noisy frames are removed based on these features using unsupervised learning based approach. We have also proposed 6 new angle and area related features along with the physical build of a person for the supervised learning based identification. Experimental results indicate that the proposed algorithm is able to achieve 96% recognition accuracy and outperforms all the stat-of-the-art methods suggested by Sinha et al. and Preis et al.

1 Introduction

Biometric Person identification in an image or video is of crucial importance and it is critical to determine the presence of a particular person for applications where automatic person recognition is a key enabler such as security and surveillance, elderly people care etc. People are mainly identified based on different physical and behavioral features e.g. iris, fingerprint, speech, face etc. But biometric identification based on these modalities are intrusive as they require direct human interaction. In addition, extracting face, iris or fingerprint characteristic from a large distance or in poor lighting condition are indeed a challenging job. This paper aims at developing a novel person identification algorithm based on only physical build characteristic of a person. As the overall physical structure of a person can be extracted at a large distance and it is very difficult to imitate or hide, the method has clear advantages over the exiting ones. One approach to determine physical characteristics of person is to capture skeleton joint co-ordinates over time. But to accomplish this, we need to have multiple positional cameras to obtain skeleton information. Fortunately, Microsoft provides us a 3D (RGB-D) sensor platform called "Kinect" which can directly provide the 20 skeleton joint co-ordinates. As we are only using skeleton information instead of video or RGB-D image, our proposed method can properly ensure user's privacy and security issue.

After obtaining the skeleton information, the physical build (features) of a person like body dimensions, height, length of two legs, arms etc. can be easily computed from the data. As human being is capable of identifying a person from

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 215–223, 2014.

his/her physical or structural build, any standard statistical learning method (supervised or unsupervised) can be used to map these unique features to a particular object class repressing a person. It needs to be mentioned that person identification using skeleton information already exists in the literature. Preis et al. [1] used physical build of a person like height, length of torso etc. and dynamic gait information like step length and velocity for person identification from constrained side walking pattern. Adrian et al. [2] proposed an unsupervised learning (K-Means) based identification algorithm based on dynamic angular information related to the gait pattern using Kinect and obtained 43.6% accuracy for 4 subjects. Manual gait cycle extraction used by Adrian et al. is not possible in any realtime system. While Naresh et al. [3] tried to model arbitrary walking pattern using only physical build characteristics, Sinha et al. [4] proposed a robust pose and sub-pose based modeling approach for the same. But none of them tried to identify a person from their only static posture using skeleton information obtained from Kinect.

For some applications like TV viewership monitoring or monitoring blackboard activity in school or college, it is very much important to recognize a person from his/her static posture. The static posture may be interpreted as standing, sitting, lying or anything else. To address the above usecases, this paper aims at proposing a novel framework for supervised learning based person identification using only frontal standing pose. We have done the frame level performance analysis as well as comparison our proposed method with respect to existing solution. The key contributions of the paper are given below

- Frontal standing pose based person identification using skeleton data.
- New area and angle related features are proposed for person identification.
- Noisy skeleton data removal using physical characteristic of the person.
- Multiclass Support Vector Machine (SVM) with RBF kernel [2] is employed for supervised learning based person identification.

Rest of the paper is organized as follows. The proposed methodology is described in the Section 2. The detailed results are provided in Section 3 followed by conclusion in Section 4.

2 Proposed Methodology

In this paper, we have presented a frontal standing posture based person identification using only skeleton data obtained from Microsoft Kinect sensor [5]. Kinect provides human skeleton data for 20 skeleton joints at 25 frames per second in real time. The framework shown in the flowchart (Fig. 1) has mainly five modules as given below.

- Acquisition of skeleton data
- Feature extraction
- Noisy frames removal
- Decision Making using Supervised Learning

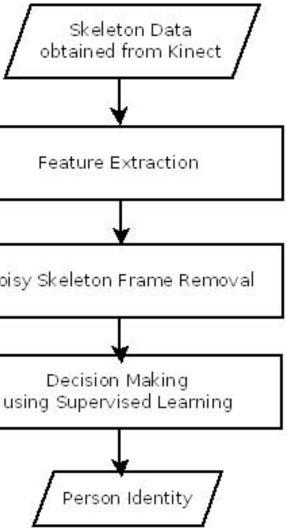
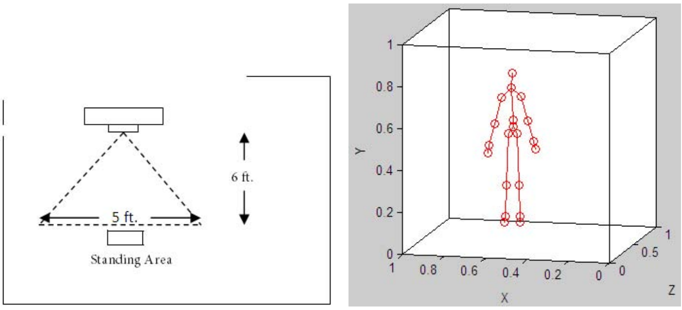

Skeleton Data obtained from Kinect
Feature Extraction
Noisy Skeleton Frame Removal
Decision Making using Supervised Learning
Person Identity

Fig. 1. Flowchart of Our Proposed Algorithm

2.1 Acquisition of Skeleton Data

For data-capture we have marked a fixed position in front of the Kinect where an individual is requested to stand for training and testing. We have used the 20 joints of skeleton data for a person captured at 30 frames per second in frontal-standing posture (figure 2). Each joint consists of 3D world co-ordinates i.e. {x,y,z} tuple in meters considering the Kinect camera as origin of the world coordinate system. We have used Microsoft SDK version 1.5 for the data-capture.



(a) Recording Setup

(b) 3D world co-ordinate points for 20 joints of a subject

Fig. 2. Kinect experimental setup

2.2 Feature Extraction

Feature extraction is one of the main steps for any machine learning based approach. In this case, we have tried to model physical build or structure of a person using a feature vector $\boldsymbol{f}$ which includes
Area Feature ($\boldsymbol{f}_{area}$) - Area occupied by the polygon formed by the joints 1. shoulder left, shoulder right and shoulder center, and 2. hip left, hip center and hip right are unique features for any individual because they do not vary with pose or time. We have considered both of these as one of our candidate features i.e. $\boldsymbol{f}_{area} \in \mathbf{R^2}$. If co-ordinates of i^{th} ($i \in 20$) joint is (x_i, y_i), then the area A enclosed by the N joints can be computed using eqn. 1

$$A = \frac{1}{2}\sum_{i=0}^{N-1}(x_i * y_{i+1} - x_{i+1} * y_i) \tag{1}$$

Angle Feature ($\boldsymbol{f}_{angle}$) - We have calculated four angles mentioned below
1. angle between shoulder left, shoulder center and spine.
2. angle between shoulder right, shoulder center and spine.
3. angle of the shoulder center and spine with respect to the vertical axis.
4. angle between hip left, hip center and hip right.
As these four angles are unique for any individual and also invariant to pose or posture, we have used the same as one of the candidate features ($\boldsymbol{f}_{angle} \in \mathbf{R^4}$).
Features Related to the Physical Build - We consider height, length of upper and lower legs, length of arms etc to describe physical build of a person. For this, we have used all the static features ($\boldsymbol{f}_{static} \in \mathbf{R^{12}}$) mentioned in [1].

We have used feature vector $\boldsymbol{f} = \{\boldsymbol{f}_{area}, \boldsymbol{f}_{angle}, \boldsymbol{f}_{static}\} \in \mathbf{R^{18}}$ for training and testing using SVM.

2.3 Noisy Frames Removal

The skeleton data obtained from Kinect is itself very much noisy. So noise cleaning is required to achieve good recognition accuracy. The noisy frames are identified and removed based on the static feature vector $\boldsymbol{f}_{static}$. For noise cleaning, we assume that the mutual Euclidean distance between two joints should not vary with time. So if it varies significantly from one frame to another, we mark those frames as noisy frames and remove all the features ($\boldsymbol{f}$) corresponding to those frame for further processing. In our implementation, this is done using unsupervised clustering algorithm [4]. The cluster or group with sparsely distributed points (representing the static feature vectors) is identified as a noisy one, and the frames associated with the sparsely distributed static feature vectors are referred to as the noisy frames. The cluster centers are initialized in the following manner.

- We compute A histogram of B bins on $\boldsymbol{f}_{static}$ where each bin is defined by (2) where k represents the bin-index, for i^{th} static features and j^{th} frame level data points (D).

$$Bin_k^{ij}, 1 \leq k \leq B, 1 \leq i \leq 12, 1 \leq j \leq D \tag{2}$$

- The bin k^i_{max} containing the maximum number of data points for i^{th} feature is used to calculate the i^{th} dimension of the first center.
- An mean of all the points belonging to the k^i_{max} bin represents the center (C^i_1)of the first cluster for the feature i and it can be defined as (3), where P^i is the number of feature points $\in k^i_{max}$. The first center is defined as $\boldsymbol{C_1} \in R^{12}$.

$$C^i_1 = \frac{\sum\limits_{j \in k^i_{max}} \boldsymbol{f}^i_{static_j}}{P^i}, 1 \leq i \leq 12 \tag{3}$$

- The second cluster center ($\boldsymbol{C_2}$) is the data point ($\boldsymbol{f}_{static_j}$) representing the static feature vector that is at a furthest distance with respect to the first center ($\boldsymbol{C_1}$). We have selected $\boldsymbol{C_2}$ based on the initialization of the K-Means++ [6] algorithm.

2.4 Decision Making Using Supervised Learning

We have used multicalss Support Vector Machine as supervised learning algorithm for decision making process.

Given N-class training data in the form of D = ($\boldsymbol{f_1}, y_1$), ($\boldsymbol{f_2}, y_2$), ($\boldsymbol{f_3}, y_3$), ..., ($\boldsymbol{f_n}, y_n$) where $\boldsymbol{f_i} \in \mathbf{R^n}$ is feature vector representing a class, a supervised learning algorithm [7] [8] requires a function g which maps the input/feature space (X) into decision or output space (Y) g:X → Y. Here g is the element of hypothesis space G. Some times g is also expressed as scoring function f(x,y): $X \times Y \rightarrow$ R, such that g is defined as g(x) = $\arg\max_y f(x, y)$. For probabilistic learning model g is defined either by conditional probability g(x) = P(Y|X) or by joint probability model f(x,y) = P(x,y). Empirical risk minimization (ERM) and structural risk minimization (SRM) [9] are commonly used for choosing g and f. In structural risk minimization based approaches the problem of over fitting is prevented by incorporating regularization penalty.

Support Vector Machine (SVM) [10] [11] [12] a very well known supervised learning algorith was first proposed by Vapnik. SVM is developed based on the structural risk minimization [9] principle derived from computational learning theory. SVM separates objects into different classes by defining a hyper plane in multidimensional space. SVM employs mathematical operator ϕ for mapping the training datapoints from input space to higher dimensional space. These mathematical opertor are often referred as Kernel. Then iterative training algorithm is used to define the separating hyperplane in that higher dimensional space by optimizing (minimizing) an error function. Based on the selection of the error function, SVM can also be categorized as i) C-SVM ii) nu-SVM iii) epsilon-SVM regression and iv) nu-SVM regression. For example, C-SVM has the error function

$$e = \frac{1}{2} * w^T w + C \sum_{i=1}^{N} \epsilon_i \tag{4}$$

subject to the constraints:

$$y_i(w^T\varphi(x_i) + b) \geq 1 - \epsilon_i \quad and \quad \epsilon_i \geq 0, \quad i = 1,..,N \quad (5)$$

Though linear hyperplane was originally proposed by Vapnik, but in 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik also modified the kernel function [13] to maximum-margin hyperplanes [14] for building a nonlinear classifier. Various types of kernel functions already exist in the literature for different applications e.g linear Kernel, Radial Basis Function (RBF), polynomial function etc.
Polynomial (homogenius) kernel- k$(x_i, x_j) = (x_i.x_j)^d$
Polynomial (inhomogenius) kernel- k$(x_i, x_j) = (x_i.x_j + 1)^d$
Gaussian radial basis function (RBF) kernel - k$(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$, for $\gamma > 0$. Sometimes $\gamma = 1/(2\sigma^2)$, where σ is described as the area of influence occupied by the support vectors over input data space. Several approach had already been proposed for multiclass SVM, few of them include

- one of the class label with respect to rest (one-versus-all)
- between each and every pair of classes (one-versus-one)
- Directed Acyclic Graph SVM [15]
- error-correcting output codes [16]

Crammer and Singer also proposed a multiclass SVM by considering entire classification objective as a single optimization problem rather than dividing it into multiple binary classification problems.

3 Experimental Results

We have taken 10 persons (7 male + 3 female subjects) dataset for training and testing. Initially, a single kinect is positioned at a fixed position to record the skeleton information at a distance of 6 feet from the subject. Then feature vector $\boldsymbol{f} \in \mathbf{R^{18}}$ is extracted at frame level for 10 subjects (A-J). As discussed earlier then we perform the noise removal using $\boldsymbol{f}_{static}$. After removing noisy frames, we store the feature vector $\boldsymbol{f}$ for rest of the frames in a dataset D. The dataset D is used for training model generation using multiclass SVM.
We have done the frame level performance analysis as well as comparison on the basis of F-score (6), which is defined as the harmonic mean of precision and recall. Here N is the number of subjects.

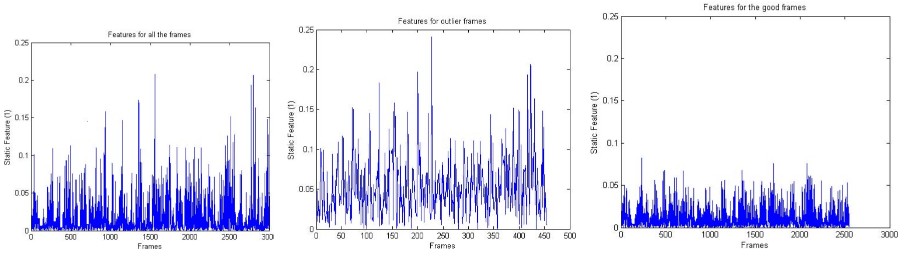
$$Fscore_i = \frac{2 * precision_i * recall_i}{(precision_i + recall_i)} \quad \forall i, 1 \leq i \leq N \qquad (6)$$

The performance analysis is done in 2 sections

- Effect of outlier removal
- Comparison with state-of-the-art systems.

3.1 Effect of Outlier Removal

As discussed in the section 2.3, skeleton information obtained from Kinect is very noisy [17]. Euclidean distance based outlier detection algorithm [4] is used to remove noisy frames. It is based on the fact that mutual distance between two physical joints should be constant over frames. Thus if the the joint-distance varies significantly from one frame to another frame, we remove those noisy frames from further processing. K-Means++ [6] algorithm is used for clustering $\boldsymbol{f}_{static}$ into two clusters - one containing noisy frames and other containing clean data. Figure 3 shows the cluster based analysis of noisy skeleton data. Figure 3(a), 3(b) and 3(c) represent the skeleton information for all the frames, for noisy frames and outcome of our proposed noise removal algorithm (i.e. noise clean skeleton data), respectively.



(a) for all frames (b) for noisy frames (c) noise clean data

Fig. 3. Sample static feature for different frames of a subject. The horizontal axis represents different frames and the vertical axis represents the normalized feature values.

3.2 Comparison with State-of-the-Art Systems

Performance evaluation of our proposed system is done at frame level with 30 seconds training data and 20 seconds testing data. A sample confusion matrix for 10 subjects marked as 'A' to 'J' is shown in the table 1. The diagonal entries of the matrix (shaded in grey) indicate correctly identified frames. Performance comparison is also performed with [18], [4] [1]. The results are tabulated in table 2. From table 2, it is very much clear that our proposed algorithm with noise removal technique is able to achieve 95.75% in real time and outperforms all the state-of-the-art methods.

4 Conclusion

Results indicate that our proposed angle and area related features with SVM based classification technique are having good contribution as the recognition accuracy has increased to 96% when only frontal standing pose is used for person identification. We are planing to do Kinect calibration to identify a multiple person in arbitrary poses using supervised learning.

Table 1. Confusion matrix for the proposed algorithm with 10 subjects

Subjects	A	B	C	D	E	F	G	H	I	J
A	2644	0	0	0	0	0	0	0	0	0
B	0	2647	0	0	0	0	0	0	0	0
C	0	17	2297	0	0	0	0	0	0	350
D	0	57	0	2608	0	0	0	0	2	1
E	0	0	0	0	2610	0	0	0	7	0
F	0	0	0	0	0	2654	0	0	0	0
G	0	0	0	0	0	0	2655	0	483	0
H	0	1	0	0	0	0	0	2653	0	0
I	0	0	0	1	0	0	0	0	2653	0
J	0	309	0	4	0	0	0	0	0	2334

Table 2. Performance (F_{score} in %) comparison with [18], [4] and [1]

Our Proposedwith frontal-standing	Using [18]	Using [4]	Using [1]
95.75	56	69	29

References

1. Preis, J., Kessel, M., Werner, M., Linnhoff-Popien, C.: Gait recognition with kinect. In: 1st International Workshop on Kinect in Pervasive Computing (2012)
2. Ball, A., Rye, D., Ramos, F., Velonaki, M.: Unsupervised clustering of people from skeleton data. In: 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 225–226. IEEE (2012)
3. Kumar, M., Babu, R.V.: Human gait recognition using depth camera: a covariance based approach. In: Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, p. 20. ACM (2012)
4. Sinha, A., Chakravarty, K.: Pose based person identification using kinect. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 497–503. IEEE (2013)
5. Microsoft: Kinect SDK (2012), `http://www.microsoft.com/en-us/kinectforwindows/develop/developer-downloads.aspx` (accessed February 6, 2014)
6. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, pp. 1027–1035 (2007)
7. Mohri, M., Afshin Rostamizadeh, A.T.: Foundations of Machine Learning. The MIT Press (2012)
8. Wikipedia: Supervised learning — wikipedia, the free encyclopedia, `http://en.wikipedia.org/wiki/Supervised_learning/` (accessed February 6, 2014)
9. Vapnik, V.N.: The nature of statistical learning theory. Statistics for engineering and information science. Springer, Berlin (1999)
10. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
11. Cortes, C., Vapnik, N.V.: Support-vector networks. Machine Learning, 20

12. Wikipedia: Support vector machine — wikipedia, the free encyclopedia, `http://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=548461902/` (accessed February 6, 2014)
13. Aizerman, M.A., Braverman, E.M.: Rozonoer: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837 (1964)
14. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM (1992)
15. Platt, J.C., Cristianini, N., Shawe-taylor, J.: Large margin dags for multiclass classification. Advances in Neural Information Processing Systems 12(3), 547–553 (2000)
16. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286 (1995)
17. Newcombe, R.A., et al.: Kinectfusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 127–136. IEEE (2011)
18. Sinha, A., Chakravarty, K., Bhowmick, B.: Person identification using skeleton information from kinect. In: ACHI 2013, The Sixth International Conference on Advances in Computer-Human Interactions, pp. 101–108 (2013)

A Virtual Handwriting Tablet Based on Pen Shadow Cues

Chin-Shyurng Fahn, Bo-Yuan Su, and Meng-Luen Wu

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan 10607, Republic of China
{csfahn,a9815001,D10015015}@mail.ntust.edu.tw

Abstract. The handwriting tablet is an electronic product, which is a kind of human-computer interfaces acting as a computer input device comprising a set of a special pen and a tablet. The user holds the pen to draw contents within a region of the tablet as inputs, which imitates handwriting and is a replacement of mouse inputs. Some handwriting tablets not only imitate the handwriting and mouse functions, but also detect the pen tilts and pressures. The tilt and pressure information can be applied to some drawing software which can also render the thickness and depth of strokes. However, since the handwriting tablet is a piece of precise equipment, it has some drawbacks- fragile, not easy to carry, and the weight is often heavy. Therefore, in this paper, we propose a new concept based on the computer vision technology to simulate the handwriting tablet. We put a rectangular plane in the FOV of a video camera to emulate a tablet, and use a conventional pen to emulate the stylus. Many experiments have been made for evaluating the effectiveness of the proposed methods. The performance of such a virtual handwriting tablet is very satisfactory and encouraged.

Keywords: virtual handwriting tablet, shadow cues, computer vision, human-computer interface.

1 Introduction

With the evolution of information technology, there are various means of human-computer interfaces. In order to make computers more convenient and interactive to control, several ways of controls have been developed, such as joysticks, hand gestures, and pens. Nowadays, human-computer interfaces without keyboards and mice are a trend because they provide users different experiences. For example, in racing games, wheel shaped joysticks are created to simulate handlers for actual cars, which gives players a more entertaining experience. Input texts via handwriting tablets supplies users a feel of writing.

The handwriting tablet is a kind of computer input devices, which is an electronic product composed of a special pen and a tablet. The user holds the pen to draw contents within a region of the tablet as inputs, which imitates handwriting and replaces mouse inputs. Some handwriting tablets not only imitate the handwriting and

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 224–233, 2014.

mouse functions, but also detect pen tilts and pressures. Both the tilt and pressure information can be applied to some drawing software that the thickness and the depth of strokes can be also rendered [1].

However, since the handwriting tablet is a piece of precise equipment, it has some drawbacks- fragile, not easy to carry, and too heavy. Therefore, we adopt the computer vision technology together with a few materials to simulate the handwriting tablet. This novel concept is to develop a virtual handwriting tablet as shown in Figure 1, where we put a rectangular plane (that can be a cardboard, corrugated paper, and so on) in the FOV of a video camera to emulate a tablet, and use a conventional pen to emulate the stylus (the pen for handwriting tablet) [2],[3].

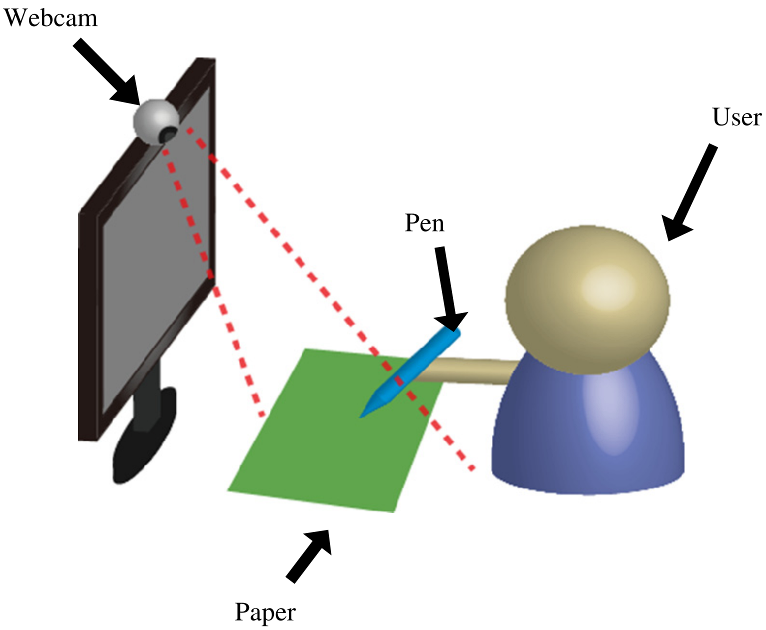


Fig. 1. The concept of the virtual handwriting tablet.

The development of such a virtual handwriting tablet is divided into three parts, including tablet detection, stylus detection, and pen shadow detection. First, the user selects an object in the FOV as the tablet, and the system determines whether the selected object meets the conditions on the form of a tablet. After the tablet object is confirmed, the system detects a pen in the tablet region. If so, the shadow of the pen is then detected. The tablet, pen, as well as shadow information are saved for the system. Figure 2 graphically shows the main flowchart of our virtual handwriting tablet system.

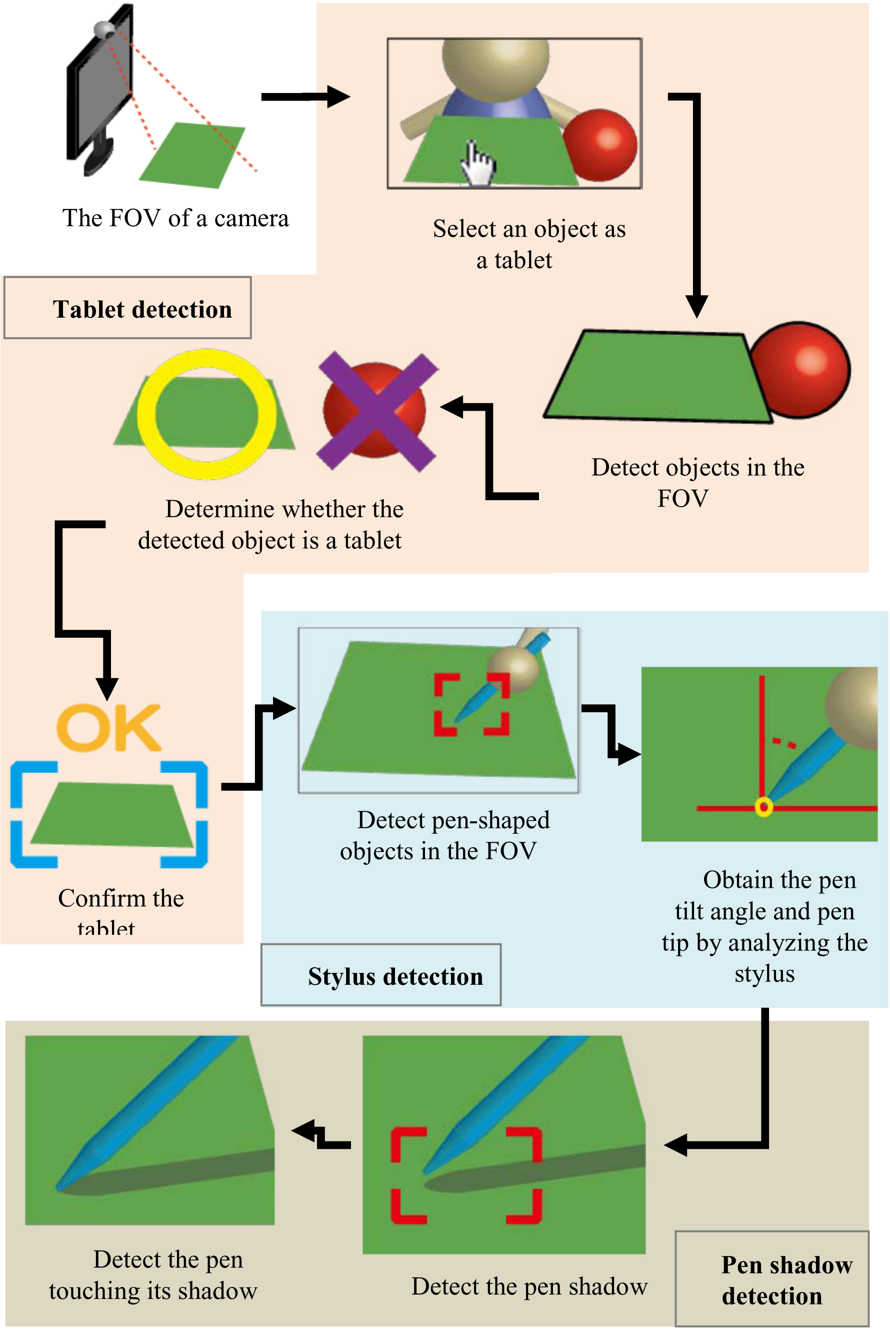

The FOV of a camera
Select an object as a tablet
Tablet detection
Detect objects in the FOV
Determine whether the detected object is a tablet
OK
Confirm the tablet
Detect pen-shaped objects in the FOV
Obtain the pen tilt angle and pen tip by analyzing the stylus
Stylus detection
Detect the pen touching its shadow
Detect the pen shadow
Pen shadow detection

Fig. 2. The main flowchart of our virtual handwriting tablet system

2 The Relation of Pen and Its Shadow

The following elaborates how we acquire the distance between the pen and its shadow used for the detection of a pen touching or detaching a tablet. After the area of the shadow is obtained, we can use the bounding box of the shadow to get the boundary and position of the area. By analyzing the variation of the distance from a pen to its shadow, we can detect the relation between the pen and the tablet. Figure 3 illustrates the pen under different tilt degrees in several sampling frames that the pen approaches the tablet and leaves it gradually.

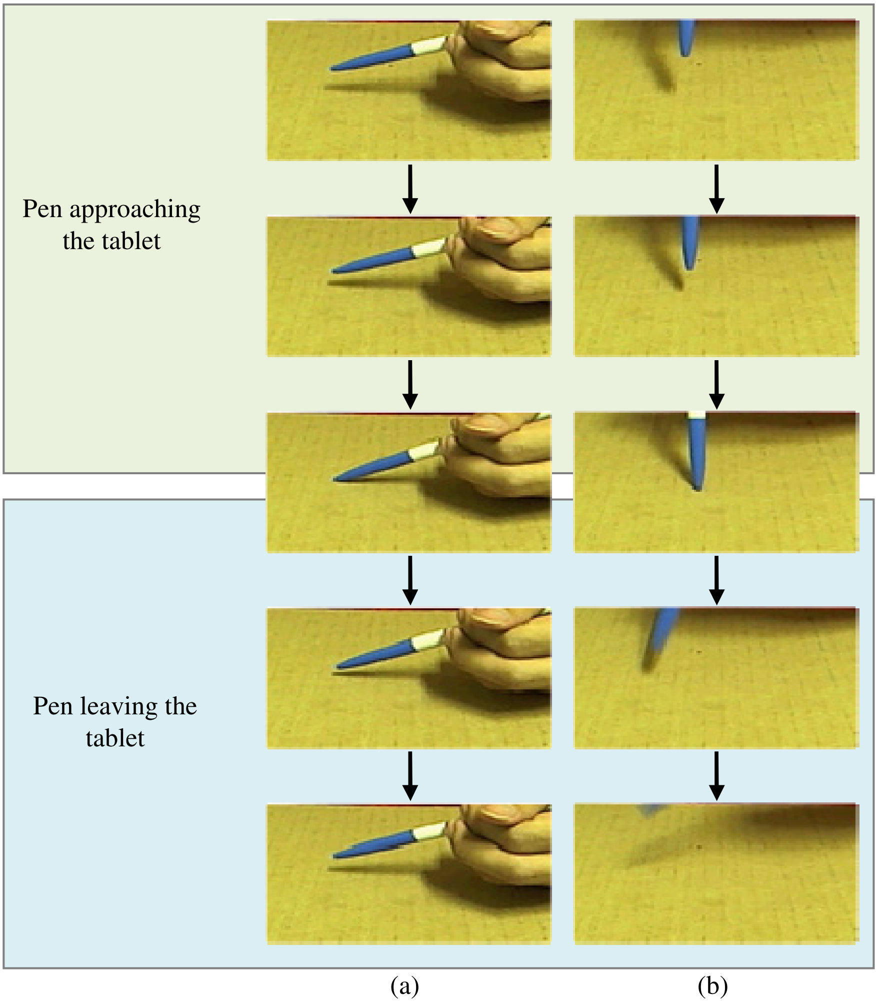


Fig. 3. A pen approaching a tablet and then leaving it: (a) the pen is tilted by 63 degree; (b) the pen is upright; that is, tilted by 0 degree

As seen from Fig. 3, when the pen approaches and leaves the tablet, there is an interval that the pen approaches and leaves its shadow. This interval can be obtained from capturing a sequence of consecutive frames by calculating the distance between the pen and its shadow in each of the frames respectively. Equation (1) formulates the distance between the pen and its shadow, where i stands for the index of a frame, SB is the bounding box of the shadow, and PB is the bounding box of the pen. The distance is acquired from Eq. (1), where SB_i^{Top} is the upper side of shadow's bounding box and PB_i^{Bottom} is the bottom side of pen's bounding box as Fig. 4 shows.

$$D(i) = SB_i^{Top} - PB_i^{Bottom} \tag{1}$$

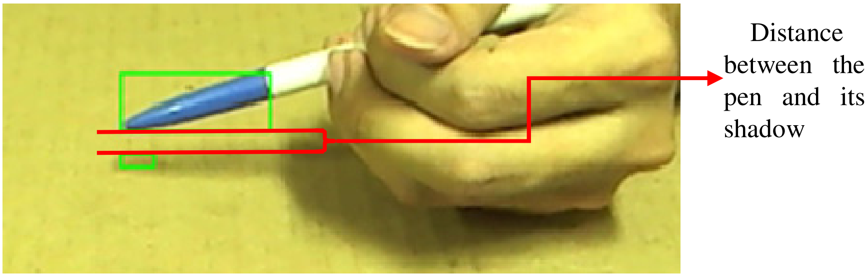


Fig. 4. Illustration of the distance between the pen and its shadow

Let n be the index of the current frame, and $D(n)$ be the distance between the pen and its shadow in the current frame. We calculate the differences between $D(n)$ and $D(n-1)$, $D(n-1)$ and $D(n-2)$, $D(n-2)$ and $D(n-3)$, and so on. Then the average of the differences is computed, which is called the average distance variation $\overline{D}_{var}$, as expressed in Eq. (2), where m is the number of frames needed for computation.

$$\overline{D}_{var} = \frac{1}{m}\sum_{k=0}^{m-1} D(n-k) - D(n-k-1), \qquad m \geq 2 \tag{2}$$

By means of the average distance variation, we can detect the pen action conducted by the user. This detection approach is depicted in Eq. (3) and stated as follows. Theoretically, $\overline{D}_{var}$ is positive when the pen is leaving the tablet. Conversely, $\overline{D}_{var}$ is negative when the pen is approaching the tablet. And T_{adv} is the threshold for detecting the variation, which is employed to discard frames with minor variation. If the variation is smaller than the threshold, the current result will not be updated.

$$Pen\ Action = \begin{cases} up, & if\ \overline{D}_{var} > T_{adv} \\ down, & if\ \overline{D}_{var} < -T_{adv} \end{cases}, \qquad T_{adv} \geq 0 \tag{3}$$

While the shadow detection is set to a very small area, we can deem these two actions as a pen touching and detaching the tablet, respectively. This is because when the detection area is very small, the pen needs to be put very close to the tablet. While the pen is quite near the shadow, it implies that the pen is very close to the tablet. Consequently, the characteristic of this behavior can be used for detecting a pen touching or detaching the tablet.

3 Experimental Results and Discussions

Many experiments of tablet detection, stylus detection, and pen shadow detection have been carried out to demonstrate the effectiveness of our proposed methods. Table 1 lists the developing environment for creating a virtual handwriting tablet system. In this experimental system, we adopt a webcam mounted on the monitor of a notebook computer, which is modeled Logitech's HD Pro Webcam C910 to support Full HD 1080p recording as shown in Fig. 5(a). The video camera is set to capture an image every 50 ms, which means the FPS set to 20. And the image resolution is set to 640×480.

Additionally, we take a traditional ball pen for emulating a digital stylus, and choose a piece of corrugated paper for simulating a tablet. The experimental set is shown in Fig. 5(b), where a piece of corrugated paper is put in front of the notebook computer, and a keyboard is placed between the piece of corrugated paper and the monitor. Figure 6(a) shows a real FOV of the camera, and the experimental ambient environment is illuminated by LED lights as shown in Fig. 6(b).

Table 1. The Developing Environment for Creating Our Virtual Handwriting Tablet System

Hardware	
CPU	Intel(R) Core(TM) i7 CPU Q740 @ 1.73GHz 1.73GHz
RAM	DDR3 4.00GB
Software	
Operating System	Microsoft Windows 7 Ultimate 32-bit
Developing Tools	Microsoft Visual Studio 2010 Microsoft .NET Framework 4 C# Language OpenCV (EmguCV) 2.4.2

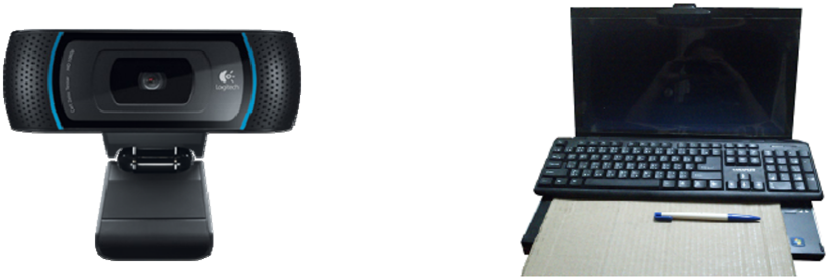
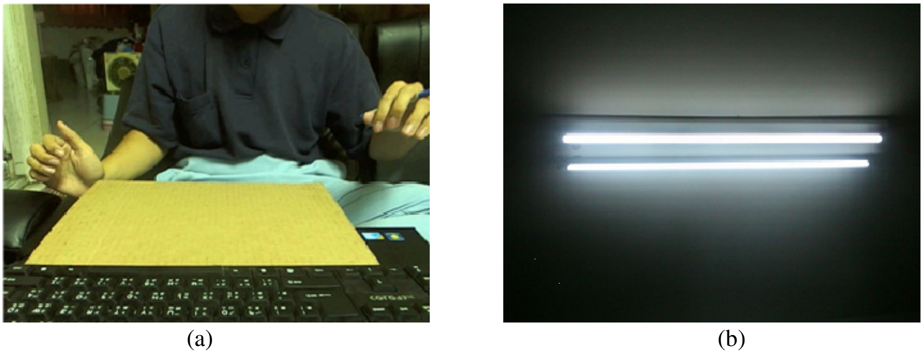

(a) (b)

Fig. 5. Experimental equipment: (a) the used camera modeled Logitech HD Pro Webcam C910; (b) the set of our virtual handwriting tablet system



(a) (b)

Fig. 6. Experimental environments: (a) an FOV seen from the webcam; (b) an ambiance illuminated by LED lights

Because our virtual handwriting tablet system takes advantage of pen's shadow for stylus detection, the position of a light source and the placement of its constituting components are very important. Figure 7 illustrates the deployment of the light source and each component of the system.

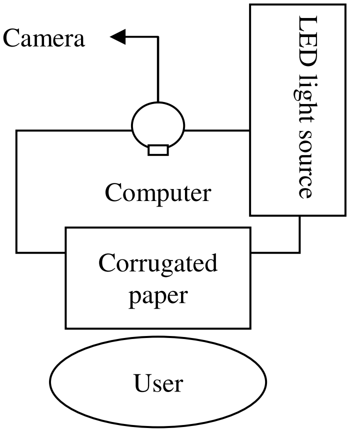
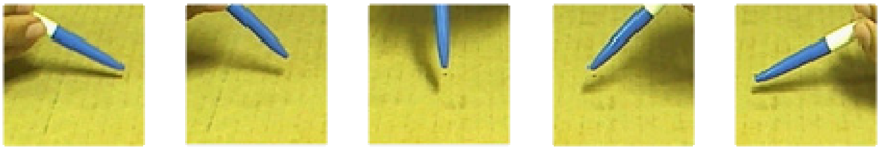


Fig. 7. The deployment of all the light source and components constituting the virtual handwriting tablet system

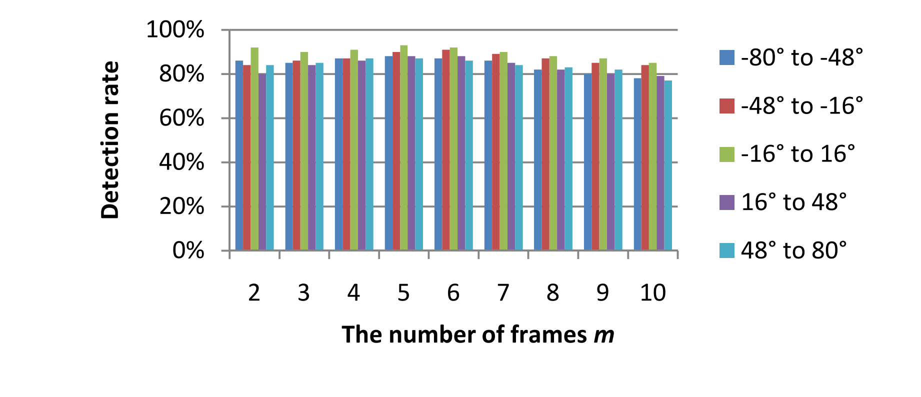
The following performs experiments on the methods proposed in Section 2 to detect pen moving directions with unlike parameters. The form of pen's shadow is varied along with different pen tilt angles as shown in Fig. 8. Therefore, we divide the pen tilt angles (-80° to 80°) into five equal portions, each of which possesses an angle of 32 degrees; that is, ranged from -80° to -48°, from -48° to -16°, from -16° to 16°, from 16° to 48°, and from 48° to 80°. Then we make the pen approach and leave the tablet for 100 times under different pen tilt angles, where the distance between the pen and the tablet is about 1cm when approaching and leaving. We record the detection rate for each of the above five ranges, respectively. The detection of pen moving directions is correct only when the shadow is detected accurately on both approaching and leaving at that time. We want to find which parameter achieves the best detection rate.



-80° to -48° -48° to -16° -16° to 16° 16° to 48° 48° to 80°

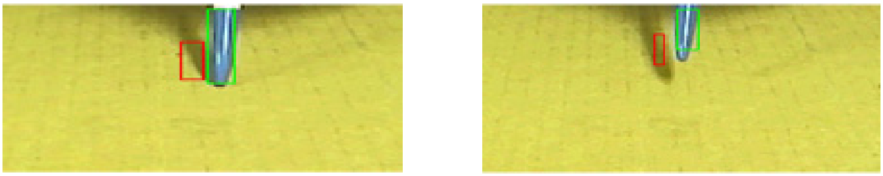
Fig. 8. Pen shadows varied with different pen tilt angles

Figure 9 shows the best detection result is acquired from the pen tilt angles between -16° and 16°. This is because the shadow appears in the side of the pen shown in Fig.10 for these angles when the pen approaching and leaves the tablet.



m	2	3	4	5	6	7	8	9	10
Average detection rate	85.20%	86.00%	87.60%	89.20%	88.80%	86.80%	84.40%	82.80%	80.60%

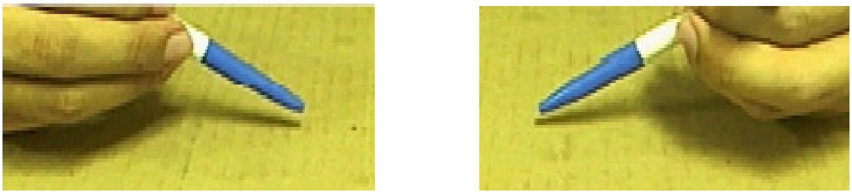
Fig. 9. The detection rate under different pen tilt angles for $m = 2, 3, \ldots, 10$



(a) (b)

Fig. 10. The correct detection result: (a) the pen touching the tablet; (b) the pen leaving the tablet under its tilt angles between -16° and 16° for $m = 5$

Besides, it is easily found that the detection rate is higher when the degree of the pen tilt angle is negative. This is caused by the light source located at the upper right corner. As a result, the produced shadow in Fig. 11(a) is more saturated than that in Fig. 11(b), since in Fig. 11(a), the direction of the pen is perpendicular to that of the light source; conversely, in Fig. 11(b), the direction of the pen is parallel to that of the light source.



(a) (b)

Fig. 11. Varied saturation of the produced shadow: (a) a more saturated shadow (the pen tilt angle = -63.43°); (b) a less saturated shadow (the pen tilt angle = 63.43°)

After observing the result, it is clear that the detection rate starts to decrease from $m = 6$. We find this is owing to the relation between the FPS and user's pen movement speed. When the user moves the pen to leave the tablet needs 5 frames, but we set to 10 frames, the additional 5 frames decreases the average distance variation. On the contrary, if it is set to 2 frames, the pen will be detected to touch the tablet before it really does. In this case, any small movements of the pen will cause wrong detection. For the same FPS, if m is small, the pen movement must be fast to reach an ideal detection rate. The best condition is that the number of frames for movement is equal to m.

4 Conclusions and Future Works

Nowadays, the human-computer interface is a popular research field, such as gesture recognition and virtual keyboard. These new interfaces bring users different kinds of experiences. Our proposed system is a new human-computer interface which adopts computer vision technology to imitate the function of a handwriting tablet, and improves the shortcomings that a traditional handwriting tablet may have; for example, not easy to carry, heavy weight, as well as fragile.

We have further proposed a method, which can detect a pen overlapping a tablet in the FOV of a camera. By detecting the variation of pen's shadows, the moving direction of a pen can be detected, which makes it possible to detect the pen touching a tablet with a single camera.

The system can be divided into three parts, including tablet selection, stylus detection, and pen shadow detection. The users select an object to be the tablet, and the system decides whether the selected object meets the conditions of a tablet. The system then finds the pen within the region of the tablet, and detects the position of the shadow of the pen. At last, the obtained information is saved for our system to send to the computer.

In the experiments, we employ different parameters for the proposed system, and analyze the experimental results. We find the optimal parameters and discuss about them. This paper has presented a novel human-computer interface and its applications. There are many research directions for the future work:

(1) Multi-touching

The proposed system can only detect one pen at a time. In the future, the system can be extended into multiple pen detection, in order to reach the goal of multi-touching.

(2) Right button click detection

Currently, our system is designed to emulate the movement of a mouse and the click of a left button. A right button click detection mechanism can be designed in the future.

(3) Detection by shadow only

Because the direction of a light source may vary, we cannot detect whether a pen has touched the tablet by pen shadow detection only. Therefore, we choose to set the range of capturing shadows to help detect the movement of the pen. However, this approach provides only a detection of the pen approximately touching a tablet. In consequence, we will intend to develop other effective methods to detect a pen touching a tablet.

(4) Writing assistance

Currently, we merely provide an object detection method. In the future, we will further combine a writing assistance system, which can help smooth the moving path, tablet calibration, as well as stop word detection.

Acknowledgement. The authors thank the National Science Council of Taiwan (R. O. C.) for supporting this work in part under Grant NSC-101-2221-E-011-143-MY3.

References

1. Wacom Intuos Pen | Wacom | Wacom Taiwan, `https://www.wacom.asia/intuos5/wacom-grip-pen` (accessed on January 12, 2014)
2. Intuos Pro | Wacom | Wacom Taiwan, `https://www.wacom.asia/tw/intuos-pro` (accessed on January 12, 2014)
3. Product List | Wacom | Wacom Taiwan, `https://www.wacom.asia/tw/products-price` (accessed on January 12, 2014)

HOUDINI: Introducing Object Tracking and Pen Recognition for LLP Tabletops

Adrian Hülsmann and Julian Maicher

University of Paderborn, Department of Computer Science, Germany
adrian.marius.huelsmann@uni-paderborn.de,
jmaicher@mail.uni-paderborn.de

Abstract. Tangible objects on a \tabletop offer a lot of different opportunities to interact with an application. Most of the current tabletops are built using optical tracking principles and especially LLP tabletops provide very good tracking results for touch input. In this paper we introduce HOUDINI as a method for LLP object tracking and pen recognition, which is based on three different sizes of touch points that help us to identify touch points belonging to fingers, objects and pens. As a result, the whole recognition process is performed at the level of touch information rather than frame by frame image analysis. This leads to a very efficient and reliable tracking, thus allowing the objects to be moved very fast without being lost.

Keywords: tabletop, interactive surface, object tracking, LLP, pen recognition.

1 Introduction

Tangible objects provide a natural way of interaction between tabletop and user. They offer opportunities which pure touch input does not offer, for example adding tactile feedback [20], allowing intuitive map navigation [15], a more precise adjustment of parameters [18] [6] [11], solving input conflicts [13] or improving awareness in co-located collaborative group settings [17].

Tabletops are commonly based upon optical tracking technologies. In the process, the tabletop surface is enriched with infrared (IF) light that gets reflected down into one or more IF-camera(s) as soon as fingers or objects hit the surface. Then it is further processed within the tracking software before the extracted touch information is sent to the application.

The predominant methods for optical tracking are *Diffused Illumination* (DI), where the IF-illumination takes places from below the tabletop surface, *Frustrated Total Internal Reflection* (FTIR)[5], in which IF-light is brought into the surface from the side where it is trapped (because of the equally named physical principle) until a touch of a finger allows it to reflect down into the camera(s), *Diffused Surface Illumination* (DSI)[2], which similar to FTIR uses IF-light coming from the side, but also requires a special Endlighten™ acrylic as surface material, and *Laser Light Plane* (LLP), where laser beams establish a plane of IF-light just above the tabletop surface that gets scattered at every touch point [16].

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 234–244, 2014.

Each of those principles has its own advantages and disadvantages and raises special problems. Besides touch input, DI also allows object recognition through attached fiducial symbols that are easily made out of printed paper [9]. But for good tracking results, DI depends on an equally illuminated surface, which is not easy to achieve from below, so that this method sometimes suffers from false inputs. Whereas FTIR is not able to detect fiducial symbols, it is very good for the tracking of fast finger movements, because of the IF-lighting delivering a camera image that is rich in contrast. However, FTIR needs an acrylic surface to ensure the physical effect of total internal reflection. As a consequence, very big tabletops are hardly to achieve, since acrylic is not as stable in shape as glass and deflects with bigger dimensions. To counter this, the acrylic needs to be considerably thicker; this rapidly increases its price.

DSI also allows object recognition through fiducial symbols, but it needs a rather expensive Endlighten™ acrylic with the same downside concerning form stability mentioned above. Since in LLP tabletops the surface is not directly illuminated and the plane of IF-light only gets scattered during touch input, the camera image is very rich in contrast, thus enabling a fast and reliable tracking of fingers, even during fast movements.

In addition, LLP tabletops can be built with much bigger dimensions, because stable glass panes can be used, which are also significantly cheaper than acrylic. By the use of lasers, the illumination also becomes independent from the tabletop size, in contrast to DI, FTIR and DSI, where a bigger surface requires more or stronger IF-light emitters.

In summary, LLP tabletops enable a fast and reliable tracking of fingers and can be built with much bigger dimensions. But due to the missing possibility to track objects they are not very common. That is why in this paper we will introduce a first method for LLP object tracking and also pen recognition, which is based on touch recognition and therefore results in a reliable tracking, even when objects are manipulated very fast.

2 Related Work

reacTIVision [9] is a computer vision framework using fiducial symbols for object tracking. The symbols consist of black and white patterns that are recognized by a camera and through image analysis these patterns are subsequently transferred into region adjacency graphs for object differentiation. Fiducials only work for tabletops based on DI and DSI since the marker patterns must be equally illuminated from the bottom.

ToyVision [12] also uses fiducials. In addition, the authors added features to augment the objects, giving the opportunity to manipulate objects while using them. To give an example, a button was added to the object which can be pressed to interact with the application.

There are other tracking approaches which use RFID tags to identify objects. In [14] this idea is combined with the camera image of the surface. Thus, giving the opportunity to identify which shadow represents which object and being able to track the objects by simply tracking the shadows.

Similar to our approach, in [7] markers are placed onto acrylic disks to allow object recognition. Nevertheless, this approach uses conventional FTIR and DI tracking instead of LLP. Objects are recognized on the basis of image analysis, the markers are elastic and used for pressure sensing.

3 Basic Idea for LLP Object Tracking and Pen Recognition

LLP tabletops are characterized by the fact that every finger, but also every general object breaking the laser light plane, causes a reflection of IF-light down into the camera and thus creates a touch point (blob) in the tracking software.

With this paper, we introduce a solution to distinguish between blobs caused by *fingers*, *objects* or *pens*. Our solution is based on *three different sizes of blobs* in combination with *specific design constraints* that help us to identify blobs belonging to objects and pens.

The presented HOUDINI system is designed as TUIO [10] proxy between the widely spread tracking software *Community Core Vision* (CCV) [3] and an arbitrary tabletop application. Basically, from the set of blobs delivered by CCV, we at first exclude blobs belonging to objects and pens. While bypassing all other blobs to the application, we at second add messages for the identified object and pens, hence splitting the initial blob set into messages for fingers, objects and pens.

Before we describe the design decisions and the implementation in more detail, we at first introduce some fundamental requirements that have been the center of the development.

4 System Requirements and Goals

Our goal was to implement a system that meets several requirements, which we consider to be fundamental for tangible interaction on tabletop displays.

- The system should support tangibles of different sizes.
- Multiple objects on the tabletop surface should be reliably tracked at the same time.
- Tracking data should include translation, rotation and basic state information (pressing a button) of tangible objects.
- Object tracking should not limit the speed of processing touch input data of fast moving fingers
- The system should be able to support translucent tangibles that offer additional possibilities for application design and reduce the amount of surface occlusion.
- The tangibles themselves should be passive, i.e. without additional electronics or batteries. This also holds for the pens used as input devices on the tabletop surface.

These are core requirements in the sense of general-purpose tangibles to be manipulated on the tabletop surface mostly by translation and rotation. The option to extend these basic tangibles with a physical button triggering a state change could be used in many application scenarios, for example to confirm a prior selection in the graphical user interface.

5 Design of Tangible Objects and Pens

Similar to other object recognition techniques, the fundamental idea of our approach is to attach markers at the bottom of objects forming individual *patterns* that can be identified by the system. But due to the characteristics of LLP, we do not rely on visual symbols and instead specify patterns as a set of touch points arranged in a certain way. This allows us to perform the object recognition at the level of touch information rather than analyzing the video stream from the camera frame by frame, as it is done by reacTIVision using fiducial markers [9].

5.1 Object Pattern Definition and Derivation of Properties

Basically, our tangible objects consist of acrylic discs and attached beveled markers which reflect the infrared light plane down into the camera and hence create touch points (see Figure 1). As mentioned before, we use different blob sizes to distinguish between fingers, objects and pens. In this context, we specify an *object pattern* as a set of touch points consisting of several *small blobs* and exactly one *big blob.*

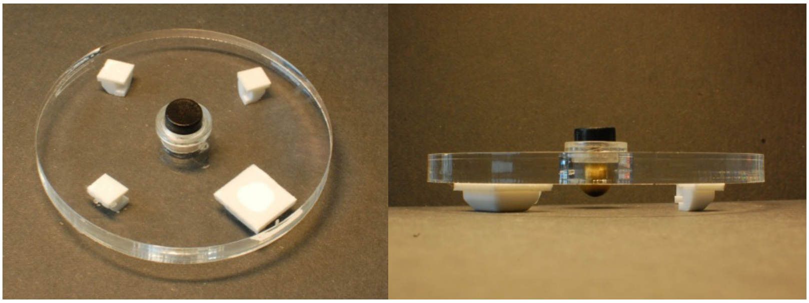


Fig. 1. Tangible object with attached markers for creating blobs of different sizes

During the manipulation of objects on a tabletop, the most significant properties to track are the object's *position* and *rotation angle*, which in our approach both must be derived from the object's pattern, e.g. a set of several small blobs and exactly one big blob. The position of the object is treated as the center of the *minimum covering circle* of all blobs which can be calculated in linear time [21]. The rotation angle α of an object can be computed as shown in Figure 2. It is clearly determined by the vector from the object's position towards the position of the big blob and the x-axis of the underlying coordinate system.

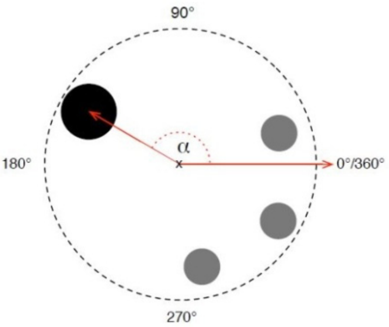


Fig. 2. Computation of the rotation angle. The rotation angle α of an object is determined by the vector from the center of the minimum covering circle to the position of the big blob.

As shown in Figure 1, objects can also have a physical button in order to trigger events. This is very useful for situations where an object is for example used to invoke a menu in which rotation of the object is used for selecting one of the menu items and the physical button is used for confirmation [12]. When pressing the button an *activation blob* appears on the surface underneath the object. The activation blob represents the button and is stored in addition to the other blobs in the pattern. When it is recognized, the object is considered to be active. Then the marker ID of the object is negated to notify applications that the button has been pressed.

5.2 Pen Pattern Definition

The same principle of different blob sizes applies to the identification of pens that are used for tabletop interaction. Normally, when writing with a pen, the palm of the hand naturally rests on the writing surface. We turn this to our profit, by interpreting a *very big blob* as a resting palm. An occurring small blob located in a circular area with a certain distance to this very big blob is then interpreted as the tip of a pen (see Figure 3). As a result, the small blob's marker ID is modified to notify applications that it belongs to a pen tip touching the surface.

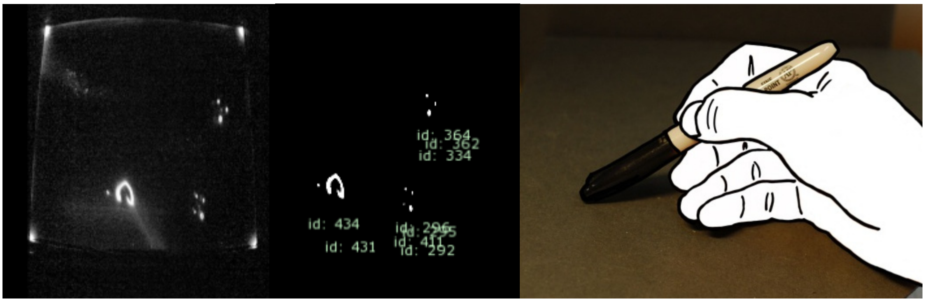


Fig. 3. Palm triggers pen recognition (right). Blob patterns of two objects and one pen (left).

The advantage of this method is in the lightweight design without any additional electronic parts and the possibility to use standard office pens. Obviously, these should be capped or used with the end of the pen. Another benefit is that there is no need for a special writing surface overlay, which considerably reduces contrast of the visual display, for example used with the widely spread Anoto Digital Pen [1].

6 Implementation

HOUDINI is designed as TUIO [19] proxy between the widely spread tracking software *Community Core Vision* (CCV) [3] and an arbitrary tabletop application. As mentioned before, the system differentiates between *small*, *big* and *very big* blobs. We therefore use a modified version of CCV, which adds the size of the blobs to the data of touch points sent via the TUIO protocol using the */tuio/2Dblb* message profile.

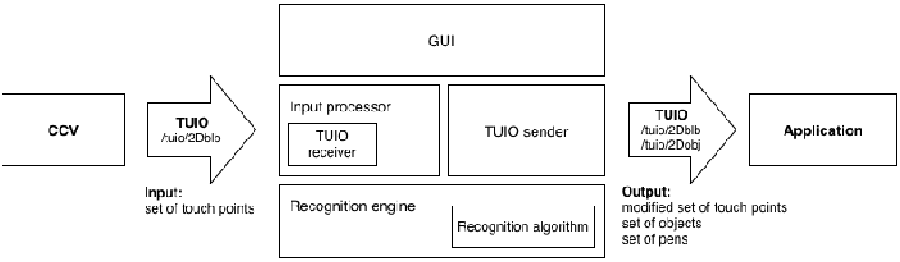


Fig. 4. Overview about the architecture of HOUDINI

The system listens for touch input on the port used by CCV and, if necessary, augments the input with object or pen-related information before sending it on a different port to an application using the */tuio/2Dblb* message profile for touch points belonging to no pattern (fingers) and the */tuio/2Dobj* message profile for identified objects or pens.

6.1 Architecture

HOUDINI is composed of four main building blocks which are shown in Figure 4. The *input processor* is responsible for receiving TUIO bundles, analyzing and, if necessary, filtering therein contained SET and ALIVE messages and afterwards forwarding them to the *TUIO sender* for dispatching. The *recognition engine* controls the recognition algorithm, supplies its input and processes the results of a recognition cycle. The *GUI* provides configuration options, a live view and pattern management functionalities for the user.

6.2 Recognition Algorithm

The recognition algorithm identifies objects by comparing the blobs on the surface to the patterns in the database and identifies pen input by looking for the separately specified pen pattern. The algorithm basically can be divided into two parts.

First, the *recognition part*, wherein the algorithm is working on all current blob messages delivered by CCV and tries to find blobs belonging to objects and pens. For already known objects or pens, the algorithm creates a new message with updated parameters which is sent to the application. For new objects and pens, the algorithm stores the IDs belonging to the blobs of the object or pen for later identification and creates a new object message containing position, rotation angle (for objects only) and ID of the object or pen. Afterwards this message is sent to the application.

Since objects have a unique pattern, whereas pens all share the same pattern, the system is not able to differentiate between pens after they lost contact to the surface. The following pseudocode illustrates the recognition process of objects in more detail.

```
for all bigBlob ∊ BigBlobs do
  for all pattern ∊ Patterns do
    collect blobs matching distance in pattern
    if more blobs found than needed then
      create subsets from the set of found blobs
    else if # of blobs equals # of blobs in pattern then
      put all found points into subset.
    end if
    for all s ∊ Subsets do
      compute minimum covering circle for s
      compute vector from center of circle towards
      big-Blob
      for all blob ∊ s do
        check whether angle of blobs matches angles
        in pattern
      end for
      if all blobs match position of pattern then
        pattern found! add to results.
      end if
    end for
  end for
end for
```

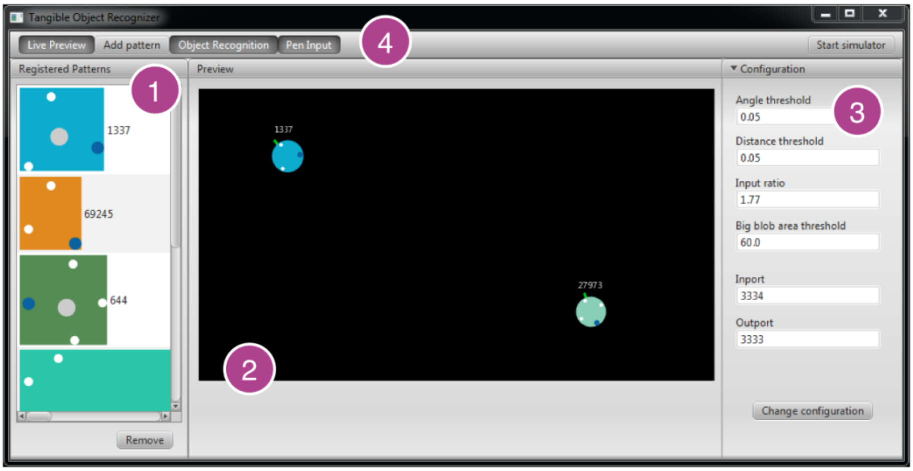
After receiving the blob messages from CCV, the algorithm collects all big blobs on the surface that have not been assigned to an object. For each of these big blobs the algorithm tries to find small blobs around it, which match the distance specified in one of the patterns. Since in this step only the distance towards the big blob is taken into account, there may be multiple points matching the specified distances. If this is the case, subsets are created, such that for each pattern point exactly one blob, which matches the distance, is taken into account. Hence, each subset consists of the big blob and one blob for each pattern point. As a next step, subsequently for each subset, if at least one is found, the minimum covering circle is computed. After that, the algorithm checks for each small blob if it is at the correct position. In order to do so, the vector from the center of the minimum covering circle towards the big blob is computed. This vector is then used to compute all angles for all other blobs and to compare them to the corresponding angles in the pattern. If both, the distance and the angle matches, the blob is at the correct position. The vector from the center to the big blob also determines the rotation angle of the object. The IDs of the blobs which have been assigned to an object are stored for future reference. These blobs can be ignored by the algorithm when searching for new objects.

The second part of the algorithm is the *reassignment of missing touch points*. As written above, if an object is recognized, all touch points and their IDs are assigned to this object. That way, if a blob disappears for a short time and then reappears with a new ID, the algorithm is able to replace the missing ID with the new one. The benefit of trying to reassign blobs compared to directly removing the object and recognizing it later is that this method reduces object-added and object-removed events, thus provides the application with more stable objects.

6.3 Graphical User Interface

HOUDINI's graphical user interface allows an easy setup and on-the-fly configuration of the tracking environment. The four main components are shown in Figure 5.

1. The registered patterns are shown at the left side including a thumbnail containing the blobs which form the pattern. Patterns are color-coded in order to give a visual connection between patterns and actual recognized objects in the live preview. Patterns can be added and removed from the library easily e.g., when a pattern should not be tracked anymore.
2. The live preview is a miniature view of the whole tabletop. It visualizes all blobs and derived tangible objects and pens.
3. Configuration parameters are shown at the right side and allow the modification of thresholds between the different blob sizes.
4. The main functions are arranged in a menu bar at the top. The live view, object and pen recognition can be switched on and off and new patterns can be registered.



Angle threshold
0.05

Distance threshold
0.05

Input ratio
1.77

Big blob area threshold
60.0

Inport
3334

Outport
3333

Fig. 5. HOUDINI's graphical user interface showing the main components

7 Conditions for Proper Tracking

At first, marker points attached to objects must be at a certain distance to each other in order to avoid melting of two blobs into one. Therefore, tangible objects must have a certain size, whereby bigger objects provide more space for the marker points. As a result, there is a clear link between the object size and the number of distinguishable patterns. An exact number of distinguishable patterns is difficult to estimate, but we successfully constructed ten different patterns using acrylic discs with a diameter of 80mm (with a tabletop dimension of 100cm x 160cm) that can be used at the same time without any problems and we see potential for a variety of more. At second, on LLP tabletops many marker points may lead to shadowing of blobs. We therefore recommend using at least one IF-laser per corner for proper results.

Concerning the user experience, we argument that a reliable tracking of less objects is more important than being able to differentiate between many objects at the expense of tracking quality. The real benefit of HOUDINI is that it is based on pure blob data, which are very stable and reliable in LLP tabletop environments, because of high contrast between touched and untouched surface areas. HOUDINI builds upon that feature by implementing object and pen recognition on TUIO messages only and with no need for complex image analysis or processing raw image data.

Therefore, HOUDINI is most suitable for applications scenarios with fewer objects, ranging from using tangibles to control parameters of music mixing interfaces [4], to physically navigate across a map [22], activate widgets and menus [8] or playing games which make use of fast moving objects like air hockey.

8 Conclusion and Future Work

We have presented HOUDINI as a TUIO proxy between the tracking software CCV and an arbitrary tabletop application and therefore implemented a solution for an object and pen recognition technique that is designed for LLP tabletops. It also completely fulfills the initially mentioned requirements.

As a result, it supports translucent tangibles of different sizes, whereby multiple objects and pens are reliably tracked at the same time without limiting the speed of processing touch input data. Object messages include translation, rotation and basic state information (pressing a button), while pens only contain information about the position and whether they touch the surface or not. Further, multiple pens can be tracked at the same time, but they cannot be distinguished from each other, because they all share the same blob pattern. This also results from the last requirement which asks for a solution without additional electronics or batteries. Instead, standard office pens can be used, which contribute to a better user experience and seamless integration of scenarios using tabletop interaction in combination with writing on real paper.

By focusing on patterns based on touch points and only using the TUIO messages for calculating the object properties, we were able to realize reliable tracking results, even during fast movements of the tangible objects.

Necessarily, our approach leads to a dependency between the object size and the amount of distinguishable patterns. We therefore do not propose our system for applications that intend to use many different tangibles.

Instead, we think HOUDINI is appropriate for the remaining majority of application scenarios, where only a few tangible objects are used, but which then can be moved very fast without being lost during the tracking.

References

1. Anoto Digital Pen, `http://www.anoto.com`
2. Akechi, N., Mizumata, T., Sakamoto, R.: Hovering fingertips detection on diffused surface illumination. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2011, p. 1. ACM, New York (2011)
3. Community Core Vision, `http://ccv.nuigroup.com/`
4. Gelineck, S., Büchert, M., Andersen, J.: Towards a more flexible and creative music mixing interface. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2013, pp. 733–738. ACM, New York (2013)
5. 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. ACM, New York (2005)
6. Hancock, M., Hilliges, O., Collins, C., Baur, D., Carpendale, S.: Exploring tangible and direct touch interfaces for manipulating 2d and 3d information on a digital table. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2009, pp. 77–84. ACM, New York (2009)
7. Hennecke, F., Berwein, F., Butz, A.: Optical pressure sensing for tangible user interfaces. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2011, pp. 45–48. ACM, New York (2011)
8. Jetter, H.-C., Gerken, J., Zöllner, M., Reiterer, H., Milic-Frayling, N.: Materializing the query with facet-streams: A hybrid surface for collaborative search on tabletops. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2011, pp. 3013–3022. ACM, New York (2011)
9. Kaltenbrunner, M., Bencina, R.: Reactivision: A computer-vision framework for table-based tangible interaction. In: Proceedings of the 1st International Conference on Tangible and Embedded Interaction, TEI 2007, pp. 69–74. ACM, New York (2007)
10. Kaltenbrunner, M., Bovermann, T., Bencina, R., Costanza, E.: Tuio: A protocol for tabletop tangible user interfaces. In: Proceedings of the 2nd Interactive Sonification Workshop (2005)
11. Lucchi, A., Jermann, P., Zufferey, G., Dillenbourg, P.: An empirical evaluation of touch and tangible interfaces for tabletop displays. In: Proceedings of the Fourth International Conference on Tangible, Embedded, and Embodied Interaction, TEI 2010, pp. 177–184. ACM, New York (2010)
12. Marco, J., Cerezo, E., Baldassarri, S.: Toyvision: A toolkit for prototyping tabletop tangible games. In: Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS 2012, pp. 71–80. ACM, New York (2012)
13. Olson, I.C., Atrash Leong, Z., Wilensky, U., Horn, M.S.: It's just a toolbar!: Using tangibles to help children manage conflict around a multi-touch tabletop. In: Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction, TEI 2011, pp. 29–36. ACM, New York (2011)

14. Olwal, A., Wilson, A.D.: Surfacefusion: Unobtrusive tracking of everyday objects in tangible user interfaces. In: Proceedings of Graphics Interface 2008, GI 2008, pp. 235–242. Canadian Information Processing Society, Toronto (2008)
15. Piovesana, M., Chen, Y.-J., Yu, N.-H., Wu, H.-T., Chan, L.-W., Hung, Y.-P.: Multidisplay map touring with tangible widget. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 679–682. ACM, New York (2010)
16. Schning, J., Hook, J., Bartindale, T., Schmidt, D., Olivier, P., Echtler, F., Motamedi, N., Brandl, P., von Zadow, U.: Building interactive multi-touch surfaces. In: Müller-Tomfelde, C. (ed.) Tabletops. Human-Computer Interaction Series, pp. 27–49. Springer (2010)
17. Terrenghi, L., Kirk, D., Richter, H., Krämer, S., Hilliges, O., Butz, A.: Physical handles at the interactive surface: Exploring tangibility and its benefits. In: Proceedings of the Working Conference on Advanced Visual Interfaces, AVI 2008, pp. 138–145. ACM, New York (2008)
18. Tuddenham, P., Kirk, D., Izadi, S.: Graspables revisited: Multi-touch vs. tangible input for tabletop displays in acquisition and manipulation tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 2223–2232. ACM, New York (2010)
19. TUIO, `http://www.tuio.org`
20. 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, pp. 481–490. ACM, New York (2009)
21. Welzl, E.: Smallest enclosing disks (balls and ellipsoids). In: Maurer, H. (ed.) New Results and New Trends in Computer Science. LNCS, vol. 555, pp. 359–370. Springer, Heidelberg (1991)
22. Wu, A., Reilly, D., Tang, A.,, M.: Tangible navigation and object manipulation in virtual environments. In: Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction, TEI 2011, pp. 37–44. ACM, New York (2011)

Detecting Address Estimation Errors from Users' Reactions in Multi-user Agent Conversation

Ryo Hotta, Hung-Hsuan Huang*, Shochi Otogi, and Kyoji Kawagoe

Graduate School of Information Science & Engineering,
Ritsumeikan University, Japan
hhhuang@acm.org

Abstract. Nowadays, embodied conversational agents are gradually getting deployed in real-world applications like the guides in museums or exhibitions. In these applications, it is necessary for the agent to identify the addressee of each user utterance to deliberate appropriate responses in interacting with visitor groups. However, as long as the addressee identification mechanism is not completely correct, the agent makes error in its responses. Once there is an error, the agent's hypothesis collapses and the following decision-making path may go to a totally different direction. We are working on developing the mechanism to detect the error from the users' reactions and the mechanism to recover the error. This paper presents the first step, a method to detect laughing, surprises, and confused facial expressions after the agent's wrong responses. This method is machine learning base with the data (user reactions) collected in a WOZ (Wizard of Oz) experiment and reached an accuracy over 90%.

Keywords: Multi-party conversation, human-agent interaction, Gaze.

1 Introduction

Various kinds of kiosk information systems are used in public places, such as shopping malls, museums, and visitor centers. The typical situation in which such systems are used is that a group of people stand in front of the kiosk and operate it in order to retrieve the information they request while talking with one another. Therefore, in order to implement conversational agents that can serve as an information kiosk in public places, multi-party conversation functionality for simultaneous interaction with multiple users is indispensable. In dyadic dialogues where only two participants are involved, it can be assumed that in most cases, one participant is the addressee when the other one participant is speaking. In multi-party dialogues, however, distinctions must be made among the roles of the participants, such as speaker, addressee, and listeners standing by. When a person is involved in a human-human-agent triadic conversation,

* Corresponding Author.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 245–253, 2014.

he/she may speak to the agent or may talk to the other person (his/her partner). When the person speaks to the agent, the agent needs to respond to that utterance. However, when the person speaks to the partner, the agent should not mistakenly respond to that utterance. Therefore, one of the basic functionalities a conversational agent needs in order to engage in multi-party conversation is the ability to identify the addressee of each user utterance.

Based on this need, this paper presents a work that aims to determine the addressee of user utterances in triadic conversations among two users and an agent. In the literature, [1–3], it has been reported that in addition to their explicit verbal utterances, humans use nonverbal signals such as their gaze, nods, and postures to regulate their conversation, e.g., to show their intention to yield or willingness to take turns in speaking. If there are specific patterns in the user's gaze behavior that depend upon whom he/she is talking to, it would be possible for the kiosk agent to automatically identify the addressee. Thus, as regards eye-gaze approximation, this study will exploit head direction information obtained from a face-tracking system. It has also been found that in human multi-party conversations and human-robot communication, not just visual cues such as eye gaze and head direction are useful in predicting the addressee, but prosodic cues of the voice as well [4, 5].

In previous stages of this project, we have developed a fully autonomous kiosk agent who can engage with two users at the same time by utilizing non-verbal information only [6–8]. The accuracy of the addressee estimation component was 80%, that means the estimation mechanism has a 20% error rate. Even a human can make such a mistake in multi-party conversation, it can not be expected that the error rate can be reduced to 0%. If the agent decides its actions in responding to the users according to an assumption that all of its perceptions are correct, the conversation afterward will crash and proceed to an unexpected path in the state transition model. Therefore, for further improvement on the system, the mechanisms for detecting and recovering the errors are required. This paper presents the analysis results of users' facial expressions after a wrong estimation of addressee by the agent.

2 Related Work

Research on human communication showed that the eye gaze is an important communication signal in face-to-face conversations. The speaker looks at the addressee to monitor her/his understanding or attitude, and, the addressee looks at the speaker in order to be able to offer positive feedback in return [1, 9]. Eye gaze also plays an important role in turn taking. When yielding his/her turns to speak, the speaker looks at the next speaker at the end of his/her utterances [2]. Vertegaal [10] reported that the gaze is a reliable predictor of addressee-hood. Likewise, Takemae [11] provided evidence that the speaker's gaze indicates addressee-hood and plays a regulatory role in turn management.

Similar results were found in mixed human-human and human-computer conversations in [4, 12, 13]. In these works, perception experiments were conducted in which subjects guessed who the addressee of a given utterance was. It was found that prosodic and visual cues were about equally effective, and that the combination of auditory and visual cues resulted in better performance. Moreover, the motivation of [5] was quite similar to that of this study, in that researchers proposed a method of identifying the addressee in a human-human-robot interaction by combining prosodic and visual cues. As a visual cue, they used the horizontal head orientation to distinguish addressees. They reported that in 35% of the cases, a person talked to the other human while looking at the robot. They then addressed the fact that visual cues alone might not be sufficient, and proposed the further incorporation of prosodic cues. As prosodic cues, they identified a number of linguistic features obtained from the speech recognition system, such as sentence length, typical phrases, and language models, and used them to distinguish the addressee from the other human. They reported that the speech addressed to the robot was detected with an F-measure of 0.72.

In this study, we share a similar motive but tackle the problem using a different approach. First, to estimate the head direction, we add more parameters, namely the position and rotation of the head. We also focus on shifts in head direction during an utterance. As for the prosodic cues, while [5]focused on linguistic features, we assume that the user's tone of voice may be different depending on whether he/she is speaking to the agent or to the partner. Thus, to measure the user's tone of voice, this study focuses on pitch, intensity (volume), and the rate of speech as the most important prosodic features [14]. It has already been found that prosodic features are useful in the recognition of emotion, and can thus be expected to be useful in characterizing the tone of voice. Considering all these aspects discussed in previous studies, this study employs machine learning techniques to create an automatic classifier for estimating the addressee via the integration of visual and prosodic information.

However, in our current prototype, after the decision making component receive the output of the addressee estimation component, it can not confirm whether the estimation is correct or not. It can only assume that the estimation is correct and go proceed the decision making. We then go forward to the next step of addressee estimation to deal with the situation when there was an error in this estimation. The idea is to get the reactions (feedbacks) from the users in a short period after the agent takes its action in responding to a user utterance. Follow the basic ideas of addressee estimation, we focus in using nonverbal feedbacks from the user, facial expressions and prosodic information.

3 Corpus Collecting Experiment

Regarding to addressee estimation, the possible errors are defined as the following situations:

Unexpected Response (UR): the addressee of a user utterance should be another user, but the agent mistakenly responded.

No Response (NR): the addressee should be the agent, but the agent did not response.

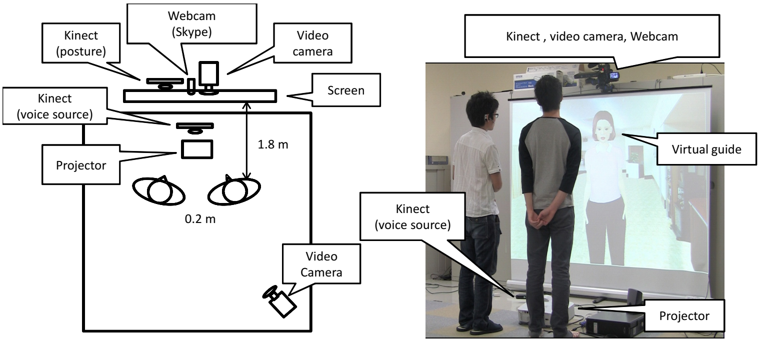


Fig. 1. Setup of the WOZ experiment to collect data for the interaction corpus

In order to collect users' reactions on wrong addressee estimations, a WOZ experiment on three collaborative decision making tasks was conducted. We expect that the subjects' reactions toward the agent may differ to how they talk with a human information provider. To observe the natural interaction with humans and agents, we chose the WOZ experiment setting instead of a human-human one.

Pairs of experiment participants were instructed to interact with a life-size female character on a screen. They had to retrieve information from the character in order to make a decision regarding the given tasks until the agreement between them achieved. As shown in Figure 1, the subjects stood about 1.8 m away from the screen where the character was projected. Two video cameras were used to record the whole experiment, one from the front to take the upper bodies of the participants and the other one takes the whole scene including the participants and the character from the rear. One Webcam was used for the telecommunication software Skype to connect the WOZ operator to the experiment room. The microphone array of one Microsoft Kinect sensor was to identify the voice source (left or right user) of user utterances. The other Kinect sensor was used to record body postures with depth images. The conversation experiment was conducted with the following premises:

- The participants want to make a decision base on their agreement from multiple candidates with the help of the agent who is knowledgeable about that task domain.
- The participants have a rough image of what they want, but they do not have idea about particular candidates in advance.
- The participants discuss on their own and acquire new information from the agent.
- The conversation ends when the participants made the final decision.

A total of 15 pairs of college students were recruited as the participants in the experiment, all of whom were native Japanese speakers. The students came from various departments ranging from economics, life science to engineering at average age, 19.2. 11 of the all 15 pairs were male ones and the other four were female ones. Each pair was instructed to complete three decision-making tasks: travel planing, lecture registration, and part-time job finding.

Travel Planning: the participants were instructed to pretend to be in a situation where they had a coupon from a travel agency that allows them to visit three of 14 sightseeing spots in Kyushu for free. The information, which includes a brief history, highlights, nearby restaurants, for each location was defined in advance. The sightseeing spots were selected from four of all seven prefectures inside Kyushu. The participants were instructed to complete their task by freely retrieving information from the travel agent, and to discuss their decisions on their own.

Lecture Registration: the participants were instructed to choose three out of 12 lectures to attend together in the next semester. The information about the lecturer, textbook, course difficulty, prerequisites, etc. for each lecture was defined in advance. The lectures were divided into four categories: information science, engineering, languages and communication, and social science. The subjects could freely ask the "tutor" agent for any information about the lectures or the agent itself, and then discuss this on their own in order to make the final decision.

Part-Time Job Hunting: the participants were instructed to request help in choosing three out of 14 part-time jobs to work together near the university. The information about the salary, location, workload, work type, etc. for each job was defined in advance. The jobs were divided into four categories: convenient stores, book shops, restaurants, and gas stations. The subjects could freely ask the agent for any information about the part-time jobs or the agent itself, and then discuss this on their own in order to make the final decision.

These tasks were chosen because the student participants are supposed to be familiar with these issues. In order to stimulate more active discussion, the participants were instructed to make rankings on the three final choices. All participant pairs were assigned to take all of the three tasks in three separate sessions, one task for one session. The sessions were conducted in all possible orders to cancel order effects. One student who major in computer science was recruited to operate the WOZ agent. He was chosen due to his familiarity with operating a GUI-based WOZ application, which ensured that there would be smooth interaction. The operator was asked to practice on the WOZ user interface for two hours prior to the experiment to further ensure that the agent's response time was quick enough. All the sentences that the agent could speak during the experiment were listed in a menu where relevant sentences were grouped for the WOZ operator to select from more easily. There was also a text field that allowed the operator to type arbitrary utterances, in the cases when they were needed but were not defined. The WOZ operator was instructed to try to end the interaction sessions in ten minutes, if possible.

The order of the task for each session was changed to achieve counter-balance, but the wrong responses were intentionally inputed according to the following rules:

- Errors are inputed around every three minutes
- Intentionally make mistakes in the situation when the agent was able to response in the interaction so for
- In the UR situations, the WOZ operator responded as keyword matching manner in simulating the autonomous agent

The wrong responses were only inputed in the third session to allow the user to get used to the conversation with a CG agent, to have time to approximate the agent's ability (100% accurate in the first two sessions), and to notice the errors more easily.

4 Features for Detecting Addressee Estimation Errors

The assumption of the error detection is, users should have some emotional reactions to the agent's errors. For example, the two cases: the users may feel surprised or funny if the agent responded to an utterance that is should not do; the users may feel confused if the agent should answer a user question but it did not, can be considered. The preliminary analysis on four groups was focused on the facial expressions of the users' reactions: laughed, confused, surprised comparing to neutral. Table 2 shows the results of facial expression annotation. The results showed that the users had high possibility to change their facial expressions after an error within five seconds (29 times among 40 error instances). Table 1 shows the relationship between each facial expression and the errors.

Table 1. Relationship between each facial expression and intentionally triggered errors in the experiment

	Neutral	Laughed	Confused	Surprised
Agent did not respond when user speak to it	8	7	13	0
Agent mistakenly responded to an utterance issued to the other user	3	5	1	3
Facial expression changes when there is no error	214	141	12	31
Percentage when there is an error	4.8%	7.8%	53.8%	8.8%

From the annotation process, we had the following three findings:

- The facial expression, confused appears at higher frequency than the other expressions when there are errors

- When one of the users showed surprised facial expression and the last speaker is the other user, there is relatively high possibility that the agent's response was not an error
- When one of the users showed confused facial expression and the last speaker is the agent, there is relatively high possibility that the agent's response was an error

Table 2. Summary of label instances of each subject

Expressions	Max.	Min.	Avg.	Std. Dev.
Neutral	35	9	21.62	7.06
Laughed	24	7	15.37	4.87
Confused	5	1	2.50	1.58
Surprised	7	0	2.75	2.04

We then used hand labeled data and FACS Action Units [15] recognized by visage|SDK [16] as the feature set (Table 3) to train a random forest with Weka [17]. The 10-fold cross-validation accuracy was over 90%. The results of the classification for each facial expression base on the proposed feature set is shown in Table 4. The detection of related facial expression itself is possible, but it is difficult to detect the agent's error merely by facial expressions. This is due to the fact that the users have frequent facial expression changes even when there is no error. The facial expression detection itself works well and should contribute to the detection of errors. However, since the users also make these facial expressions when there is no error, errors can not be detected merely by facial expressions, other features are required.

Table 3. Facial and head movement features used in classifying the facial expressions

Nose wrinkler (AU9)	Lip corner depressor (AU13/15)	Rotate eyes down (AU64)
Jaw drop (AU26)	Outer brow raiser (AU2)	Position distance
Lower lip drop (AU16)	Inner brows raiser (AU1)	Rotation distance
Upper lip raiser (AU10)	Brow lowerer (AU4)	Rotation (angle)
Lip stretcher (AU20)	Rotate eyes (AU61/62)	

5 Conclusion and Future Work

It is necessary for the agent to identify the addressee of each user utterance to deliberate appropriate responses in interacting with multiple users. However,

Table 4. Classification results of each facial expression

Expressions	Precision	Recall	F value
Neutral	0.866	0.866	0.866
Laughed	0.887	0.890	0.888
Confused	0.962	0.960	0.961
Surprised	0.947	0.944	0.945
10-fold cross validation	90.90%		

the agent can not confirm whether the estimation is correct or not. This paper presents a work on the mechanism to detect the error from the users' reactions. The first step, a method to detect laughing, surprises, and confused facial expressions after the agent's wrong responses. This method is machine learning base with the data (user reactions) collected in a WOZ (Wizard of Oz) experiment and reached an accuracy over 90%. From the analysis results, errors can not be detected merely by facial expressions, other features are required.

As future works, we are analyzing the situations when the addressee estimation component is prone to make errors. Also, we would like to consider other modalities like voice features or postures to improve the accuracy.

References

1. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychologica 26, 22–63 (1967)
2. Duncan, S.: Some signals and rules for taking speaking turns in conversations. Journal of Personality and Psychology 23(2), 283–292 (1972)
3. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language 50(4), 696–735 (1974)
4. Terken, J., Joris, I., Valk, L.D.: Multimodalcues for addressee-hood in triadic communication with a human information retrieval agent. In: Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI 2007 (2007)
5. Katzenmaier, M., Stiefelhagen, R., Schultz, T.: Identifying the addressee in human-human-robot interactions based on head pose and speech. In: Proceedings of the 6th International Conference on Multimodal Interfaces, ICM 2004 (2004)
6. Huang, H.H., Baba, N., Nakano, Y.: Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation from nonverbal information. In: 13th International Conference on Multimodal Interaction (ICMI 2011), pp. 401–408 (2011)
7. Baba, N., Huang, H.H., Nakano, Y.: Addressee identification for human-human-agent multiparty conversations in different proxemics. In: 14th International Conference on Multimodal Interaction (ICMI 2012), 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality (2012)
8. Nakano, Y., Baba, N., Huang, H.H., Hayashi, Y.: Implementation and evaluation of multimodal addressee identification mechanism for multiparty conversation systems. In: 15th International Conference on Multimodal Interaction, ICMI 2013 (2013)

9. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press (1976)
10. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 301–308 (2001)
11. Takemae, Y., Otsuka, K., Mukawa, N.: Video cut editing rule based on participants' gaze in multiparty conversation. In: 11th ACM International Conference on Multimedia (2003)
12. Lunsford, R., Oviatt, S.: Human perception of intended addressee during computer-assisted meetings. In: Proceedings of the 8th International Conference on Multimodal Interfaces, ICMI 2006, pp. 20–27. ACM, New York (2006)
13. Dowding, J., Alena, R., Clancey, W.J., Sierhuis, M., Graham, J.: Are you talking to me? dialogue systems supporting mixed teams of humans and robots. In: AAAI Fall Symposium (2006)
14. Rodriguez, H., Beck, D., Lind, D., Lok, B.: Audio analysis of human/virtual-human interaction. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 154–161. Springer, Heidelberg (2008)
15. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system (facs). Website (2002)
16. Visage Technologies AB: Visage|SDK. Website (2008), `http://www.visagetechnologies.com`
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. ACM SIGKDD Explorations 11(1), 11–18 (2009)

Evaluation of Leap Motion Controller with a High Precision Optical Tracking System

Grega Jakus, Jože Guna, Sašo Tomažič, and Jaka Sodnik

University of Ljubljana, Ljubljana, Slovenia
{grega.jakus,joze.guna,saso.tomazic,jaka.sodnik}@fe.uni-lj.si

Abstract. The paper presents an evaluation of the performance of a Leap Motion Controller. A professional optical tracking system was used as a reference system. 37 stationary points were tracked in 3D space in order to evaluate the consistency and accuracy of the Controller's measurements. The standard deviation of these measurements varied from 8.1 μm to 490 μm, mainly depending on the azimuth and distance from the Controller. In the second part of the experiment, a constant distance was provided between two points, which were then moved and tracked within the entire sensory space. The deviation of the measured distance changed significantly with the height above the Controller. The sampling frequency also proved to be very non-uniform. The Controller represents a revolution in the field of gesture-based human-computer interaction; however, it is currently unsuitable as a replacement for professional motion tracking systems.

Keywords: Leap Motion Controller, motion capture system, consistency, accuracy

1 Introduction

Gesture-based user interfaces in combination with the latest technical advances, incorporating accurate and at the same time affordable new types of input devices, offer new opportunities for specific application areas such as entertainment, learning, health and engineering [1]. One of the latest technological breakthroughs in gesture sensing devices is Leap Motion Controller [2]. The device, approximately the size of a matchbox, allows for precise and fluid tracking of multiple hands, fingers or small objects in free space with sub-millimeter accuracy. With its enhanced interaction possibilities, the Controller could trigger a new generation of much more useful 3D displays and possibly complement the mouse as a secondary input device [3].

An interesting study of the Controller in [4] shows its potential for gesture and handwriting recognition applications. The acquired input data are treated as a time series of 3D positions and processed using the Dynamic Time Warping algorithm. The authors report promising recognition accuracy and performance results.

Professional optical motion capture systems are also often used in the domain of human-computer interaction for motion detection and gesture recognition. An example of their application is presented in [5], where the authors focused on an adaptive

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 254–263, 2014.

gesture recognition system while developing a gesture database to eliminate the individual factors that affect the efficiency of the recognition system. In particular, hand gestures were investigated.

The motivation for the research presented in this paper is to analyze the accuracy and consistency of the Controller with the aid of a professional optical motion capture system Qualisys [6]. The major goal of the study was to determine the Controller's suitability for a possible replacement of a professional motion tracking system. We were primarily interested in the following aspects:

- the consistency of the measurements at fixed spatial positions (spatial dispersion of measurements through time),
- the accuracy of the measured position in relation with its spatial position (the spatial dependency of accuracy), and
- the uniformity of data sampling.

The rest of the paper is organized as follows: Following the introduction, technical setup and measurement methodology are presented in Chapter 2. The detailed results are presented and analyzed in Chapter 3. Finally, key conclusions are discussed in Chapter 4.

2 Experimental Design

2.1 Technical Setup

Due to patent and trade secret restrictions, very few details are known about the Leap Motion Controller's inner structure and its basic operational properties. It is, however, clear that infrared imaging is used for object tracking.

According to the official specification [2], the Controller's sensory space is in the shape of an inverted pyramid positioned at the central point of the device. The Cartesian and spherical coordinate systems used to describe the tracked positions are shown in Fig.1. Cartesian coordinate system consists of the x axis running along the longer Controller's side (the length). Height is measured above the Controller and described as the y axis. Finally, the depth runs along the shorter Controller's side and is described as the z axis.

Our pre-experiment trials determined the Controller's useful sensory space, which is approximately within the boundaries of -250 mm < x < 250 mm, -250 mm < z < 250 mm and 0 mm < y < 400 mm.

The spherical coordinates consist of radius (r), azimuth (φ) and elevation (θ). In our case, the azimuth angle is defined under the assumption of the symmetry in the Controller's performance over x and z axes. The angle is therefore measured from the x axis (and not from the z axis as it is a common practice) to the line connecting the coordinate origin and the projection of the measured location in the x-z plane.

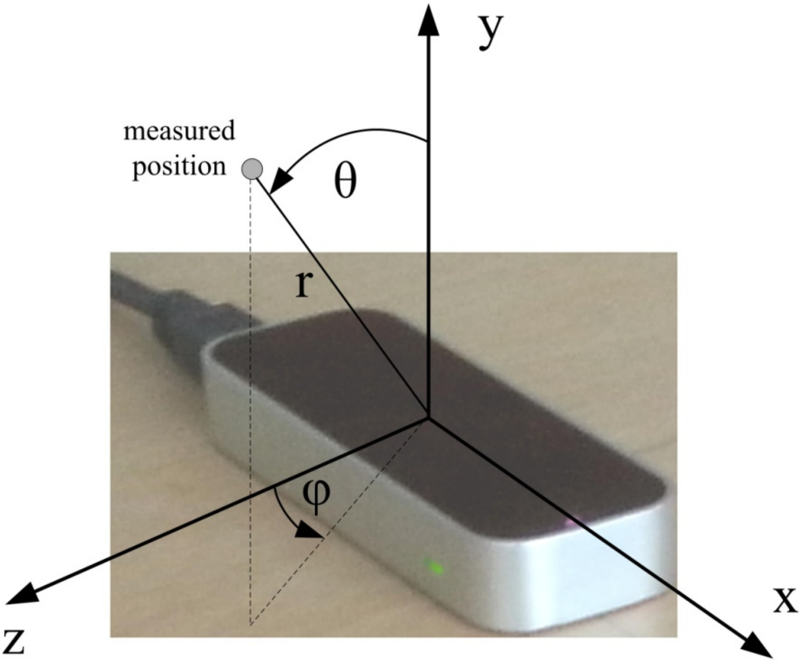


Fig. 1. The Cartesian and spherical coordinate systems used to describe positions in the Controller's sensory space

Programmatically, the Controller can be accessed through various Application Programming Interfaces. For the purpose of this study, a special real-time data acquisition and logging software has been developed in Python programming language. Each measurement of the Controller was logged with the corresponding timestamp. The latter enabled us to determine the exact time gap between two sequential samples and to calculate the corresponding sample frequency. Data processing and analysis was performed in Matlab.

A high-precision optical tracking system Qualisys [5] was used as a reference system. It consisted of eight Oqus 3+ high-speed cameras and the Qualisys Track Manager software. Such systems are widely used for fast and precise tracking of various objects in industrial applications, biomechanics, and media and entertainment applications. The tracking precision depends on the number of cameras used, their spatial layout, the calibration process, and the lighting conditions.

2.2 Methodology

The Controller's performance was evaluated through two types of measurement scenarios. In the first scenario, 37 stationary points in space were tracked for a longer period of time in order to evaluate the consistency and uniformity of the measurements. The locations of the stationary points were chosen systematically through the entire Controller's sensory space.

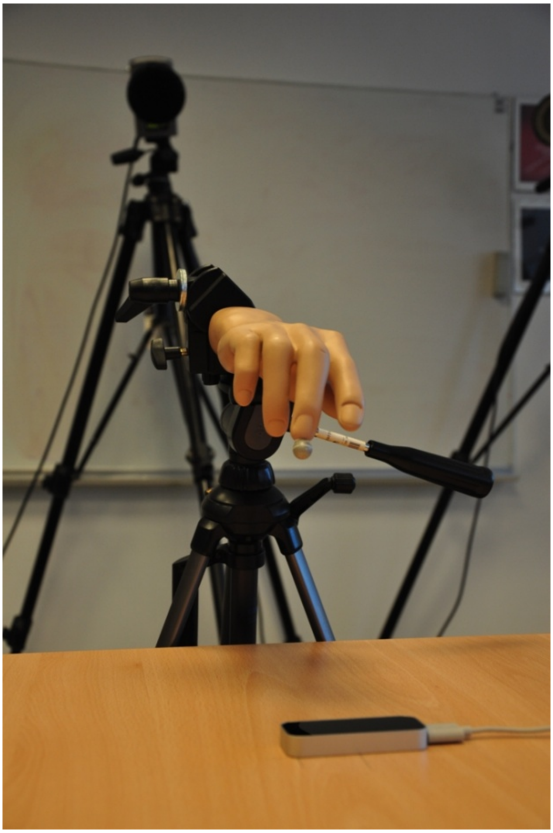


Fig. 2. Experimental environment showing the static measurement scenario setup. One of the cameras of the motion tracking system is visible in the background.

The tracked object was a passive reflective marker, attached to the middle finger of a plastic hand model. The marker was tracked by the Controller and by the reference optical motion tracking system simultaneously. The experimental setup is shown in Fig. 2.

In the second measurement scenario, a constant distance was provided between two points (markers) with the aid of a special V-shaped tool. It consisted of two markers and a supporting rigid structure ensuring constant distance (21.36 mm) between the markers. The exact distance was acquired with the reference system. The tracking accuracy of the Controller was evaluated based on the deviation of the distance between the two markers after moving the V-tool freely around the sensory space. This experimental setup is illustrated in Fig. 3.

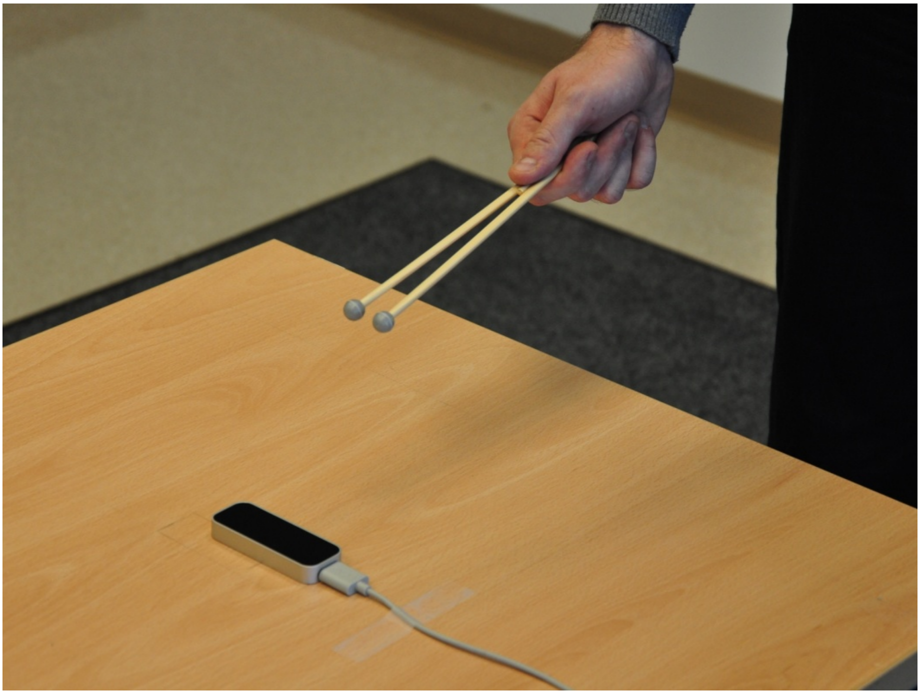


Fig. 3. Experimental environment showing the dynamic measurement scenario setup

3 Results and Interpretation

3.1 Static Measurements

In the static setup, standard deviations calculated for all 37 points varied from 8.1 µm to 490 µm. Generally speaking, the lowest standard deviations were measured in space directly above the Controller, while the highest standard deviations were measured at the leftmost and the rightmost positions.

Fig. 4 shows the probability density of deviation of the measured location from the actual location for the individual axes. The results indicate that, when making a single measurement, a smaller deviation of the location is expected along the x axis (alongside the Controller) compared to the directions away from the Controller (along the z axis) or in height (along the y axis).

Further analysis of the linear correlation revealed that standard deviation increases significantly with the distance from the Controller (r = 0.34, p = 0.044) and azimuth angle (leftmost and rightmost sides of the Controller, r = 0.43, p = 0.051). No significant correlation was found when changing inclination of the tracking objects.

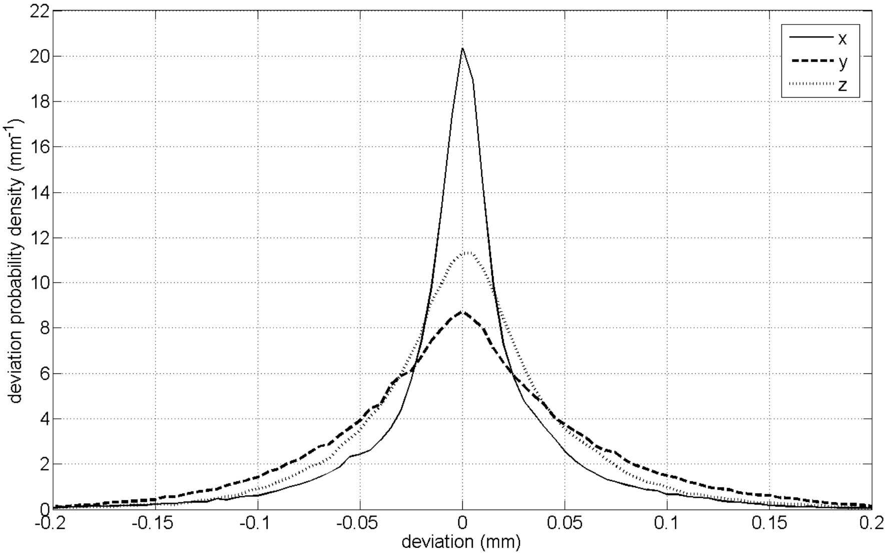


Fig. 4. The deviation probability density for individual axes including all 37 locations

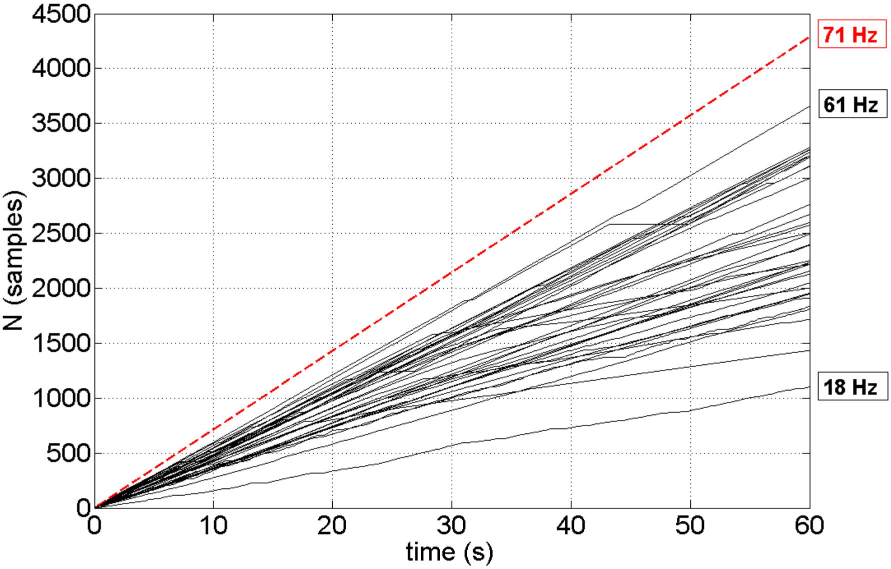


Fig. 5. The sampling performance of the Controller in the static setup

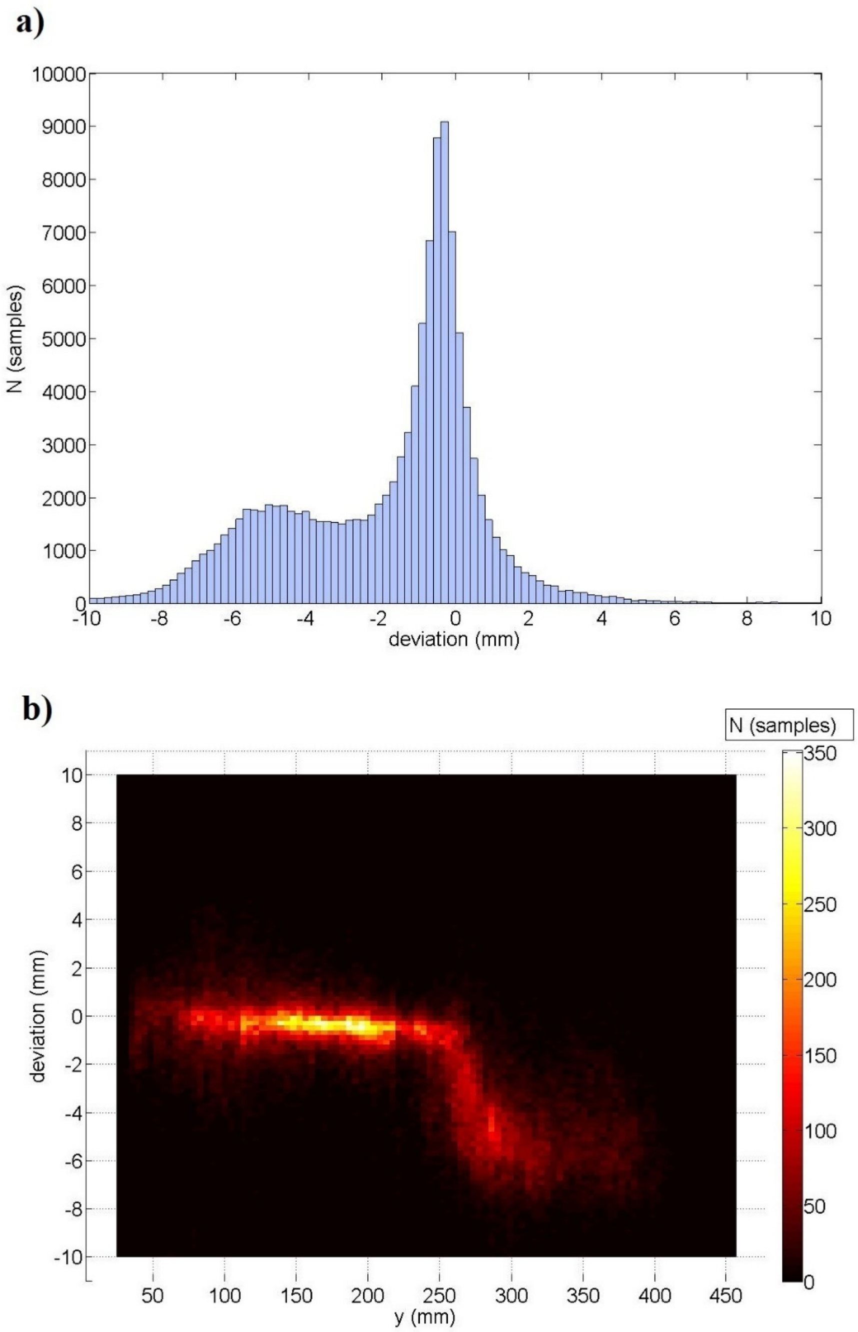


Fig. 6. Distribution of distance deviation between two points: the overall distribution (a) and the distribution on the y axis (b)

Our further investigation was focused on the Controller's sampling performance. Fig. 5 demonstrates the progress of data acquisition in the first minute of measurements at each of the 37 measured positions. The figure indicates very non-uniform sampling as the sampling frequency varies both in space and time. The minimal logged period between two samples was 14 ms, which corresponds to the reference sampling frequency of 71 Hz (indicated with the dashed line in Fig. 5). The actual average sampling frequencies were calculated based on the number of samples acquired in the first minute and ranged between 18 and 61 Hz. The average sampling frequency across all locations was 39 Hz. The standard deviation was 13 Hz.

3.2 Dynamic Measurements

In the dynamic setup, two markers with constant inter-distance were tracked while moved freely in the sensory space. Fig 6a shows the distribution of deviation of the distance between the two tracked markers. The distribution consists of two local peaks, one at 0 mm and one at approximately -5 mm. The latter corresponds to the points located more than 250 mm above the Controller, which can be noted from Fig 6b.

The spatial dependency of the distance deviation was determined by computing linear correlations between the deviation and individual spatial dimensions. The analysis revealed statistically significant moderate negative linear correlations between the distance deviation and the height above the Controller (r = 0.61, p < 0.000) and the distance from the Controller (r = 0.60, p < 0.000). The distance deviation was not correlated with other spatial dimensions.

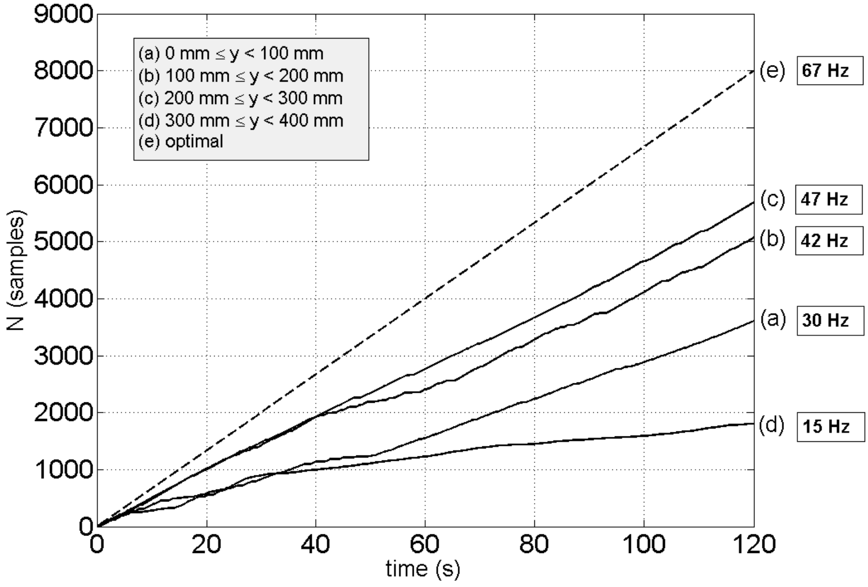


Fig. 7. The sampling performance of the Controller in dynamic scenario

The sampling frequency varied significantly in the dynamic setup as well. Fig. 7 displays the Controller's sampling performance when tracking moving objects in four separate layers in height. The actual sampling performance is compared against "optimal" sampling performance (indicated by a dashed line). The latter corresponds to the constant sampling period of 15 ms, the minimum time interval between two consecutive samples logged in the dynamic measurements.

The figure indicates the best sampling performance between the heights of y = 100 mm and y = 300 mm, while it gets significantly less efficient below and particularly above this height.

4 Discussion and Conclusions

The Leap Motion Controller proved to be very accurate for tracking static points in a predefined sensory space, but less accurate when objects move around. In both cases, the measurement consistency and accuracy varied significantly at different spatial positions. For example, in the static scenario, objects placed directly above the Controller were tracked with the highest consistency (lowest standard deviation) but with a much lower consistency (highest standard deviation calculated) at the leftmost and rightmost positions. In the dynamic scenario, the distribution of deviation of the distance between the two markers increased significantly when tracking higher than 250 mm above the Controller.

Our experiment clearly demonstrated that the Controller's consistency and accuracy are spatially dependent. This fact limits the Controller's suitability for precise tracking of various objects in space. The additional limitations are also a relatively modest sensory space (only approximately one tenth of a cubic meter) and a varying sampling frequency for both static and dynamic measurements. These drawbacks, however, do not influence the Controller's primary purpose which is to be used as an alternative interaction device.

To conclude, the Leap Motion Controller undoubtedly represents a revolution in the field of gesture-based human-computer interaction, but it is currently unsuitable as a replacement of professional motion tracking systems with a sufficient accuracy and sampling uniformity. Based on the insights gained from the study, our future research involving Leap Motion Controller will be focused on using the device in practical applications such as gesture-based interfaces.

References

1. Bhuiyan, M., Picking, R.: Gesture-controlled user interfaces, what have we done and what's next? In: Proceedings of the Fifth Collaborative Research Symposium on Security, E-Learning, Internet and Networking (SEIN 2009), Darmstadt, Germany, pp. 26–27 (2009)
2. Leap Motion Controller, `https://www.leapmotion.com` (accessed on October 29, 2013)
3. Hodson, H.L.: Motion hacks show potential of new gesture tech. New Scientist 218(2911), 21 (2013)

4. Vikram, S.: Handwriting and Gestures in the Air, Recognizing on the Fly. CHI 2013 Extended Abstracts (2013)
5. Aziz, A., Khairunizam, W., Zaaba, S.K., Shahriman, A.B., Adnan, N.H., Nor, R.M., Ramly, M.F.: Development of a Gesture Database for an Adaptive Gesture Recognition System. International Journal of Electrical & Computer Sciences IJECS-IJENS 12(4) (2012)
6. Qualisys, `http://www.qualisys.com` (accessed on October 29, 2013)

Proposal of a Method to Measure Difficulty Level of Programming Code with Eye-Tracking

Tomoko Kashima[1], Shimpei Matsumoto[2], and Shuichi Yamagishi[2]

[1] Faculty of Engineering, Kinki University,
1 Takaya Umenobe, Higashi-Hiroshima City, Hiroshima, 739-2116, Japan
[2] Faculty of Applied Information Science, Hiroshima Institute of Technology,
2-1-1 Miyake, Saeki-ku, Hiroshima 731-5193, Japan

Abstract. In recent years, guaranteeing the educational quality is required in university education of Japan. With this situation in mind, we built study support environment with the information technology. As a result, we utilized the result for programming education and obtained the effect. There are various technical elements in the programming skill. However, many evaluations have adopted a comprehensive evaluation method. Therefore, a student's attainment to each technical element is indefinite. Some students become difficult to perform learning activities. So, in this research, programming notes the point which is the implicit thinking skill which is strongly related in study. Accumulation experience analyzes strongly related eye movement, and we aim at the standard construction for skill.

Keywords: programming, difficulty level, educational support, eye-tracking.

1 Introduction

Recently, university education of Japan is demanded for the educational quality. Based on this situation, authors developed study training supporting environment using Information technology. Such learning environment has been utilized for the education of computer programming. We have some efficacy in those studies. However, there are various technical elements in programming skill. By such programming, the present condition is that a learner's evaluation is estimated by only overall points. Therefore, a learner is difficult to get to know the attainment level to each own technical element. So, in this study, we propose the method of presuming the difficulty of programming which applied the eye-tracking system [1] which is biological information.

Most investigations are classified with learner's analysis and discovery of the feature [2], teaching method proposal of programming instruction [3] and development of programming instruction support software [4]. However, almost all study target a beginner. Most investigations are classified with learner's analysis and discovery of the feature, teaching method proposal of programming instruction and development of programming instruction support software. Almost all study targets a beginner. Usually, a beginner's study training speed differs.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 264–272, 2014.

Therefore, it is necessary to change the study training supporting method and the teaching method according to the level of each learning. It is important to provide the various study training methods according to a learner's level.

The research task is divided into three phases in this study.First, we create a learner's evaluation index. Programming technique is subdivided and the skill standard for checking the learning situation of the technology assessment covering the many dimensions is developed. Skill items and those degrees of difficulty are built referred to item reaction theory, and existing previous study and books, and create a skill judging examination. Next, eye-tracking measurement experiment and its analysis are conducted. Measuring with a skill check is difficult for the process of the thinking according to the degree of experience. A learner's eye-tracking is measured and a learner's pattern of thinking is classified. A goal (predominance set) is set up using a Data Envelopment Analysis (DEA) model [5] by making into an input item the data obtained with the skill check, and the data obtained from biological information, and the teaching method according to a goal is proposed. A learner's skill improvement is supported by repeating these steps. Finally, we measure the learning efficiency according to each learner. And a learner's skill level and the carrer path which responded properly are shown. Here, the creation method of a concrete learner's evaluation index is explained. The technical element of programming is clarified and development of the learning materials for evaluation of the skill level of each element and a skill check table are developed. In order to learn the programming skill of a computer, a learner needs the ability of programming of not only the grammar of a programming language but an algorithm, or others. We clarify all skill elements.

Next, we develop the learning materials for checking the skill check about a learner's programming skill item. Learning materials are based on the structure which we developed until now. The structure takes up the module of the minimum unit. For example, the question of code complement form with which a starved area is compensated in an input, an output, and processing is used. It is the structure which can educate the thinking power according to the learner's technical element by the question. It feeds back to a learner by teaching and testing using the above structure. In the process in which this cycle is repeated, learner degree of comprehension and the degree of difficulty are quantitatively computed by item reaction theory. Based on a result, a learner group is defined according to a learner's achievement. Learning materials and a judgment examination are completed. A skill judging test applies the lesson module of Moodle, and collects learning histories efficiently.

It is easy to make a judgment mistaken for a correct answer in the proposed learning materials. However, even if a learner makes the same mistake, it is possible that the levels of a learner's understanding differ. Then, acquisition of programming skill pays attention to the point being strongly related in the logical thinking based on study training experience. This accumulation experience pays attention to the eye movement expressed strongly. By analyzing a learner's eyes, the process of consideration is clarified until it obtains the answer to each question. The degree of comprehension to experience and the component engineering which cannot be measured in a test is clarified by measuring a learner's eyes and classifying a learner's pattern of thinking. In this study, the analysis result of the eyes data especially done to various learners is introduced. The validity of the eyes data which is biological information as a method of measuring the degree of comprehension of program technology by that cause is described.

2 Programming Education

In the institution of higher education relevant to an information technology, programming skill is especially positioned as an important subject. Since the skill standard of programming is not fully defined, the contents of instruction differ for every organization. An attainment target, technical elements, and those setting levels are dependent on a teacher's educational philosophy in many cases. Student's results are decided by one-dimensional evaluation. In this case, mathematical logic thinking power is required strongly in many cases. The learner who makes this mathematical thinking power elated is not dependent on the teaching method or the contents, and good evaluation is obtained. On the other hand, other learners are evaluated by programming as a proper layer which is not. Historically, analysis of the achievement indicates that two layers certainly exist universally. This trend is not concerned with the difference of age, sex, and an academic level, but generating equally is known experientially. On many works, the presentation to a data input/output and a user is in the mainstream. It is rare to require advanced logic thinking power. A function, a class design, and the definition of a data structure are different technical elements from logic thinking power. The teacher should also evaluate these points. The improvement in software development power of Japan is indispensable in an international competition. Therefore, the learner whose learning evaluation of programming is not good should not judge that there is nothing properly. A learner is evaluated from many sides and it is thought that it is necessary to show the place of activity.

2.1 Previous Study

The study for programming is divided roughly into three kinds.

1. A learner's analysis and discovery of the feature
2. The teaching method proposal of programming instruction
3. Development of programming instruction support software

No. 1, the observation of an error pattern that a learner falls easily, the item and the coping-with method which bar a learner's understanding and the feature of the learner who makes programming unskillful. These are mostly in agreement in the report of previous study [6]. However, many study has stopped at suggestion and a proposal for the beginner. Those study has proposed neither the study training supporting method for a beginner to advance a study training in maturity according to the level of each learning, nor the teaching method according to the level of a learner's learning.

No. 2, Hofuku et al. has perceived that there are many points which must be learned when studying programming, and has proposed the learning method which arranges a study training step in detail [7]. However, they have not mentioned the point of preparing the directivity of various study trainings according to a learner's attainment level.

No. 3, It is considered as the support in a lecture, and the research tasks with main self-study learning environment construction. The LMS development based on Web by the study for self-study environmental construction is in use. The advantage of the study here is at the point which can feed back the result which may have had the use example of self-teaching environment analyzed to the teaching method [8]-[10]. And, in order to aim at evoking interest, development of the function which paid its attention to the leisurely element, and presentation of an intuitive concept are done in many cases. The example of representation has Squeak, Alice, a pro grameen, and the Argo logic. As for the above, many results of research outstanding for the beginner have been reported. However, the support according to the achievement of the learner of the level which stepped up from the beginner is hardly tried in previous study. In order to be targeted at all the learner, it is necessary to formalize the step of the thinking according to programming experience. The trial into which skill of the technical element of programming introduced eye-tracking apparatus at this point depending on experience of coding for expression of experience is not checked as long as it is our investigation.

3 Proposal

In this study, we roughly divide a research task into the next three items, and are planning it.

3.1 Making of Evaluation Index

Programming technique is subdivided and the skill standard for checking the learning situation of the technology assessment covering the many dimensions is developed. Skill items and those degrees of difficulty are built referring to item reaction theory, and existing previous study and books, and create a skill judging test.

Specifically, a learner's evaluation index is created first. The technical element of programming is clarified and development of the learning materials for evaluation of the skill level of each element and a skill check table are developed. As shown in Table 1, the work which pulls out the item in connection with the ability of programming of not only the grammar of a programming language but an algorithm or others is done. Next, the learning materials for doing the skill check about a learner's programming skill item are developed. Learning materials are based on the structure developed by achievements [11]. The module of the minimum unit is taken up, and it is a question of code complement form with which a starved area is compensated in an input, an output, and processing, and can educate the thinking power according to a technical element. It feeds back to a learner by teaching and testing using the above. In the process in which this cycle is repeated, learner degree of comprehension and the degree of difficulty are quantitatively computed by item reaction theory. Based on a result, a learner layer is defined according to a learner's achievement. Learning materials and a judgment test are completed by the above. A skill judging test utilizes the lesson module of Moodle, and collects learning histories efficiently.

3.2 Eyes Measurement and Analysis

About the process of the thinking according to the degree of experience which cannot be measured with a skill check, eyes are measured and a learner's pattern-of-thinking classification is done. A target (predominance set) is set up using a Data Envelopment Analysis (DEA) model by making these into an input item, and the teaching method according to a target is proposed.

It is easy to make a judgment mistaken for a correct answer in the learning materials done in the last fiscal year. However, while it is the same, it is possible that a level differs also in changing. Then, paying attention to a point strongly related in the logical thinking based on study training experience, this accumulation experience pays attention to acquisition of programming skill at the eye movement expressed strongly. By analyzing a learner's eyes, the process of consideration is clarified until it obtains the answer to each question. The degree of comprehension to experience and the component engineering which cannot be measured in a test becomes clear by measuring a learner's eyes and doing a learner's pattern-of-thinking classification. Fig. 1 shows the result of the eyes analysis experiment of the learner who carried out in advance of this study. The users 1 and 2 of Fig. 11 are learners who are insufficient of the skill which guesses the output of the program of Fig. 1.However, it becomes clear that it is an understanding level which is different when eyes are measured and analyzed. Although it turns out that he can understand the grammar used as the key to a program, and a value about the user 1, it is a difficult learner to arrive at an answer, without the ability to understand fine grammar. On the other hand, if it arrives at the contents which are not understood to some extent about the user 2, it read over from the 1st line again and has repeated this operation. Thus, it becomes possible to read in an eye-tracking history the degree of comprehension which is not reflected in a result. The pattern of thinking of the learner who measured and got eyes is utilized for the teaching method making corresponding to the component engineering. It is assumed that the comment in question was the cause by which a learner layer with difficult follow continues existing in the point only depending on the explanation from a teacher's expert viewpoint traditional. Technique changes in the layer from which an experience level is different.

A gap exists explaining skillful work to a beginner. I think that the layer whose experience level of us is shallow will be what the technique of a somewhat

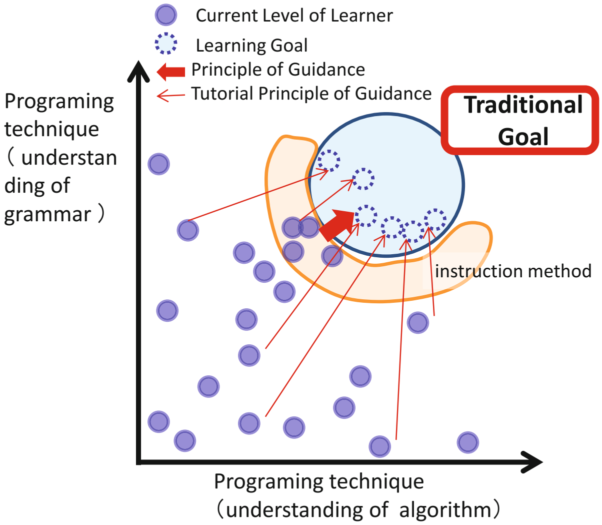

Current Level of Learner
Learning Goal
Principle of Guidance
Tutorial Principle of Guidance
Traditional Goal
Programing technique (understan ding of grammar)
instruction method
Programing technique (understanding of algorithm)

Fig. 1. The existing educational approach

high layer is imitated for (view place of eyes), and it will be important for it to promote step-up. With eyes analysis, the definition of learner layers is made strict and learner layers are simultaneously associated with a path. Grouping of the learner with each technologic-abilities difference is carried out based on the vector of many dimensions. A target (predominance set) is set up for every group, and it aims at proposing the teaching method according to a target. A Data Envelopment Analysis (DEA) model is used. Unlike the former, a setup of many targets is attained by this, and the target according to each learner's skill can be set up Fig. 2. In Fig. 2, three targets are set up and two or more teaching methods also exist. Thus, a setup of the target for developing the ability made elated for a learner is attained.

3.3 Evaluation of Learning Efficiency Nature According to a Larner

Here, a learner evaluates whether the study training is done efficiently. And, a learner's skill level and the carrer path which responded properly are shown. Specifically, each learner continues a study training toward a target in accordance with the teaching method. As a result, learning efficiency is evaluated according to the final place at which the learner arrived. Furthermore, matching with a learning stage and a career is done. As shown in Fig. 3, the check of the future target according to ability of a learner is attained, and he leads to the improvement in greediness for learning. And, in order to attain at a target, the check of the technical element which run short at present is attained, and becomes possible at any time about the directivity and the current position of a study training.

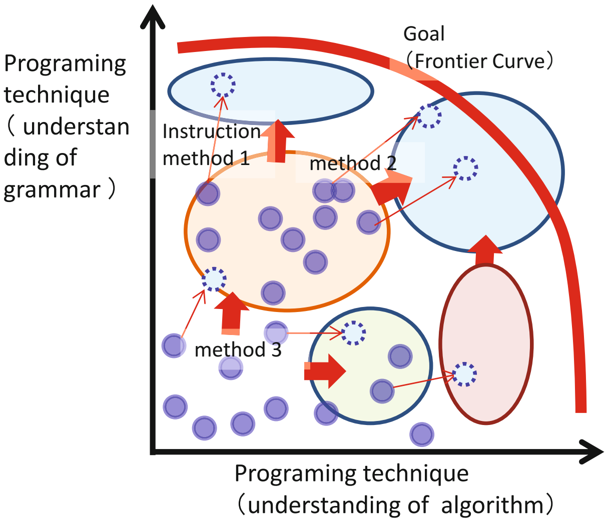
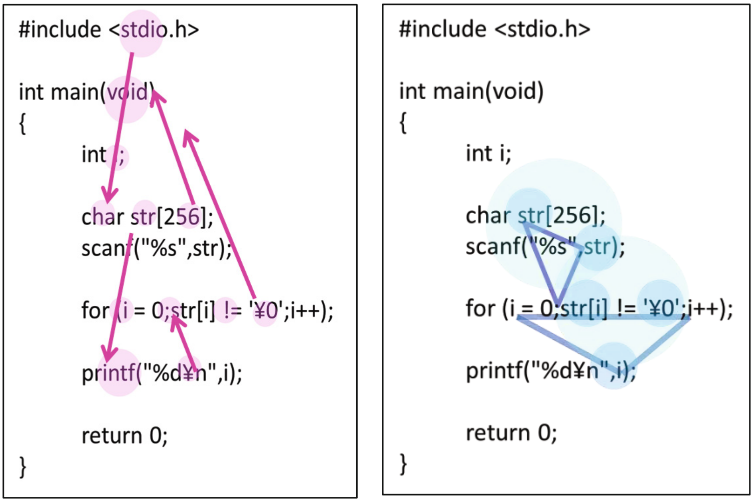

Programing technique (understan ding of grammar)
Goal
（Frontier Curve）
Instruction method 1
method 2
method 3
Programing technique
（understanding of algorithm）

Fig. 2. The example by a Data Envelopment Analysis



```
#include <stdio.h>

int main(void)
{
        int i;

        char str[256];
        scanf("%s",str);

        for (i = 0;str[i] != '¥0';i++);

        printf("%d¥n",i);

        return 0;
}
```

```
#include <stdio.h>

int main(void)
{
        int i;

        char str[256];
        scanf("%s",str);

        for (i = 0;str[i] != '¥0';i++);

        printf("%d¥n",i);

        return 0;
}
```

Fig. 3. Learner's eyes data

4 Result

The proposal of a learner's evaluation index was first created with the proposal approach. As shown in Table 1, the technical item of programming was subdivided. It roughly divided with an understanding of grammar, and an understanding of the algorithm. About the detailed classification, knowledge needed for the C language which independent administrative agency Information-technology Promotion Agency shows was created to reference [8]. It becomes possible to also classify a learner according to these classifications. Next, eyes measurement experiment and analysis were conducted. It let the learner read a program and created two or more questions to which I get it to reply what kind of result to be displayed. The thinking process of the program was measured this time using the noncontact eyes measuring instrument. The eyes data shown in Fig. 3 is the learner who answered the inaccurate solution both.

However, it is shown by the learner from eyes data that degree of comprehension differs. Specifically, it turns out that the user 1 understands the line of the important key of a program. Although the data flow of the for sentence of a loop is also understood, not having led in the answer is shown. About the user 2, he cannot understand the important point of a program, but it can observe signs that it rereads repeatedly to carry out one sentence and one-sentence understanding carefully repeatedly. Thus, it became clear that a learner's degree of comprehension and character can be read in eyes information according to a learner.

5 Conclusion

In this study, the thinking process of programming based on an eyes course was presumed for the purpose of making of the study training index which is useful for programming instruction. As a result, it became clear that it becomes possible to obtain the achievement level over the technical element according to the learner who was not evaluated until now. We would like to conduct many experiments, to propose the optimal career according to detailed skill investigation and skill, and to evaluate whether it is applicable in the future.

Acknowledgments. This work was partly supported by the Grant-in-Aid for Young Scientists (B-25750092), The Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Gog, T., Scheiter, K.: Eye Tracking as a Tool to Study and Enhance Multimedia Learning. Learning and Instruction 20(2), 95–99 (2010)
2. Lau, W., Yuen, A.: Modelling Programming Performance: Beyond the Influence of Learner Characteristics. Computers & Education 57(1), 1202–1213 (2011)

3. Kiss, G.: Teaching Programming in the Higher Education not for Engineering Students. Procedia - Social and Behavioral Sciences 103(26), 922–927 (2013)
4. Othman, M., Othman, M., Hussain, F.: Designing Prototype Model of an Online Collaborative Learning System for Introductory Computer Programming Course. Procedia - Social and Behavioral Sciences 90(10), 293–302 (2013)
5. Emrouznejad, A., Parkerb, B., Tavares, G.: Evaluation of Research in Efficiency and Productivity: A Survey and Analysis of the First 30 Years of Scholarly Literature in DEA. Socio-Economic Planning Sciences 42(3), 151–157 (2008)
6. Fushida, K., Tamada, H., Iguki, H., Fujiwara, K., Yoshida, N.: Coding Pattern Analysis for Novice Programmer in Programming Exercise, Technical Report of the Institute of Electronics, Information and Communication Engineers, Vol.111(481), SS2011-68, 67-72 (2012) (in Japanese)
7. Hofuku, Y., Cho, S., Nishida, T., Kanemune, S.: A Class Analysis by "De-Gapper": The Tool to Detect Gaps between Programs. Information Processing Society of Japan, SIG Notes 2013-CE-121(8), 1–6 (2013) (in Japanese)
8. Ogino, A., Tamada, H., Ueda, H.: Phynocation: A Prototyping of a Teaching Assistant Robot for C Language Class. In: Stephanidis, C. (ed.) Universal Access in HCI, Part IV, HCII 2011. LNCS, vol. 6768, pp. 597–604. Springer, Heidelberg (2011)
9. Tamada, H., Ogino, A., Ueda, H.: A Framework for Programming Process Measurement and Compiling Error Interpretation for Novice Programmers. In: Proc. of 2011 Joint Conference of the 21st International Workshop on and 6th International Conference on Software Process and Product Management, pp. 233–238 (2011)
10. Ishikawa, H., Maruyama, K., Terada, M.: Programming Study Support Using the Review Function of SNS. Proc. of Forum on Information Technology 11(3), 617–622 (2012) (in Japanese)
11. Ito, K., Sato, T., Tsubakimoto, M.: Analyzing Case Examples of a Self-Study Environment for Programing Education. Research Report of Japanese Society for Information and Systems in Education 27(4), 9–13 (2012) (in Japanese)

Expressing Observation Direction through Face and Body Rotation in a Multi-user Conversation Setting

Satoshi Mieda, Shiro Ozawa, Munekazu Date, Hideaki Takada,
Yoshiaki Kurokawa, and Akira Kojima

NTT Media Intelligence Laboratories, NTT Corporation
{mieda.satoshi,ozawa.shiro,date.munekazu,takada.hideaki,
kurokawa.yoshiaki,kojima.akira}@lab.ntt.co.jp

Abstract. In this paper we clarified the range of observing direction by rotating the 2D human image and it is possible to express the observing direction by face direction. We conducted two subjective experiments about direction expression of the person on an image. In the first experiment, we compared two types of human image expression, rotated 2D human image of rotated 2D and direction correct. In the second experiment, we evaluated the effect of human image rotation and the criterion for judging the direction. We showed that the direction of the user's face is the main factor in expressing the observation direction. Results clearly showed that it is possible to express the observation direction, which is required for effective communication, by using only the rotation of human facial image.

Keywords: communication, remote, human expression.

1 Introduction

We are interested in enhancing communication between persons in locations remote from each other through the transmission of nonverbal information, such as the direction in which users are looking and the hand gestures they make. In order to show users and their objects of interest, it is very important to convey the direction in which they are looking when they are observing other users. With current video conference systems, it is difficult to accurately transfer the observation directions when two or more users are involved, so remote site users may feel that other users are looking at themselves.

2 Related Work

One approach to solving this problem is the use of multiple cameras to acquire and display multiple gaze directions [1]. Another is the use of a 3D display to show the directions [2] [3] [4]. However, these approaches require the use of many cameras or a special display device.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 273–280, 2014.

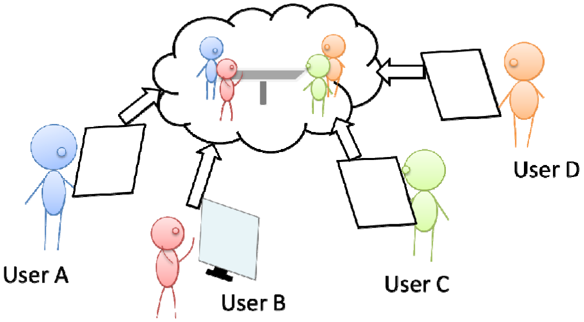


Fig. 1. Proposed communication system. Multiple users are arranged in the virtual space and can communicate with each other as if they were actually meeting together.

In another approach that has been proposed, remote participants are arranged in an actual space and the gaze direction is presented by giving a physical direction [5]. However, it requires special displays with a pan and tilt mechanism for remote participants. It is said that the normally 2D human image is perceived as always facing to the observer [6]. This is called the Mona Lisa effect.

The approach we propose involves a multiuser communication environment we are researching, which is arranged in a virtual space (Fig. 1). It comprises a desktop conference system in which simple equipment (a web camera and display) is used and only images showing a user's front view can be taken. In this paper, we report an experiment we conducted to validate our approach, in which a remote user's observation direction is expressed by using his or her front side facial image in a simple setting and rotating the image in the virtual space.

3 Evaluation of User's Observing Direction

Our goal is to build the environment where multiple people join and talk smoothly. So, we conduct a subjective experiment about the person's direction expression. We evaluated the effect the direction in which the remote user is thought to be observing and the effect of human image rotation, the criterion for judging the remote user's observing direction. We conduct two experiments. As first experiment, we confirm the effect of rotating human image. Secondly, we evaluate the effect of the difference the face and the body image.

3.1 Experimental Settings

Figure 2 shows the experimental setting. In this experiment we assume the environment where four users talk surrounding the round table. In order to display a remote person at actual size, we use 40 inch LCD (1,920 x 1,080 pixels, 880 x 500 mm). We set the distance between the display to the user 600mm, and the distance to the facing

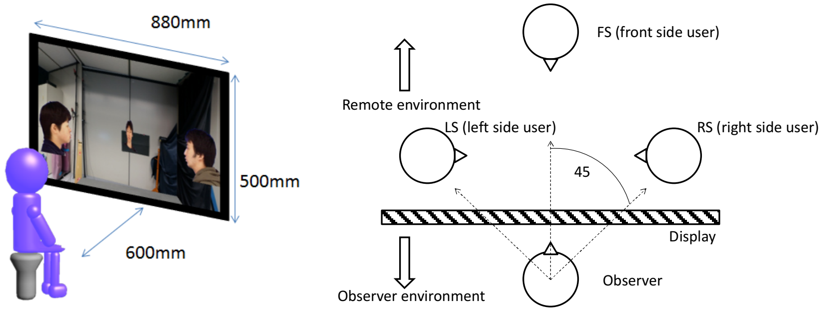


Fig. 2. Arrangement of users in experimental setting. In the left side, arrangement of observer and a display is shown. In the right side, The FS (front side user) is located in front of the observer, while the LS (left side user) and RS (right side user) are located at a 45 degree angle from the observer.

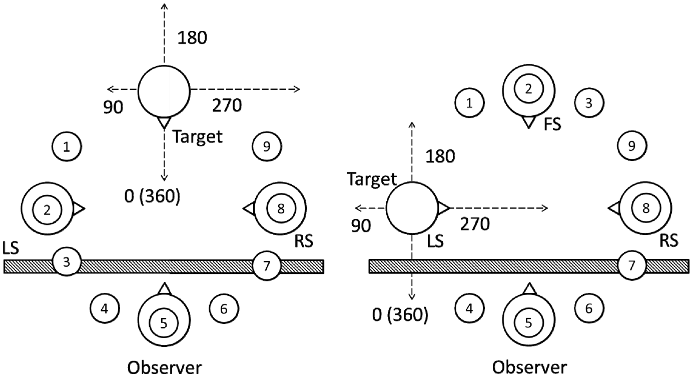


Fig. 3. The left side figure shows the answer direction when the target user is located in front of the observer. The right side figure shows the answer direction when the target user is located at left side of the observer.

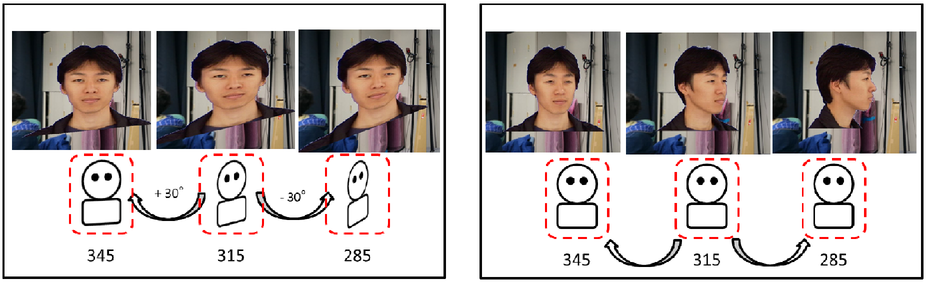


Fig. 4. Images of remote user located at a 45 degree angle to observer in a 3D virtual environment. Images in the left box are obtained by rotating human images; those in the right box are obtained with human image of directional correct.

partner who is displayed in the monitor is virtually 1,200 mm, which is possible distance to talk natural. The subject is seated in front of the display which is placed on 820mm high table. We construct 3D virtual space using OpenGL, and human images are arranged at the 3D position in the 3D virtual space. In the experiment we don't consider about displaying 3D such as binocular parallax or motion parallax because we are interested in plane and three-dimensional rotations of 2D images which can show high resolution and use easily.

3.2 Task and Experimental Design

We conducted a judging the person's observing direction task in a virtual environment. The experimental factors were displayed human image (rotating image, direction correct image), position of target user to answer (left side, front) and user directions. We simply show user's front image which is rotated depending on observing direction. We also show the direction correct user's image as observing direction to compare. The number of target user's positions was two because we considered about symmetry. The user directions were when the user was the front, two targets, and, when the user was the left side, three targets. We showed the human image which angle of images rotated at from -45 deg. to 45 deg. every 15 deg. to each target (7 patterns to each targets). Figure 4 shows two kinds of human image and target user's positions. Using all combinations, subjects did 70 trials three times for a total of 210 trials. We presented the each image randomly. Before starting experiment, the seat height was adjusted to the eye height of human image. In the task, a subject answered which direction a specified human image is observing by a number. Figure 3 shows the relation of the number and direction. Subjects included 4 men ranging from 25 to 30 years old, and the total time was about 20 minutes. Figure 5 shows the example of actual experiment.

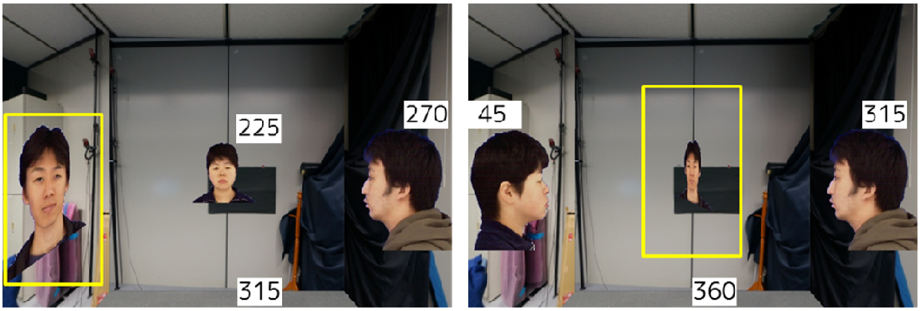


Fig. 5. Images used in actual experiment. The subject should answer the user observation direction surrounded yellow rectangle.

3.3 Result of Sotating Human Image

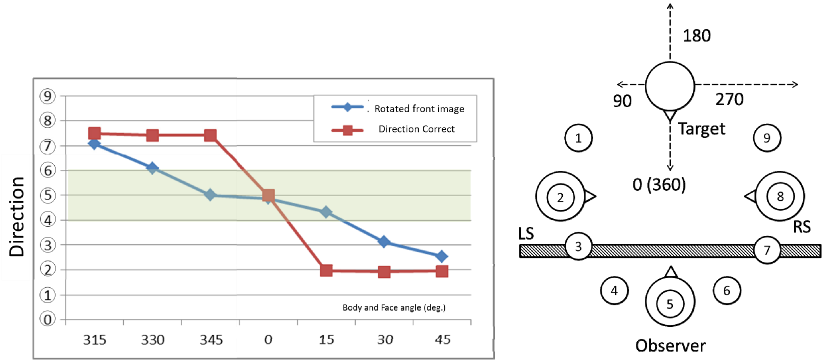


Fig. 6. The graph at left shows the perceived user direction when the target user is located in front of the observer and looking around the observer. The figure at right side shows the relation between answer number and observing direction.

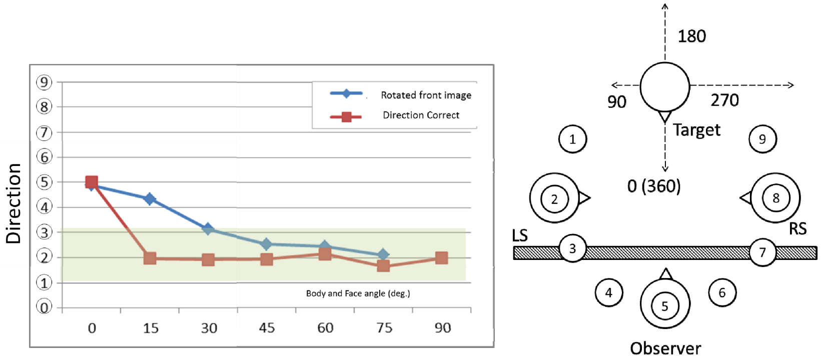


Fig. 7. The graph at left shows the perceived user direction when the target user is located in front of the observer and looking around the observer. The figure at right side shows the relation between answer number and observing direction.

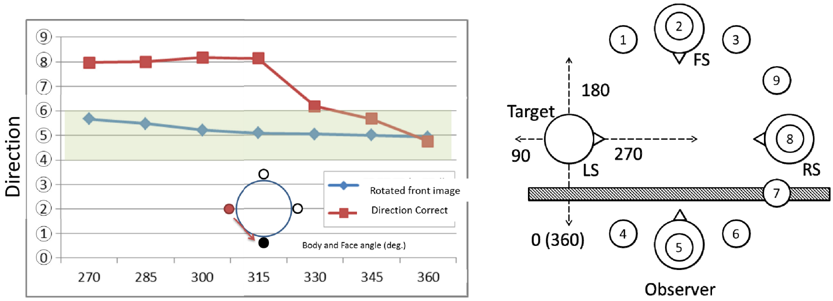


Fig. 8. The graph at left shows the perceived user direction when the target user is located at left side of the observer and looking around the observer. The figure at right side shows the relation between answer number and observing direction.

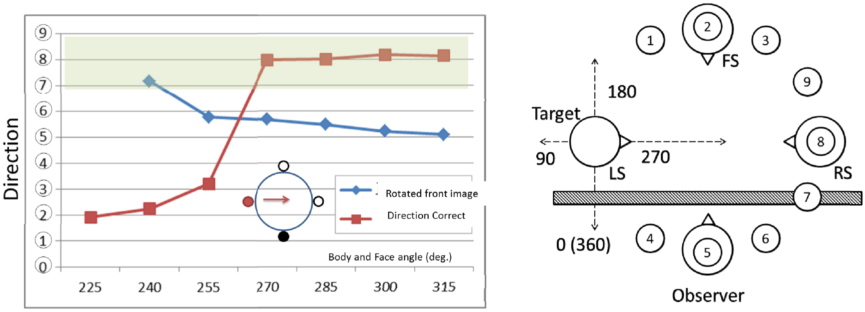


Fig. 9. The graph at right shows the perceived user direction when the target user is located at left side of the observer and looking around the observer. The figure at right side shows the relation between answer number and observing direction.

Figure 6 and 7 show the result of the answered direction when the target user is in front of subjects and figure 8 and 9 show the result of the answered direction when the target user is left side of subjects. Experiment 1 clarified that when a remote user is positioned such that he or she appears to be directly in front of the observer, the user's observation direction can be expressed only by rotating a 2D human image about z axis so that it is arranged to a 3D virtual space. Since it is difficult to acquire and carry out segmentation of the user's front view, we verify which element is most effective with respect to the observation direction of a person.

4 Evaluation of the Different Direction Face and Body

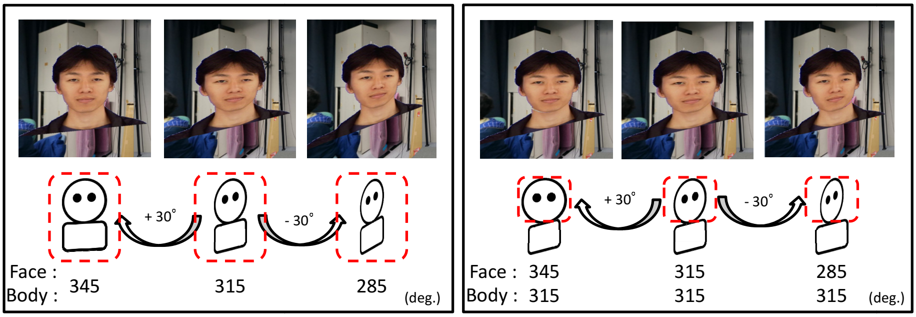


Fig. 10. Images of a remote user located at a 45 degree angle to the observer in a 3D virtual environment. Images in the left box were obtained by rotating the face and body together; those in the right box were obtained with our technique that rotates the face and body separately.

4.1 Experimental Setting

The arrangement of the users in the experimental setting is shown in Fig. 2. Four users join the environment, and one of them (the observer) observes the other three via a monitor. The observer is shown images of the three persons and reports the observation direction he or she perceives for a specific person. The face and body are given different rotation angles in the image. The facial angle, body angle, and the position of the target user were given as experimental factors. Figure 10 shows an example of images used in the experiment.

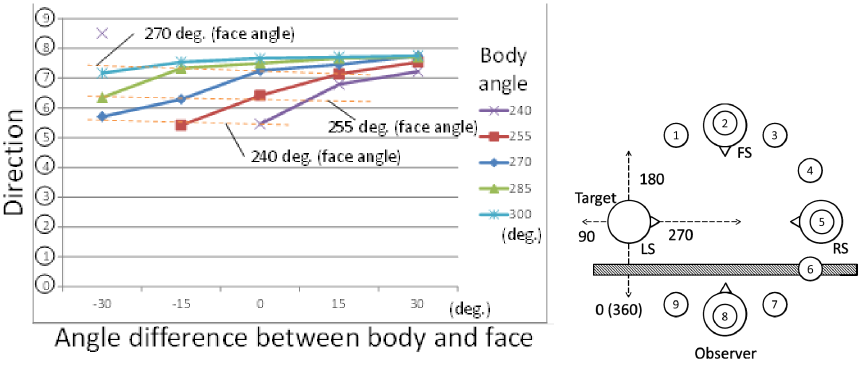


Fig. 11. The graph at left shows the perceived gaze direction relative to body angle with rotation angle difference given to the face and body. The dotted lines indicate the angle of the same face. The image at right shows the relations among the observation direction, the answer number, and the LS rotation angle.

4.2 Result of Express the Different Direction Face and Body

Figure 11 shows the experiment results obtained when the target person was the LS. It shows that the gaze direction of the LS is expressed to the observer, if the direction of the image is rotated to the observer side rather than to the direction of the RS, who is located in front of the LS, and if the direction of the front of the LS is widened, the direction other than that the observer will be shown. The face direction is dominant in expression of the observation direction; the body direction is largely irrelevant to it.

5 Discussion and Impact

We aim at achieving an environment in which four users communicate in virtual space using one camera and one display for each user (Fig. 1). We showed that the direction of the user's face is the main factor in expressing the observation direction, and that the observation direction can be shown without acquiring the image of a user's whole body.

Experiment results clearly showed that it is possible to express the observation direction, which is required for effective communication, by using only the rotation of human facial image, and to achieve virtual space communication among users coming together from remote locations in a communication system arranged in a virtual space where images of the users' front view can be taken.

References

1. Nguyen, D., Canny, J.: Multiview: Spatially faithful group video conferencing. In: Proc. CHI 2005 (2006)
2. Feldmann, I., Waizenegger, W., Atzpadin, N., Schreer, O.: Real-time depth estimation for Immersive 3D video vonferencing. In: Proc. 3DTV-CON 2010, pp. 1–4 (2010)
3. Jones, A., Lang, M., Fyfe, G., Yu, X., Busch, J., Mcdowall, I., Bolas, M., Debevec, P.: Achieving eye contact in a one-to-many 3D video teleconferencing system. ACM Transactions on Graphics 28(3), Article 64 (2009)
4. Iso, K., Date, M., Takada, H., Andoh, Y., Ozawa, S., Matsuura, N.: Video conference 3D display that fuses images to replicate gaze direction. In: Proc. IAS 2011, pp. 1–6 (2011)
5. Otsuka, K., Mucha, K.S., Kumano, S., Mikami, D., Matsuda, M., Yamato, J.: A system for reconstructing multiparty conversation field based on augmented head motion by dynamic projection. In: Proc. MM 2011, pp. 763–764 (2011)
6. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psycologica 26, 22–63 (1967)

Gaze Location Prediction with Depth Features as Auxiliary Information

Redwan Abdo A. Mohammed, Lars Schwabe, and Oliver Staadt

University of Rostock, Institute of Computer Science, Rostock, Germany
{redwan.mohammed,lars.schwabe,oliver.staadt}@uni-rostock.de

Abstract. We present the results of a first experimental study to improve the computation of saliency maps, by using luminance and depth images features. More specifically, we have recorded the center of gaze of users when they were viewing natural scenes. We used machine learning techniques to train a bottom-up, top-down model of saliency based on 2D and depth features/cues. We found that models trained on Itti & Koch and depth features combined outperform models trained on other individual features (i.e. only Gabor filter responses or only depth features), or trained on combination of these features. As a consequence, depth features combined with Itti & Koch features improve the prediction of gaze locations. This first characterization of using joint luminance and depth features is an important step towards developing models of eye movements, which operate well under natural conditions such as those encountered in HCI settings.

1 Introduction

Being able to predict gaze locations, as compared to only measuring them, is desirable in many application scenarios such as video compression, the design of web pages and commercials adaptive user interfaces, interactive visualization, or attention management systems[14,5]. However, eye movements are known to depend on task demands, in other words, information not present in the visual stimulus. As a consequence, algorithms based on the computation of salient locations from only the bottom-up visual signals have principled limitations in gaze prediction.

Eye movements have been predicted mainly using purely stimulus-driven models. Most models of saliency [6,8] are biologically inspired and based on a bottom-up computational model which does *not* take into account contextual factors or the goal of a user in a visual task. Multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted from the image at multiple scales. Then, a saliency map is computed for each of the features and combined in a linear or non-linear fashion into a master saliency map that represents the saliency of each pixel. This idea of saliency maps was used in other studies, where it was extended and further developed. For example, Mahadevan and Vasconcelos [3] proposed a discriminant formulation of center-surround

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 281–292, 2014.

saliency for static images. One can view their work as a normative approach, because they first formulate the saliency map computation as a problem, and then derive their algorithm as the solution to this problem. More specifically, they consider saliency as a decision making task informed by natural image statistics. The outcome of their work is an automatic selection of the important features. This improves the original Itti & Koch model [6], where the features selection and combination was done in a heuristic way. This was later also extended to dynamic scenes and movies using dynamic textures [8]. However, the original Itti & Koch model was also improved recently using graphs to compute saliency [4]. This shows that the concept of saliency maps is still very fruitful and can guide research in predicting eye movements. These saliency-based models are all based on low-level image features. Despite this limitation, they often predict gaze well, but mid- and high-level features also affect gaze. Therefore, Judd et al. [17] pursued a machine learning approach: They learned gaze points based on measured eye movements using a linear SVM and low-, mid- and high-level features. They reported better predictions than Itti & Koch on 1003 images observed by 15 subjects [17].

Another line of research has investigated the depth structure of natural scenes using range sensors [12,18,10]. This depth structure is not directly accessible to the human vision system and needs to be inferred using stereo vision or other depth cues. Some statistical aspects of depth images as well as the relation between depth and luminance images have been investigated before [18,10,13], but the statistical properties of depth images at the center of gaze are not clear [11]. For example, simple questions such as "Do humans look more often to high contrast edges due to depth gaps than to edges due to texture borders?" have not been addressed yet [11]. It was shown, however, that eye movements are far from a random sampling. It was even suggested that the statistics of natural images differ at the center of gaze when compared to random sampling [13]. Thus, taking into account eye movements is essential for shaping artificial vision systems via natural images. In [9] we have analyzed the saliency in 2D pixel and depth images using a very simple feature: the local standard deviation of pixels. We found that saliency in depth images is bimodally distributed with highly salient locations corresponding to low salient 2D image locations. Given that most saliency algorithms work on the 2D images, this finding points towards including depth cues into the computation of saliency maps.

In this paper, we present the results of a first experimental study to further improve the computation of saliency maps. More specifically, We have recorded the center of gaze of users when they were viewing natural scenes. We first examined the statistical characterization of depth features in natural scenes at the center of gaze. The rational for investigating depth images is that they may reveal the "saliency that matters", because when interacting with the environment we evolved by interacting with objects in a three dimensional (3D) world. Thus, we hypothesize that saliency maps respecting this will ultimately outperform saliency maps computed only on the basis of 2D pixel images in terms of predicting eye movements. We then examined the presence of depth features around gaze locations. We used machine learning to train a bottom-up, top-down model of saliency based on 2D and depth features/cues. We used different performances distance measures. We found that models trained on Itti & Koch and depth features combined outperforms models trained on other individual features or other pairs of features combined.

This paper is organized as follows: First, we describe the material and methods including the image material (Sec. 2.1) and the features we extracted from the luminance and depth images (Sec. 2.3 and 2.4). Then, we present the results of our analysis, where we first compared the distribution of depth values of patches in the center of gaze to that expected from random sampling (Sec. 3) and then gaze location prediction when viewing photos of natural scenes (Sec. 4).

2 Material and Methods

2.1 Stimulus Material

Forty images obtained originally from Make3D project Range and Image Dataset [15,16] were presented to five subjects. The 2D color pixel images were recorded with a resolution of 1704 × 2272 pixels, but the depth images with a resolution of 305 × 55 pixels. They where 40 images from “forest scene”, “city scene”, and “landscape scene”. The users were males and females between the ages of 18 and 35. Three of the viewers were researchers in institute of computer science and the others were naive viewers. All viewers sit at a distance of approximately *1.5 m* from the computer screen of resolution 1280x1024 in a dark room and used a chin rest combined with a bite bar to stabilize their head. An mobile eye tracker recorded their gaze path on a separate computer as they viewed each image at full resolution for ten seconds separated by two seconds of viewing a gray screen.

2.2 Measuring Gaze Locations

An iView X HED 4 Eye Tracking System (SMI) was used to record eye position. The eye tracker uses two cameras. The first is used to track the pupil and the second camera records the scene view. The gaze position is reported with a sampling rate of 50 Hz and a reported accuracy of 0.5°-1°. We used the default lens (3.6 mm) for the scene camera which provides a viewing angle of ±31° horizontally and ±22° vertically. The scene camera resolution is 752×480. Then, to avoid parallax error, we calibrated in a distance within 1-1.5 m. We used a calibration with five points so that the SMI recording software can compute the gaze location in scene camera coordinates from the recorded pupil images. The scene camera of the eye tracker delivers RGB frames as well as gaze locations, both with time stamps (Figure 1 a), Also we recorded information about which and when each image have been presented to the viewer. Our analysis were all done offline. First we aligned the frames temporally to the high resolution images using the information we recorded about when each image have been presented to the viewer. Then we used normalized Cross-Correlation [7] to register each

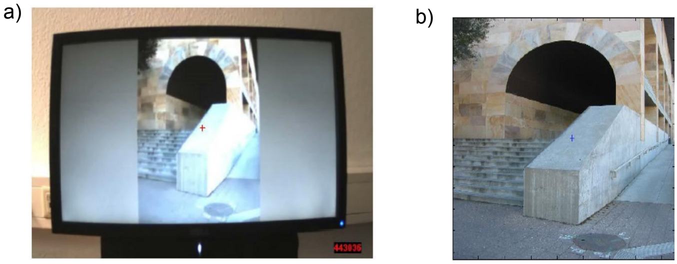


Fig. 1. Example of a gaze registration. **a)** Frame from the scene camera of the eye tracker and the corresponding gaze point (Red cross). **b)** Registered gaze point (Blue cross) on the corresponding high resolution image.

part of interest in each frame to the corresponding high resolution image. Using the transformation obtained to register each gaze point to the high resolution image (Figure 1 b), we generated a saliency map of the locations fixated by each viewer. Also, we convolve a Gaussian filter across the user's fixation locations in order to obtain a continuous saliency map of an image from the eye tracking data of a user.

2.3 Features of Luminance Images

Different low-level features were collected. For example: the intensity, orientation and color contrast channels as calculated by Itti and Koch's saliency method [6]. Also, each gray-scale image is linearly decomposed into a set of edge feature responses to Gabor filters with different orientations. We used orientations $\theta = \{0°, 15°, \ldots, 165°\}$, but only one frequency and two spatial phases. Within each image we subtracted the mean from the filter responses to each orientation, and normalized the responses to the interval between -1 and 1. We used Gabor filters responses to compare the performance with the 3D edges.

2.4 Features of Depth Images

Gap Discontinuity. A gap discontinuity in the underlying 3D structure is a significant depth difference in a small neighborhood. We measure gap discontinuity μ_{GD} by computing the maximum difference in depth between the depth of a pixel in the depth image and at its eight neighboring pixel. Here, we considered the methods presented in [19]; μ_{GD} for a point (x, y) is defined as:

$$\mu_{GD}\,(x, y) = \max\{\; |\; z(x, y) - z(x + i, y + j)\; | : -1 \leq i, j \leq 1\}\,, \qquad (1)$$

where $z\,(x, y)$ represents a depth value. This quantity is then thresholded to generate a binary gap discontinuity map. In our analysis, we have empirically chosen a threshold $\mu_{GD}\,(x, y) > T_d$ where $T_d = 0.5$. Fig. 2 (b) shows an illustration of a gap discontinuity map.

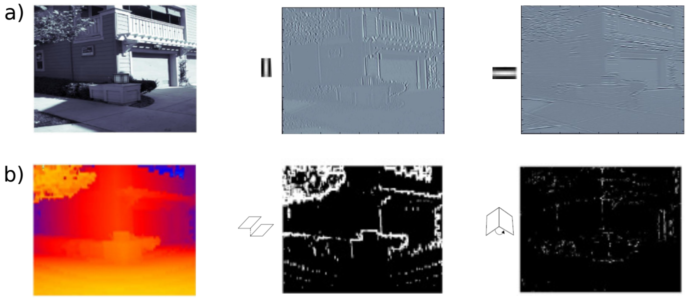


Fig. 2. Examples for features in luminance and depth images. **a)** A gray-scale image convolved with two Gabor filters selective for the same spatial frequency, but different orientation. **b)** A depth map (left) decomposed into its discontinuity maps: gap discontinuity map (middle) and orientation discontinuity map (right).

Surface Orientation Discontinuity. An orientation discontinuity is present when two surfaces meet with significantly different 3D orientations. Orientation discontinuity was measured using surface normal analysis. Here, we considered the methods presented in [1,19]. The orientation discontinuity measure μ_{OD} is computed as the maximum angular difference between adjacent unit surfaces normal. First, a three dimensional point cloud was constructed from the X, Y, Z coordinates for each pixel in a depth image. Then, each pixel is represented by a pixel patch $P_{(x,y,z)}$ compiled from the eight neighboring points in the point cloud. Finally, the unit surfaces normal are computed for each patch $P_{(x,y,z)}$ using Singular Value Decomposition (SVD).

More specifically, for an image patch $P_{(x,y,z)}$ the orientation discontinuity is defined as

$$\mu_{OD}\left(P_{(x,y,z)}\right) = \max\left\{\alpha\left(normal\left(P_{(x,y,z)}\right), normal\left(P_{(x+i,\, y+j,\, \mathrm{z+k})}\right)\right)\right\} \tag{2}$$

where $-1 \leq i, j, k \leq 1$ and $normal\left(P_{(x,y,z)}\right)$: is a function, which computes the unit surface normal of a patch $P_{(x,y,z)}$ in 3D coordinates using Singular Value Decomposition (SVD), α is a function computing the angle between adjacent unit surfaces normal. It is given by

$$\alpha\left(P_1, P_2\right) = \arccos\left(normal\left(P_1\right) \cdot normal\left(P_2\right)\right). \tag{3}$$

max is function to compute the maximum angular difference between adjacent unit surfaces normal. This measure is also thresholded, but based on two criteria, namely i) an *angular criterion*: the maximum angular difference between adjacent unit surfaces normals should be more than a threshold $T_{\theta 1}$ and less than $T_{\theta 2}$, and ii) a *distance-based criterion*: the maximum difference in depth between a point and its eight neighbor's μ_{GD} should be less than a threshold T_d.

In our analysis, we have empirically chosen $T_{\theta 1} = 20^\circ$, $T_{\theta 2} = 160^\circ$ and $T_d = 0.5$, respectively. Fig. 2b shows an illustration of an orientation discontinuity map.

2.5 Classifiers for Predicting Gaze Locations

Opposed to previous computational models that combine a set of biologically plausible filters together to estimate saliency maps, we use a learning approach to train a classifier directly from human eye tracking data. We use a linear Support Vector Machine (SVM) to find out which features are informative. We used models with linear kernels because it performed well for our specific task. Linear models are also faster to compute and the resulting weights of features are easier to understand. We divided our set of images into training images and testing images in order to train and test our model. From each image we chose 200 positively labeled pixels randomly from the top 40% salient locations of the human ground truth saliency map and 200 negatively labeled pixels from the bottom 60% salient locations. In order to have zero mean and unit variance we normalized the features of our training set and used the same normalization parameters to normalize our test data.

For each image in our dataset, we predict the saliency per pixel using a particular trained model. We used the value of $w^T x + b$ (where w and b are learned parameters and x refers to the feature vector) as a continuous saliency map which indicates how salient each pixel is. Then we threshold this saliency map at 40% percent of the image for binary saliency maps.

2.6 Error Measure

The Kullback–Leibler (KL) divergence was used to measure the distance between distributions of saliency values at human vs. random eye positions. We used KL because KL is sensitive to any difference between the histograms, where other measures essentially calculate the rightward shift of histogram1 relative to the histogram2. Also KL is invariant to reparameterizations, such that applying any continuous monotonic nonlinearity to estimated saliency map values[2]. Let $ti = 1 \cdots N$ be N human eye positions in the experimental session. For a saliency model, Estimated Saliency Map is sampled at the human saccade $X_{i,Human}$ and at a random point $X_{i,random}$. First the saliency magnitude at the sampled locations is normalized to the range [0,1]. Then histogram of these values in q=10 bins across all eye positions is calculated. $\Pr(X_{Human}(i))$ and $\Pr(X_{random}(i))$ are the fraction of points in bin i for salient and random points. Finally the difference between these histograms was measured using KL divergence is:

$$KL\left(X_{Human}; X_{random}\right) = \sum_{i}^{q} \Pr\left(X_{Human}(i)\right) \log\left(\frac{\Pr\left(X_{Human}(i)\right)}{\Pr\left(X_{random}(i)\right)}\right). \quad (4)$$

Models that can better predict human fixations show higher KL divergence.

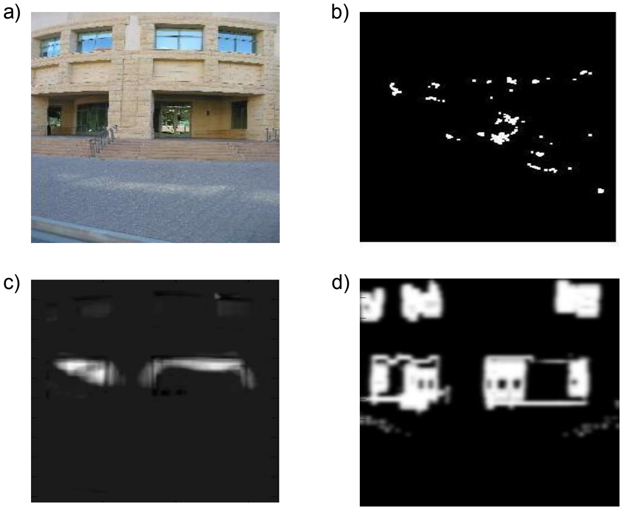


Fig. 3. Examples for features in luminance and depth images. **a)** Natural scene. **b)** Fixation map recorded with our stationary setup . **c)** Itti & Koch features. **d)** Depth discontinuity features.

3 Results 1: Depth Features at the Center of Gaze

We recorded eye movements data from subjects as they viewed static images presented on a computer monitor (see section 2). For each depth image we extracted square image patches around the subject's center of gaze. We also extracted image patches selected at random positions.

3.1 Depth Values around Gaze

We first compared the distribution of depth values of patches in the center of gaze to that expected from random sampling. It is clear that, the distribution of depth values of patches at the center of gaze statistically differ than from random sampling. Figure 4 (a) shows that the normalized histogram of the random sampling from 40 scenes, averaged over all subjects, differ than the distribution of patches in the center of gaze (see Figure 4 (b)) (with P-value = 1.091e-016 of the two-side Kolmogorov–Smirnov (K-S) test with significance level of 0.05).

Figure 4(c) shows that the normalized histogram of patches in the center of gaze over 40 scenes averaged over all subjects in the first three seconds of viewing the scenes differ than the last seven seconds (see Figure 4(d)) (with P-value = 8.6504e-065 of the two-side Kolmogorov–Smirnov (K-S) test with significance level of 0.05). We repeated the statistical test with a maximum of 50m depth and the results was validated.

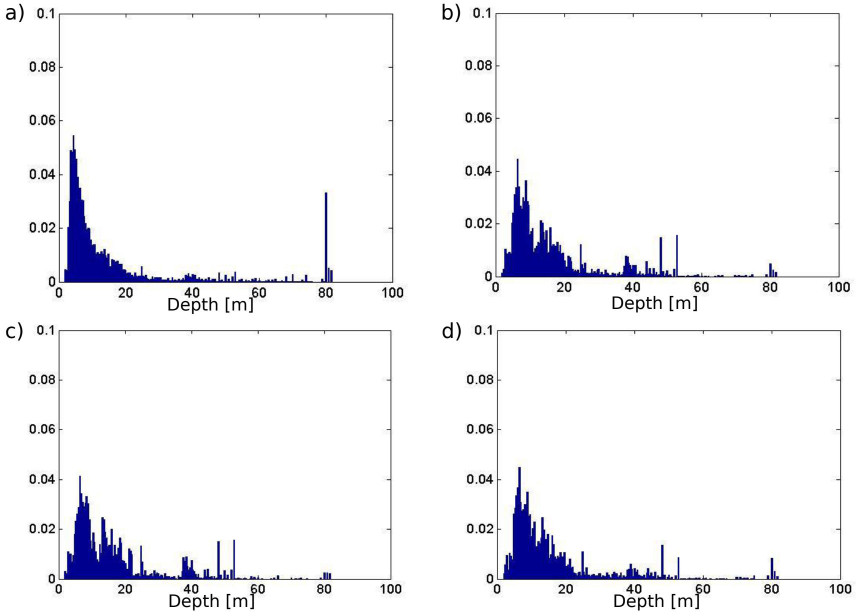


Fig. 4. a) Normalized histogram of depth values of random sampling over 40 scenes, averaged over all subjects. **b)** Normalized histogram of depth at gaze locations, averaged over all subjects. **c)** Normalized histogram of patches in the center of gaze over 40 scene for each subject in the first three seconds of viewing the scenes, averaged over all subjects. **d)** Normalized histogram of patches in the center of gaze over 40 scenes in the last seven seconds of viewing the scenes, averaged over all subjects.

3.2 Depth Features around Gaze

Before we used depth features as new information for predicting eye movements. We examined the presence of depth features around gaze locations. The result of the distribution of depth features in a different neighborhoods around the gaze location averaged over all subjects are shown in Figure 5(a) and the distribution of depth features around gaze for individual subjects are shown in Figure 5(b). It is clear that the presence of depth features around gaze locations are high. This suggest that saliency maps models respecting this will ultimately outperform saliency maps computed only on the basis of 2D pixel images in terms of predicting eye movements.

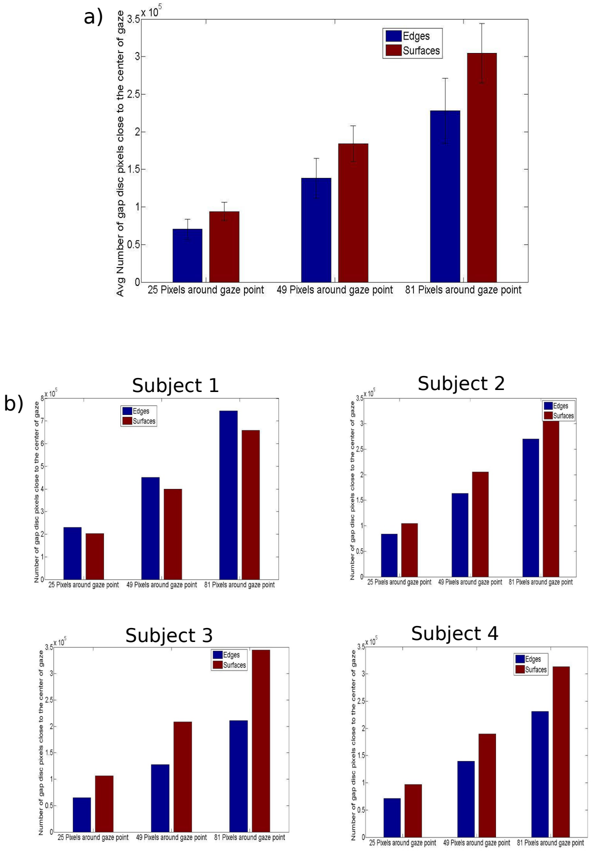


Fig. 5. The presence of depth features in a different neighborhoods around the gaze points. **a)** Bar plot for the presence of depth features in a different neighborhoods around the gaze points, averaged over all subjects. **b)** Bar plot for for the presence of depth features in a different neighborhoods around the gaze points for individual subjects.

4 Results 2: Gaze Location Prediction When Viewing Photos of Natural Scenes

We measured the performance of saliency models using KL divergence (see Section 2.6). Figure 6 describing the performance of different features models for each subject averaged over all testing images. For each image we predict the saliency per pixel using a specific trained model. We can see that the prediction differ according to the type of features we selected. While the model trained on competing saliency features from Itti and Koch perform better than the models trained on other individual features (i.e. only Gabor or only depth features). The averaged result over all subjects shows this finding (see the diagonal of Figure 7).

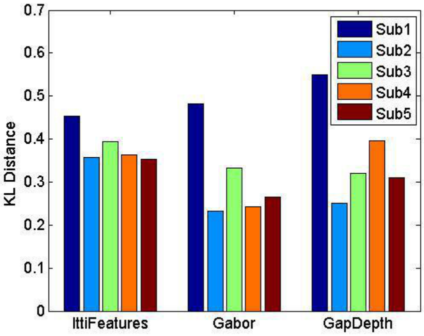


Fig. 6. The KL divergence describing the performance of different SVMs trained on each feature individually, for individual subject

Interestingly the models trained on Itti & Koch combined with depth features outperform models trained on other individual features (i.e. only Gabor or only depth features), or trained on combination of these features. (see Figure 7). It is interesting to note that, depth features combined with luminance features improve the prediction of gaze locations.

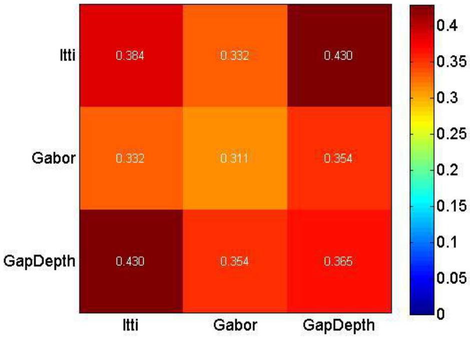


Fig. 7. The KL divergence matrix describing the performance of different SVMs models trained on set of features individually and pairs of features combined, averaged over all subjects. The main diagonal shows the performance of the models trained on individual features. The lower/ upper triangular parts of the matrix show the performance of the models trained on pairs of features combined.

Finally, the overall summary of our analysis is shown in Figure 7 where we computed the KL performance for SVMs trained with different individual features and combined together, averaged over all subjects. We perform the statistical test (t-test2) for all pairs of features (i.e. KL_Itti vs KL_Gabor, KL_Itti vs KL_GapDepth and KL_Gabor vs KL_GapDepth) with significance level of 0.05 the corresponding P-values were (0.3740, 0.9240 and 0.4488) respectively.

In Figure 7, we see the KL divergence matrix describing the performance of different SVMs models averaged over all subjects. The KL divergence matrix are symmetric with respect to the main diagonal. The main diagonal shows the performance for SVMs models trained on individual features. The lower/ upper triangular parts of the matrix show the performance for SVMs models trained on pairs of features combined.

5 Conclusion

We have analyzed the statistical of depth features in natural natural scenes at the center of gaze. We found that the distribution of depth values of patches at the center of gaze differ than from random sampling. Most interestingly, we found that the presence of depth features around gaze locations were high. This finding points us towards including depth cues into the computation of saliency maps as a promising approach to improve their plausibility.

We also used machine learning to train a bottom-up, top-down model of saliency based on 2D and depth features. We found that models trained on Itti & Koch and depth features combined outperform models trained on other individual features (i.e. only Gabor filter responses or only depth features), or trained on combination of these features. As a consequence, depth features combined with Itti & Koch features improve the prediction of gaze locations.

Our approach, of using joint luminance and depth features is an important step towards developing models of eye movements, which operate well under natural conditions such as those encountered in HCI settings.

Acknowledgments. This work was supported by the DFG GRK 1424 MuSAMA. We also acknowledge the dataset of Laser depth data and luminance Images taken from the Make3D project [15,16].

References

1. Hoover, A., Jean-Baptiste, G., Jiang, X.: An experimental comparison of range image segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 673–689 (1996)
2. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. 35(1), 185–207 (2013)
3. Gao, D., Vasconcelos, N.: Discriminant saliency for visual recognition from cluttered scenes. In: NIPS (2004)
4. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems 19, pp. 545–552. MIT Press (2007)
5. Horvitz, E., Kadie, C., Paek, T., Hovel, D.: Models of attention in computing and communication: From principles to applications (2003)

6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
7. Lewis, J.P.: Fast normalized cross-correlation (1995)
8. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. 32(1), 171–177 (2010)
9. Mohammed, R.A.A., Schwabe, L.: Scene-dependence of saliency maps of natural luminance and depth images. In: Fifth Baltic Conference "Human - Computer Interaction" (2011) (to appear)
10. Mohammed, R.A.A., Schwabe, L.: A brain informatics approach to explain the oblique effect via depth statistics. In: Zanzotto, F.M., Tsumoto, S., Taatgen, N., Yao, Y. (eds.) BI 2012. LNCS, vol. 7670, pp. 97–106. Springer, Heidelberg (2012)
11. Mohammed, R.A.A., Mohammed, S.A., Schwabe, L.: Batgaze: A new tool to measure depth features at the center of gaze during free viewing. In: Zanzotto, F.M., Tsumoto, S., Taatgen, N., Yao, Y. (eds.) BI 2012. LNCS, vol. 7670, pp. 85–96. Springer, Heidelberg (2012)
12. Potetz, B., Lee, T.S.: Statistical correlations between 2d images and 3d structures in natural scenes. Journal of Optical Society of America, A 7(20), 1292–1303 (2003)
13. Reinagel, P., Zador, A.M.: Natural scene statistics at the centre of gaze. Network 10(4), 341–350 (1999)
14. Roda, C.: Human Attention in Digital Environments. Cambridge University Press, Cambridge (2011)
15. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS 18. MIT Press (2005)
16. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 824–840 (2009)
17. Durand, F., Judd, T., Ehinger, K., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
18. Yang, Z., Purves, D.: Image source statistics of surfaces in natural scenes. Network: Computation in Neural Systems 14(3), 371–390 (2003)
19. Yokoya, N., Levine, M.D.: Range image segmentation based on differential geometry: A hybrid approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(6), 643–649 (1989)

Study and Evaluation of Separability Techniques and Occlusion in Multitouch Surfaces

Jessica Palomares, Manuel Loaiza, and Alberto Raposo

TECGRAF - Technical-Scientific Software Development Institute
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil
jessika.palomares@gmail.com,
{manuel,abraposo}@tecgraf.puc-rio.br

Abstract. Multitouch interfaces allow interacting with a virtual object directly, similar to a real object. However, there are several issues to be resolved, such as the accuracy of the manipulation, the occlusion, the separability of the manipulation, etc. Multitouch interfaces allow multiple spatial transformations that can be performed on a virtual object with only a gesture. For example, an object can be rotated, translated and scaled with two fingers with a single gesture. However, some unwanted movements may occur accidentally. Separability techniques appear with the intent to prevent unwanted movements on multitouch surfaces. Occlusion is another problem that occurs in multitouch interfaces. Often the user's hand hides the vision of the object with which he/she interacts; or the user's action on interface hinders the movement when it clicks on a bottom that triggers action. This paper proposes two techniques of separability, aiming to reduce the problems that arise due to excessive freedom of manipulation in multi-touch interfaces, and evaluates the efficiency of these techniques. The techniques developed are not only applicable in simple virtual objects; they are also for WIMP (windows, icons, menus, pointer) objects, aiming to reduce occlusion. A series of tests was performed to evaluate precision, occlusion time for completion of task, and ease of use.

Keywords: Human-Computer Interaction, multitouch interaction, Separability, Occlusion, Spatial Tranformation.

1 Introduction

Touch-sensitive surfaces appeared as a means to provide a more direct and natural human-computer interaction, allowing creating an alternative to mouse-based interfaces. Among the devices using this technology, we find from mobile individual devices to collaborative tabletop surfaces.

Multitouch tabletop surfaces offer many advantages, such as the detection of several simultaneous touch events along the display area, allowing the parallel interaction of more than one user with one or multiple programs. This technology also enables the manipulation of graphic objects in more complex ways than it is possible

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 293–304, 2014.

with the mouse. A single gesture may generate many simultaneous spatial transformations in an object. For example, with two fingers, users may translate, rotate and scale an object at the same time.

However, there is the opposite situation, where the user only wants to make a subset of these actions, and unwished movements accidentally occur, given our imprecision with fingers movements. According to Nacenta et al. [1], it is difficult only to translate and scale an object without rotating it, since the object reacts to small angle variations among the contact points. This problem can be reduced by a separability technique applied in spatial manipulations in multitouch surfaces.

In the present work, we study the characteristics of separability and occlusion for interaction in multitouch tabletop surfaces. We chose two techniques to demonstrate the importance of these problems and to propose solutions for them. The first technique, called "Handles", explicitly separates the spatial manipulations by means of areas over the object, as described in Nacenta et al. [1]. The second technique is called "Rock & Rails" and proposes the use of a set of gestures that, combined with touches over the object, separate the spatial manipulations, as described by Widgor et al. [2]. In addition to the separability problem, we show in the present work that occlusion also interferes in the correct spatial transformation of an object, especially in the case of WIMP interfaces, still currently used in multitouch applications.

We propose the modification of Handles [1] and Rock & Rails [2] techniques that, together with other separability techniques and the study of occlusion problems, aim at providing support to the reuse of WIMP interfaces in multiuser devices, such as tabletop surfaces.

This paper is organized as follows. The following section presents some related work. Then, in section 3, the techniques proposed in this work are presented. In section 4 we describe the user tests and analyze the results. Finally, section 5 concludes the paper and indicates future work.

2 Related Work

We present related work divided into two subsections approaching, respectively, the concepts of object manipulation and separability in multitouch surfaces, and the occlusion problem in applications implemented for multitouch surfaces.

2.1 2D Objects Manipulation and Separability

Wu et al. [3] created a prototype for furniture organization in a plan called RoomPlaner. This prototype runs in a DiamondTouch [4] tabletop and can be used by two persons at the same time. Kruger et al. [5] developed the Rotate'N Translate (RNT) technique, aiming to seamlessly integrate the rotation and translation of 2D objects.

Hancock et al. [8], investigated several manipulation techniques for 2D objects and proposed the so called "two-point rotation and translation" technique, also known as rotate–scale–translate (RST), or pinch zoom, when associated only with object's scaling [6]. In this technique, the first contact point is used to move the object, while the second one is used to rotate it. This technique became very popular in multitouch interaction and is the one we are going to use in the present work as the reference technique, which we call here “no restriction technique”.

In the work of Moscovich and Hughes [7], a new approach is presented. The idea is to make a better use of the number of contact points that our fingers may offer, to map these contact points into events that the user may use to manipulate 2D objects. For example, the Sticky Fingers technique [8] proposes that it is possible to scale, rotate and translate an object using two fingers. Ashtiani e Stuerzlinger [9] introduces a transformation technique with up to three touches called XNT. In this technique, the reference point for rotation and scale transformations is the midpoint calculated using the number of contact points over an object.

In the work of Nacenta et al. [1] proposed a set of interaction techniques that allow users to select a subset of degrees of freedom. The proposed techniques reduced the unwanted manipulations without affecting the performance of manipulation. On them was separated control of the degrees of freedom, thus improving the accuracy of movement that users are able to do about the objects. In the work of Widgord et al. [2] proposed a set of gestures (Rock, Rail and Curved Rail) which, in combination with touches, limited degrees of freedom in spatial manipulations of virtual objects 2D.

Ashtiani and Stuerzlinger [9] proposed a technique called XNT and XNT-S. The technique XNT-S is a variation of the technique XNT, where it is proposed to separate the manipulations by the amount of touches on the object: translation with one touch, two touches to scale and with 3 touch rotation, it was calling XNT-S.

2.2 Occlusion

Multitouch tabletop surface provide the users with a large, horizontal and shared interaction area, offering new opportunities to support collaboration, discussion, interpretation and analysis tasks with the information presented on the surface. However, in a device where input (touches) and output exhibition areas are coincident, the user hand or arm may occlude part of the screen [10]. These occlusion problems have been studied in many Works. For instance, Vogel and Baudish [11] presented the Shift technique, with the argument that it is advantageous to keep the interaction point at the local and to exhibit a copy of the occluded area in another area of the screen.

Roudaut et al. [12] introduced the TapTap technique, designed to improve the direct touch precision, based on a temporal multiplexing strategy. Brandl et al. [13] point that multitouch tables are also affected by the user position and the use of the hands. Based on their observations about occlusion in multitouch tabletops, these authors implemented a digital menu for these surfaces. This menu provides an apperture to avoid the hand occlusion over it, and can be adapted to different hand positions.

Many other solutions have been proposed to avoid occlusion in spatial manipulations, such as a tactile cursor [7], remote manipulation [2], and the handles techniques [1].

Currently, operational systems, such as Windows 7 and 8, Mac OS X, and Linux, provide support for multitouch, and have been changing the size and content of their graphical interface elements. For instance, in some programs provided with Windows 7, such as the Paint, we observe that the toolbar is larger than in the previous version, and it has a new organization of the elements. This indicates that there is a strong tendency to support tactile events, as became clearly evident in Windows 8.

Although Windows 8 provides an interface clearly adapted to multitouch devices, the use of a mouse is still required to interact with windows, which are legacy of previous versions of the operational system. In multitouch interfaces, the direct manipulation of virtual objects is susceptible to occlusion, due to the size of users' hands and fingers. Direct touches increase occlusion, as pointed by Potter et al. [14]. In the present work, we studied the occlusion and how it affects transformation operations (rotation, translation, and scaling) in WIMP interfaces, still present in Window 8.

3 Proposed Separability Techniques

The great advantage of direct manipulation is that it has the potential to increase the velocity of complex manipulations, because this kind of manipulation eliminates the necessity of making the transformation operations sequentially [2]. However, there are tasks that require a higher level of precision, and can be hindered by the control of more than one operation at the same time.

Multitouch tables were developed to allow that users work together and interact simultaneously with an application. The use of WIMP interfaces in multitouch tables requires an adaptation of current Windows, they need to rotate, translate and scale, so that users can execute their tasks more easily. In addition, there are problems related to the occlusion of interface elements. For example, one may trigger an unwanted event when touching a button or menu within the window area.

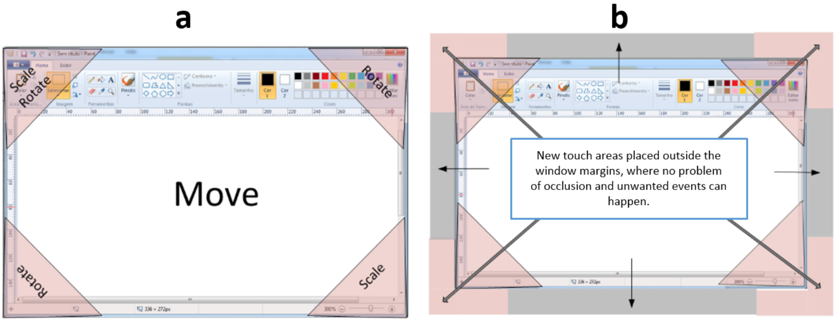
Considering these problems, the present paper proposes two new techniques to support the manipulation of WIMP interfaces in tabletop surfaces. The first technique is based on the work of Nacenta et al. [1] and is called "borders outside the object". The second one is based on the work of Wigdor et al. [2] and is called "with the help of a proxy". These techniques are described below.

3.1 Borders Outside the Object

In mouse-based interfaces, generally the manipulation or transformation of an object requires an explicit way to be executed. For example, in MS Word, if users want to rotate an image, they have to use a specific green manipulator (handle) placed at the top of the image. Handles are in fact a very common approach in traditional interfaces. They provide separability by means of the explicit selection of the transformations that can be applied over an object.

In multitouch interface, handles-based techniques were implemented in the works of Apted et al. [15] and Nacenta et al. [1]. These techniques were proposed as strategies to map and restrict direct spatial manipulation events to specific areas of the objects being manipulated in multitouch surfaces. However, both techniques above may present problems if the user manipulates objects with buttons, menus, or links, present in WIMP interfaces. A single touch over any of these elements may be interpreted as an event associated to the interface element, as a spatial manipulation event.

As an example, consider the case where the user works on a multitouch table and want to show an application window to a colleague, requiring moving and rotating that window. As shown in Figure 1a, using the original Handles techniques, as proposed by Nacenta et al. [1], when rotating the window, the user could accidentally touch window maximization or minimization buttons, or when moving the window, the user can draw something in the drawing area. For these reasons, we redefine the manipulation areas, placing them around the object, and not over it. This change avoids any misinterpretation of the events that could happen due to occlusion or to the size of the user's touch (fat fingers problem [16]).



New touch areas placed outside the window margins, where no problem of occlusion and unwanted events can happen.

Fig. 1. (a) Occlusion problem when applying Handles technique in Windows interface. (b) The first proposed technique: Borders outside the object.

Figure 1b presents our proposal. The semitransparent red areas at the four corners of the object are reserved for rotation and scaling, similar to the original Handles technique. The gray semitransparent areas at the four edges are reserved for the translation. It is important to say that, for rotation and scaling, the user has to touch simultaneously on the two areas corresponding to the transformation. For translation, the user can touch on only one gray area of Figure 1b.

Nacenta et al. [1] pointed some problems related to the areas defined for the handles. One of these problems is that the area defined for each handle is affected by the scaling of the object. If the object is reduced a lot, the user may have difficulties to select a handle. To avoid this problem in our proposal, the borders around the object (our handles area) is maintained with a fixed width, independent of the object size. This size was defined according to the work of Wang and Ren [17], where the authors indicate that interaction targets must have a size larger than 11.52mm, for square objects.

3.2 With the Help of a Proxy

Wigdor et al. [2] presented the Rock & Rails technique for multitouch manipulation based on hand gestures. In this technique, the user makes a gesture with the non-dominant hand and makes direct touches over the object with the dominant hand. Each gesture creates a different kind of object. One of these gestures, called “rock”, creates a semitransparent square called “proxy”, which allows the remote control of more than one object, and also avoids the hands occlusion over the object.

However, the Rock & Rails technique still has a problem in WIMP interfaces. It happens because the user has to keep the non-dominant hand making the selected control gesture over the surface of the object. This may cause unwanted events or occlusion in elements of the application’s interface (Figure 2).

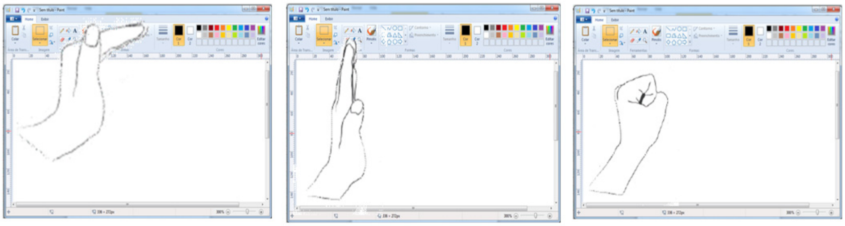


Fig. 2. Hands gesture proposed by Rock & Rails [2] technique and the occlusion problem with WIMP interfaces

In the present work, we decided to restrict the operations to the proxy, since it does not have objects within it and all its area can be used for manipulation. In the proposed technique, we implement the handles technique in the proxy. Therefore, the users may use the proxy when the object connected to it requires a more precise manipulation. The object can also be normally translated with the “no restrictions technique” without affecting the proxy; the user may use the proxy when needed (Figure 3).

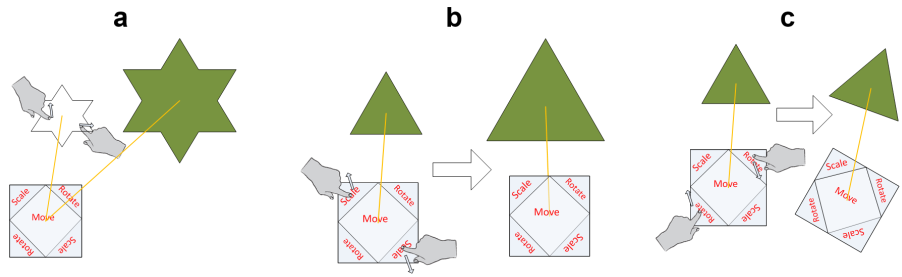


Fig. 3. Proposed technique “with the help of a proxy”. a) Translation can be done using or not the proxy. b) Scale and c) Rotation events only can be done using the proxy.

4 Evaluation of the Proposed Techniques

In this work we are going to compare three manipulation techniques for multitouch surfaces, the two techniques we propose (borders outside the object and with the help of a proxy) and the no restriction technique, which is the “standard technique”, according to Nacenta et al. [1]. To evaluate the techniques, we developed a test application with two scenarios. The first one was developed to evaluate the manipulation of a simple object, in this case, an image (Figure 4a). The second scenario has a WIMP interface to be manipulated, where other events may be triggered during manipulation (Figure 4b).

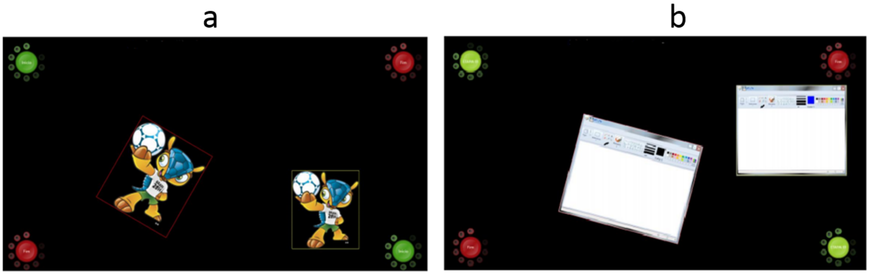


Fig. 4. Screen of scenarios a) an image and b) a WIMP interface

In the second scenario, we developed a window similar to that of the Paint program. In this scenario we want to evaluate the influence of occlusion in the spatial transformations using the proposed techniques. In this Paint-like window, we implemented some of the main characteristics of the program: the color palette (Figure 5a), the pen and eraser buttons (Figure 5b). The drawing area (Figure 5c) and buttons on the upper left corner: minimize, maximize and close (Figure 5d) also trigger events. For example, when one of the buttons is selected by the user touch, the window changes its position to the upper left corner of the screen, showing that an event happened. The application was designed to prevent that this window disappears in case of accidental touches on the close button. The adequate areas for manipulation are highlighted in Figure 5 with red borders.

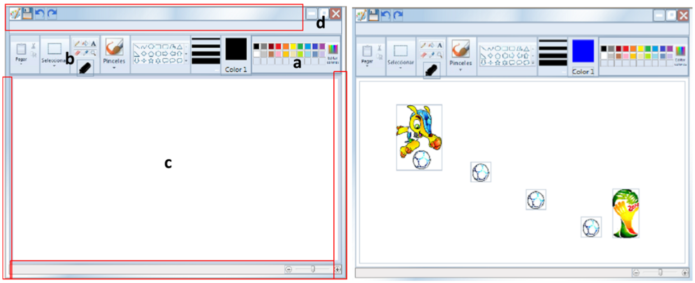


Fig. 5. Paint window style implemented to our test application

In the Rock & Rails technique [2], the user had to use a specific gesture to call the proxy. In our application, the proxy is called when the user performs a tap on the button named "Proxy" showed in Figure 6b. To call the menu, the user must execute the double tap (Figure 6a) anywhere on the table surface.

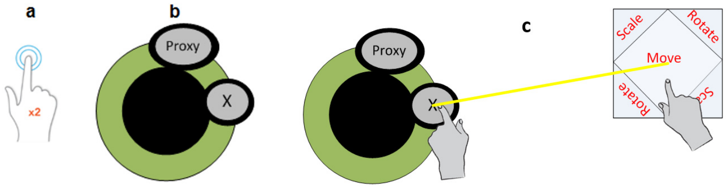


Fig. 6. - a) Double tap gesture, b) Menu interface showed after double tap gesture, c) The proxy and visual connection with the menu interface.

To connect an object to the proxy, the user must make a first tap on the proxy and a second tap on the object, when a line is drawn between the two objects, indicating the connection between them. Then, proxy movements are reflected on the object. The user removes the proxy from the surface through a connection made between the proxy windows and the "X" button on menu content (Figure 6c).

4.1 Test Application

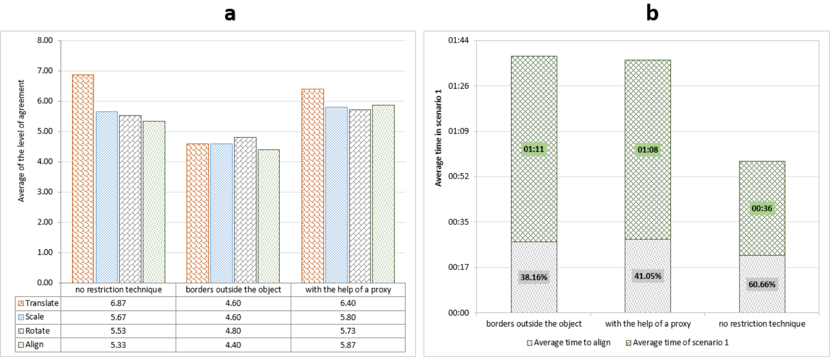
The tests implemented in this work had the objective to evaluate users' performance to complete certain tasks that require spatial transformations on 2D objects. The tests ran on a Microsoft PixelSense SUR40 multitouch table. A group of 20 users participated; evaluating three techniques in two scenarios.

In each scenario the application showed two objects, one object can be manipulated and the other one is static. The user had to align the movable object into another using spatial transformations. In the first scenario, the manipulated object was a simple picture. In the second scenario, the object was a WIMP window. In this second scenario an additional step was defined. After both objects are aligned, the user had to join six images (soccer balls) using dashes. The six images were presented in the Paint window since the beginning of the test (Figure 5). In case the user has drawn something with his/her touches in the drawing area of the Paint window during the window manipulation, he/she would have to first erase these drawings, and then draw the line joining the balls. The purpose of this second step was to reinforce the notion that the manipulation of a WIMP interface might generate unwanted events.

At the end of each scenario, the user had to fill out a questionnaire based on Likert scale where each item had 7 levels of agreement, where 1 means "strongly disagree" and 7, "completely agree". For data analysis, we define the following variables that allow measuring the performance of each technique: rotation, translation and scale errors, execution time, and time spent by occlusion.

4.2 Analysis of Scenario 1

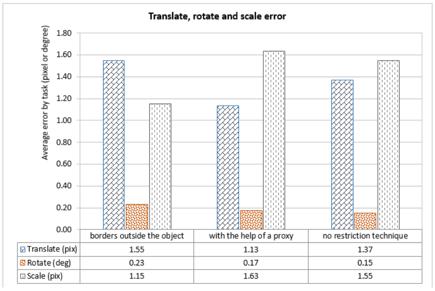
Figure 7a shows the result of the users' opinion about the ease of use of each technique in spatial transformations like rotation, translation, scaling, and alignment operation in scenario 1. One may observe that the technique with borders outside the object was considered the most difficult, compared to the other two techniques.



	no restriction technique	borders outside the object	with the help of a proxy
▨ Translate	6.87	4.60	6.40
▨ Scale	5.67	4.60	5.80
▨ Rotate	5.53	4.80	5.73
▨ Align	5.33	4.40	5.87

Fig. 7. a) Results of users' opinion considering each technique in translation, rotation, scaling and alignment task. b) Average time to complete the test using the proposed techniques.

Figure 7b shows that the technique with borders outside the objects spent more global time to finish the proposed tasks, compared to the other techniques. However, the figure shows that users spent less time, proportionally to the global time, to do the final alignment task (38.16% of total time). Similarly, despite the no restriction technique presented the lowest mean time to accomplish the task, it was the technique where users spent a larger proportion of time in the alignment task (60.66% of total time).



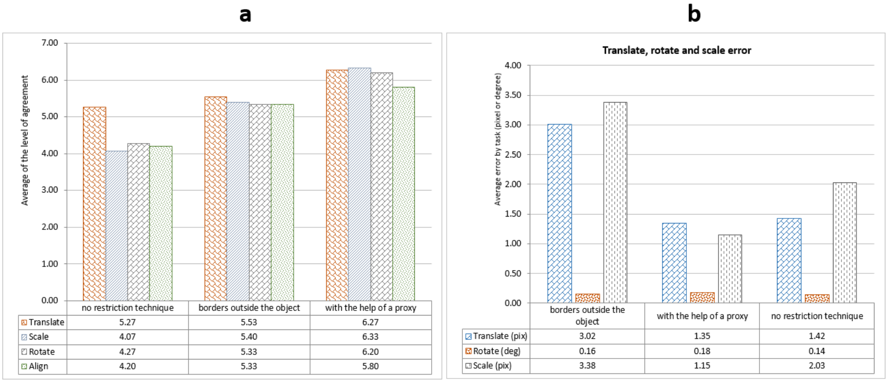
	borders outside the object	with the help of a proxy	no restriction technique
▨ Translate (pix)	1.55	1.13	1.37
▨ Rotate (deg)	0.23	0.17	0.15
▨ Scale (pix)	1.15	1.63	1.55

Fig. 8. Average error per task (translation, rotation and scaling)

In Figure 8, we observe the average errors of manipulation tasks. The technique with the help of a proxy presented more error in the scale operation. With respect to translation and rotation, the technique with borders outside the object was the less accurate. These errors measurements reinforce the users' opinion (Figure 8a) with respect to the technique with borders outside the object. Some users said that the view of the object was hampered by the borders when they try to align objects.

4.3 Analysis of Scenario 2

Figure 9a shows the users' opinion about the ease of use of each technique when working with a WIMP window. We observe that in this scenario the no restrictions technique was considered the most difficult. It had the lowest score in all types of spatial transformations. The technique with the help of a proxy obtained users' preference.



	no restriction technique	borders outside the object	with the help of a proxy
▨ Translate	5.27	5.53	6.27
▨ Scale	4.07	5.40	6.33
▨ Rotate	4.27	5.33	6.20
▨ Align	4.20	5.33	5.80

	borders outside the object	with the help of a proxy	no restriction technique
▨ Translate (pix)	3.02	1.35	1.42
▨ Rotate (deg)	0.16	0.18	0.14
▨ Scale (pix)	3.38	1.15	2.03

Fig. 9. a) Results of level of agreement of users considering each technique in translation, rotation, scaling and alignment task. b) Average error per task.

We may see a change in users' opinion, related to the use of two fingers to perform rotation and scale transformations. Users now prefer our proposed techniques that provide a wider space for object manipulation, without unexpected events executed by buttons and the drawing area of the Paint-like window interface.

Figure 9b shows the errors with respect to position, scale and rotation of the three techniques. We can see that the technique that got the lowest errors was the technique with the help of a proxy, while the technique with borders outside the object obtained the highest error values for translation and scale.

The second stage of scenario 2 was designed to evaluate how much the occlusion influenced the spatial transformations of the object. In this step, the user was asked to first remove any draft that has been drawn in step one (during spatial manipulation), and then make a trace to join the four balls. The time spent clearing the draft is the measure that we use to indicate the influence of occlusion in each technique. In Figure 10, we observe the average time measured to the occlusion problem. The technique with borders outside the object had the lowest occlusion time and the no restriction technique was the one with a longer duration. We can then infer that this last technique was the most affected by the occlusion.

The technique with borders outside the object was less affected by occlusion because the borders act as a kind of "protection" to prevent users to touch in controls inside the window. The user manipulates only the borders to align the window, for this reason, their exposure to occlusion error was minor compared with the other techniques.

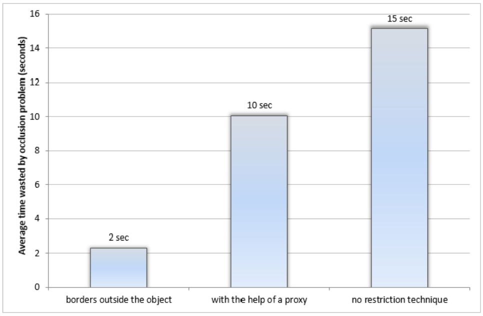


Fig. 10. Average time spent by users to fix problems derived from occlusion

5 Conclusion

In this work, we proposed and evaluated two techniques of separability for virtual objects in 2D multi-touch interfaces. These techniques were evaluated with user tests that found that the two proposed techniques improve separability and reduce occlusion in spatial transformations of simple objects and objects that contain elements of WIMP interfaces.

In relation to the accuracy of each technique, our results suggest that the technique with the help of a proxy improves the separability in both object types evaluated. However, in the evaluation of occlusion interference, it achieved an average result compared to the two other techniques. The technique with borders outside the object has similar gains in separability issue and better performance to reduce the occlusion in objects with WIMP interface. We found that the technique with borders outside the object has a better support when separability and occlusion appear together, especially when we need a better support for reuse of WIMP interfaces.

Our results also indicated the two proposed techniques spend less time in the operation of fine fitting of objects compared with the no restriction technique. Based on these results, we may indicate that the separability is a good strategy for avoiding the time spent in alignment movements and hence the users' fatigue.

Acknowledgments. Alberto Raposo thanks CNPq (process #310607/2013-2) for the support to his research work. This work is also supported by FAPERJ (process #190.243/2013).

References

1. Nacenta, M., Baudish, P., Banko, H., Wilson, A.: Separability of spatial manipulations in multi-touch interfaces, pp. 175–182 (2009)
2. Wigdor, D., Benko, J., Pella, J., Lomabardo, J., Williams, S.: Rock & rails: Extending multi-touch interactions with shape gestures to enable precise spatial manipulations. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 1581–1590 (2011)

3. Wu, M., Balakrishnan, R.: Multi-Finger and Whole Hand Gestural Interaction Techniques for Multi-User Tabletop Displays. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 193–202 (2003)
4. Dietz, P.: DiamondTouch: A Multi-User Touch Technology. In: Proceedings of UIST, pp. 219–226 (2001)
5. Kruger, R., Carpendale, S., Scott, S.D., Tang, A.: Fluid integration of rotation and translation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 601–610 (2005)
6. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1083–1092 (2009)
7. Moscovich., T., Hughes, J.F.: Multi-finger Cursor Techniques. In: Proceedings of Graphics Interface, pp. 1–7 (2006)
8. Hancock, M., Ten Cate, T., Carpendale, S.: Sticky tools: Full 6DOF force-based interaction for multi-touch tables. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, pp. 133–140 (2009)
9. Ashtiani., B., Stuerzlinger, W.: 2D similarity transformations on multi-touch surfaces. In: Proceedings of Graphics Interface, pp. 57–64 (2011)
10. Vogel, D., Casiez, G.: Hand Occlusion on a Multi-Touch Tabletop. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2307–2316 (2012)
11. Vogel, D., Baudisch, P.: Shift: A technique for operating pen-based interfaces using touch. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 657–666 (2007)
12. Roudaut, A., Huot, S., Lecolinet, E.: TapTap and MagStick: Improving One-Handed Target Acquisition on Small Touch-screens. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 146–153 (2008)
13. Brandl, P., Leitner, J., Seifried, T., Haller, M., Doray, B., To, P.: Occlusion-aware menu design for digital tabletops. In: Proceeding of CHI 2009 Extended Abstracts on Human Factors in Computing Systems, pp. 3223–3228 (2009)
14. Potter, R.L., Weldon, L.J., Shneiderman, B.: Improving the Accuracy of Touchscreens: An Experimental Evaluation of Three Strategies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 27–32 (1988)
15. Apted, T., Kay, J., Quigley, A.: Tabletop sharing of digital photographs for the elderly. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 781–790 (2006)
16. Wigdor, D., Williams, S., Cronin, M., Levy, R., White, K., Mazeev, M., Benko, H.: Ripples: Utilizing per-contact visualizations to improve user interaction with touch displays. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 3–12 (2009)
17. Wang, F., Ren, X.: Empirical Evaluation for Finger Input Properties in Multi-touch Interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1063–1072 (2009)

Human Activity Recognition from Kinect Captured Data Using Stick Model

Vempada Ramu Reddy and Tanushyam Chattopadhyay

TCS, Innovation Labs,
Kolkata, West Bengal, India
{ramu.vempada,t.chattopadhyay}@tcs.com

Abstract. In this paper authors have presented a method to recognize basic human activities such as sitting, walking, laying, and standing in real time using simple features to accomplish a bigger goal of developing an elderly people health monitoring system using Kinect. We have used the skeleton joint positions obtained from the software development kit (SDK) of Microsoft as the input for the system. We have evaluated our proposed system against our own data set as well as on a subset of the MSR 3Ddaily activity data set and observed that our proposed method out performs state-of-the-art methods.

Keywords: Human activity, Human action, Kinect, Skeleton, Activity recognition.

1 Introduction

Now a days there is a great demand for elderly people health monitoring systems as elderly people are gradually increasing since last decade [1]. A demographic revolution is underway throughout the world. Today, world-wide, there are around 600 million persons aged 60 years and over; this total will double by 2025 and will reach virtually two billion by 2050 - the vast majority of them in the developing world. According to world health organization (WHO), the count of elderly people may reach to 2000 million by the year 2050 [2]. The statistics of elderly people from the year 2002 to 2050 predicted by WHO is shown in Fig. 1 [2]. Therefore, proper caring is very much needed to monitor their health by identifying their daily activities. A lot of systems were developed for monitoring the activities of people. Broadly all the systems are classified in to two categories, namely (i) ubiquitous sensor based and (ii) computer vision based approaches [3] [4]. But the problem with ubiquitous sensors is obtrusive and some sensors are invasive, so elderly people shows no or less interest in using them. Therefore, some researchers developed mobile phone based activity recognition. But the main problem comes with if the user can forget to carry mobile phone for some activities like going to bed, watching TV in drawing room, toileting etc. In such cases computer vision based approaches proved to be best one. The comprehensive survey on the human activity recognition in [3] concludes the requirement of depth sensor to make a robust computer vision based system for

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 305–315, 2014.

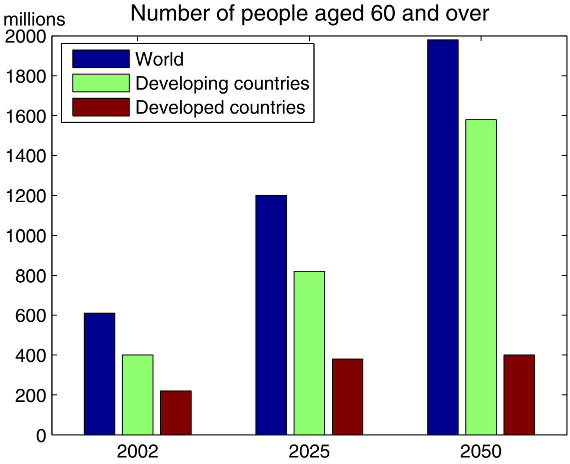


Fig. 1. Statistics of elderly people

human activity recognition. With the arrival of Kinect from Microsoft creates a new way of research in computer vision by mounting RGB camera with low cost depth sensor. Hence, we are using Kinect to monitor the activities in the bed room when the person may not carry the mobile phone. However, privacy is the major concern for monitoring bed room activities using computer vision based approach. But the advantage of deploying a Kinect based system is the availability of skeleton joint positions by using any stick model like Microsoft SDK so that the concern of privacy does not arise. We are using the floor map with the basic activities performed by the person in the bedroom recognized from the Kinect data to recognize the activities like going to toilet, coming back from toilet, going to bed, wake up from bed. But in this paper we are going to describe only our proposed method for activity recognition using the stick model obtained from the Kinect.

Rest of the paper is organized as follows: Following section presents the state-of-the-art works carried out in recognizing the activities. The data sets used for carrying out the experiments are explained in the section 3. The details of human activity recognition systems developed using different approaches is presented in the Section 4. The details of support vector machines for classifying the human activities is given in Section 5. Section 6 discuss the results of different approaches. Final section leads to summary and conclusions of the paper.

2 Prior Arts

Home activity monitoring is an interesting research topic for a long period. One such work in recent past can be found in [5]. Similarly research on human body model estimation using voxel data was started long back in the early of this century as [6]. But the onset of Kinect, the Microsoft gaming platform, [7] facilitate the access of 3-D data at a lower cost. Kinect was initially used for gesture recognition to make the user feeling more comfortable while playing games [8]. Kinect can sense the Red-Green-Blue (RGB) color value of the pixels as well as the depth (D) value using an infra red sensor within the device. Software tool kits (SDK) and some open source codes are also available to obtain a stick model of skeleton points from the RGB-D data. Human activity recognition problems are solved by taking the RGB-D or the skeleton as input. In this work we focused on monitoring the human activities by using only skeleton information obtained from Kinect. The details of some of the existing works on human activity monitoring using skeleton information is given below.

In [9], on-board mobile robot is used and position and velocity of robot is used for predicting the human motion and position relative to robot. Spine is chosen as representative point. For estimating the relative position Kalman filter is used . Circular path and zigzag motion of human is considered for experimentation.

In [10], novel features such as local occupancy pattern (LOP) feature was used based on the depth data and the estimated 3D joint positions. In addition, new temporal pattern representation called Fourier Temporal Pyramid is used to represent the temporal structure of an individual joint in an action. The features used are robust to noise, invariant to translational and temporal misalignment, and capable of characterizing both the human motion and the human object interactions. An action-let ensemble model is learnt to represent each action and to capture the intra-class variance. An actionlet is a particular conjunction of features for a subset of the joints, indicating a structure of the features. As there are an enormous number of possible actionlets, novel data mining solution to discover discriminative actionlets. Discriminative weights are learnt by kernel method.

In [11], only 10 joints of skeleton such as Head, ShoulderCenter, Spine, HipCenter, HipLeft, HipRight, KneeLeft, KneeRight, AnkleLeft, AnkleRight are considered and 4 types of features are extracted for human posture recognition in the context of a heath monitoring framework. 7 different experiments were carried out with the coordinates and different angles formed within these ten joints. With and without scaling the features is carried out. Support vector machine (SVM) was used for classifying the activities such as standing, sitting, bending and laying.

In [12], eigen joints features such as 3D position differences of joints to characterize action information including posture feature fcc, motion feature fcp, and offset feature fci in each frame and then concatenated the three features. Naive-Bayes-Nearest-Neighbor (NBNN) classifier is used for multi-class action classification.

In [13], Sequence of the Most Informative Joints (SMIJ) is used. Different sets of joints reveal discriminative information about the underlying structure of the action. At each time instant, automatically selects a few skeletal joints that are deemed to be the most informative for performing the current action. The selection of joints is based on highly interpretable measures such as the mean or variance of joint angles, maximum angular velocity of joints, etc. Histograms of Most Informative Joints (HMIJ), Histogram-of- MotionWords (HMW) and

Linear Dynamical System Parameters (LDSP), are used to demonstrate the power of the SMIJ features in terms of discriminability and interpretability for human action recognition. the quality of different features are evaluated using 1-nearest neighbor (1-NN) and support vector machine (SVM). Levenshtein distance is used for classification based on SMIJ. 2 distance is used for classification based on histogram feature representations HMIJ and HMW. Martin distance is used as a metric between dynamical systems for classification based on LDSP. One-vs-one classification scheme with Gaussian kernel is used.

In [14], histograms of 3D joint locations (HOJ3D) from 12 informative joints are used for compact representation of postures. The 12 joints includes head, L/ R elbow, L/ R hands, L/ R knee, L/ R feet, hip center and L/ R hip. Hip center was taken as the center of the reference coordinate system, and define the x-direction according to L/ R hip. The rest 9 joints are used to compute the 3D spatial histogram. Linear discriminant analysis (LDA) is performed to extract the dominant features. 3D skeletal joint locations are extracted from Kinect depth maps using Shotton method. Using LDA, the computed HOJ3D from the action depth sequences are reprojected. Later they clustered into k posture visual words, which represent the prototypical poses of actions. Discrete Hidden Markov models (HMMs) are used to model the temporal evolutions of visual words. It demonstrates significant view invariance on 3D action data set based on the design of spherical coordinate system and the robust 3D skeleton estimation from Kinect.

3 Activity Database

We evaluated the performance of our proposed method and state of the art methods on two different human activity data sets of 3D skeleton data. One data set is our own data set and second data set is standard MSR Daily Activity 3D data set. The description of each data set is given below.

Data set #1: The data set used for activity recognition is collected using Microsoft Kinect. The Kinect camera is fixed on the wall and recorded the activities. The sequence of different activities such as sitting, walking, laying and standing are recorded by covering all possible combinations or variations. The data is collected from 10 subjects consists of 5 male and 5 female. The subjects are within the age group of 23-40. For each subject we collected 2 repetitions of each action. Each repetition is about 90 sec duration, yielding a total of 120 min (90sec × 2rep × 4activities × 10subjects) of data. The frame rate of Kinect sensor of skeleton data is 30. Therefore, total number of frames for each activity from all subjects obtained is $30 \times 60 \times 30 = 54000$ frames. The skeleton data collected for each activity from each subject is dumped into text file which contains the basic 20 skeleton joint positions (HipCenter, Spine ShoulderCenter, Head, ShoulderLeft,ElbowLeft, WristLeft, HandLeft, ShoulderRight, ElbowRight, WristRight, HandRight, HipLeft, KneeLeft, AnkleLeft, FootLeft, HipRight, KneeRight, AnkleRight, FootRight) with its X, Y and Z co-ordinates.

Data set #2: The methods used in this work for identifying the human activity is also tested on the standard MSR Daily Activity 3D data set consisting of the skeleton data obtained from a depth sensor similar to the Microsoft Kinect with 15 Hz. MSR DailyActivity3Ddataset is a daily activity data set captured by a Kinect device. There are 16 activity types: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down. If possible, each subject performs an activity in two different poses: sitting on sofa and standing. The total number of the activity samples is 320. Out of 16 activities we selected the subset of 4 activities such as sit still, lay down on sofa, walk and stand up(same activities which was mentioned earlier in Data set #1) for testing our methods.

4 Proposed Method

In the proposed method we have used the skeleton joints obtained from Kinect as the input. The Microsoft SDK we have used usually returns 20 skeleton joints with their x, y and z coordinates. It was reported in the literature [13] that all the joints are not required for activity recognition. They have reported 6 major joints to be most informative. So we have analyzed the skeleton joint points and observed that most of the joint points are very noisy irrespective of the sequence and some joints are not much affected from the noise. As per our analysis, Head, Shoulder- Center, Shoulder-Left and Shoulder-Right are the most reliable joints. Therefore, in this work we carried out the experimentation using these 4 joints. In this work we accomplish the task of human activity recognition in three steps namely (i) Feature extraction phase,(ii) Training phase, and (iii) Testing phase.

4.1 Feature Extraction

In this work we have explored three different features for recognizing human activities. This phase is the core part of any classification system as the performance of the system in terms of accuracy largely depends on set of features used. The details of the feature extraction methods are given below.

PCA Based Approach. In this method, simple X, Y, and Z co-ordinates of the skeleton joint positions of Head, ShoulderCenter, ShoulderLeft and ShoulderRight are considered for every 30 frames and then we applied principle component analysis (PCA). The steps involved in feature extraction is given below.

1. In this method, the X,Y, and Z coordinates of the corresponding 4 joints i.e., Head, ShoulderCenter, ShoulderLeft and ShoulderRight are extracted for every 30 frames.

2. The extracted points are arranged in the form of matrix $H_{12\times30}$ where rows represent the three coordinates of four joints and columns represent the frames.
3. The mean value for each row of matrix $H_{12\times30}$ is calculated i.e., $M_{12\times1}$ and subtracted the matrix M from H, resulting matrix $\tilde{H}_{12\times30}$.
4. Now the covariance of the matrix $\tilde{H}_{12\times30}$ is obtained by using $\tilde{H}_{12\times30} \times \tilde{H}^T_{12\times30}$. The resulting covariance matrix is $C_{12\times12}$, where each column represent the eigen vector. Among the 12 eigen vectors, top 7 eigen vectors are considered based on the top 7 eigen values. The resulting matrix now obtained is $P_{12\times7}$.
5. Original feature matrix $H_{12\times30}$ can be represented as $H_{12\times30} = \tilde{H}_{12\times30} + P_{12\times7} \times B_{7\times30}$, where $B_{7\times30}$ is the weight matrix.
6. The weight matrix is now obtained as $B_{7\times30} = P^{-1}_{7\times12}(H_{12\times30} - \tilde{H}_{12\times30})$
7. Now concatenate all the columns of $B_{7\times30}$ into single vector f_{PCA} forming 210 dimensional feature vector.
8. Follow steps 1 to 7 for each activity from each person data.

Statistical Features. From the collected data, we have analyzed that there is a lot of variation exists between the mean values of the X, Y and Z coordinates of the above mentioned four joints for each activity. The similar phenomenon is also observed between the difference of maximum and minimum values of X, Y and Z coordinates of the four joints for each activity. Hence in this work we have explored mean values and difference between maximum and minimum values of the joint positions of the four joints as features for identification of activities such as sitting, walking, laying and standing. Let F_{mean} be the feature vector which are the mean values of the 3 coordinates of 4 joints extracted for every 30 frames. Therefore the feature vector F_{mean} is represented by 12 features. Let $F_{max-min}$ be the feature vector which are the values of difference between maximum and minimum values of the X, Y and Z coordinates of 4 joints extracted for every 30 frames. Therefore the feature vector $F_{max-min}$is represented by 12 features. An example of discrimination of X, Y and Z co-ordinates of mean feature vector and the difference of maximum and minimum feature vector for the 4 joints for the corresponding 4 activities is shown in Fig. 2.

In this study, three human activity recognition systems(HARS) are developed which are summarized as follows:

1. HARS-1: Human activity recognition system using only mean values of the joint positions as features.
2. HARS-2: Human activity recognition system using only difference between maximum and minimum values of the joint positions as features.
3. HARS-3: Human activity recognition system using combination of mean values and difference between maximum and minimum values of the joint positions as features.

The experimentation is carried out using only mean values feature vector F_{mean}, only difference of maximum and minimum values feature vector $F_{max-min}$ and

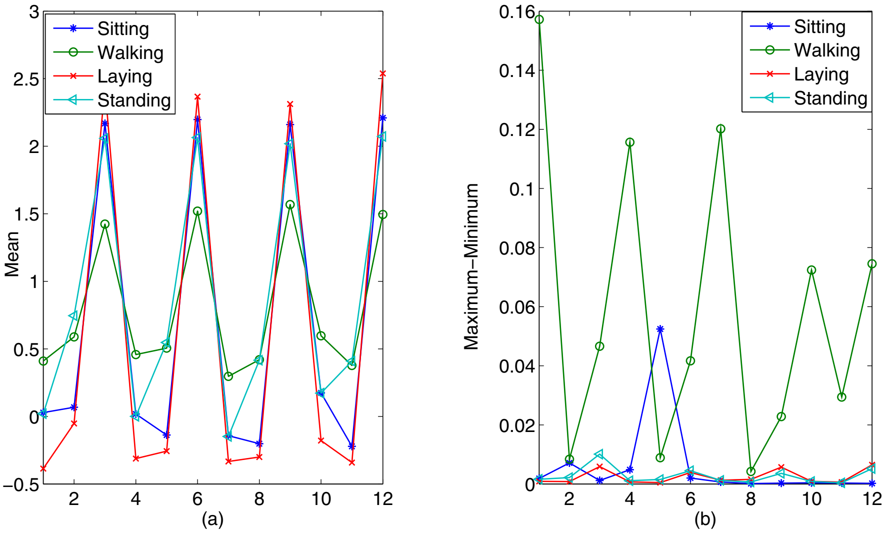


Fig. 2. Representation of feature vector (a) mean and (b) difference of maximum and minimum of the four activities sitting, walking, laying and standing

combination of both F=[F_{mean}, $F_{max-min}$]. It is observed the concatenated feature vector outperformed compared to individual feature vectors. This can be verified from the results given in the Section 6.

4.2 Training

For developing the efficient human activity recognition system, proper training need to be carried out using machine learning. In this work 5 fold cross validation is used, where 4 folds used for training and 1 fold for testing. The sequence of steps followed in training phase are given below.

1. Let A be a set of m activities (A = $a_1, a_2, \ldots, a_m$) available to us.
2. Let S be a set of annotated skeleton data available.
3. Let $\theta_k \in$ S be the set of data used for training phase, where θ_k is a subset of data of set S and θ_k is 80% of S.
4. Now use feature vectors extracted in the feature extraction phase and activity pairs as input to machine learning from the training data θ_k to generate the models.

4.3 Testing

For testing the human activity recognition systems developed using different features, we used 1 fold of data out of 5 folds. Testing is carried by using the models developed in training phase. The sequence of steps followed in testing phase are given below.

1. Let $\beta_k \in$ S be the set of data used for testing phase, where β_k is a subset of data of set S and β_k is 20% of S.
2. The data in the test set is not present in the train set i.e., $\theta_k \cap \beta_k = \phi$.
3. The same steps are followed for extraction of features from test set β_k of different methods which was mentioned in the previous subsection.
4. Now in the testing phase we provide only the features (f)as input to the system.
5. Now system predicts the activities based on the learning models developed in the training phase.
6. For each method experiment is run for 5 times such as out of 5 folds of data, each time 1 fold is used for testing and remaining 4 fold used for training.
7. Now the performance accuracy in terms of percentage is computed for each run in test phase as follows:

$$\%Accuracy = \frac{A_c}{A_t} \times 100$$

   where A_c is number of activities correctly classified out of total activities A_t
8. The overall average performance accuracy of the test data is calculated by taking the mean of all accuracies obtained in the 5 runs carried out in test phase.

5 Support Vector Machines

In this work, Support Vector Machines (SVM) are explored as a machine learning tool to discriminate the human activities. SVM classification is an example of supervised learning. SVMs are useful due to their wide applicability for classification tasks in many signal processing applications. A classification task usually involves training and testing data which consist of some data instances. In the training set, each instance contains one target class label and many attributes. The main goal of SVM for classification problem is to produce a model which predicts target class label of data instances in the testing set, given only the attributes. The SVM models for different human activities were developed as-one against-rest principle. The SVM model for the specific activity was developed, by using feature vectors derived from the desired human activity clues as positive examples and the feature vectors derived from the other human activities as negative examples. Radial basis function (RBF) kernel, unlike linear kernel, is used in this work to map the data points to higher dimensional space as it can handle the case where the relation between the class labels and attributes is nonlinear. The intuition to use RBF kernel function is due to its universal approximation properties. Also, it offers good generalization as well as good performance in solving practical problems [15]. The basic architecture of human activity classification system using SVMs with above mentioned features is shown in Fig. 3.

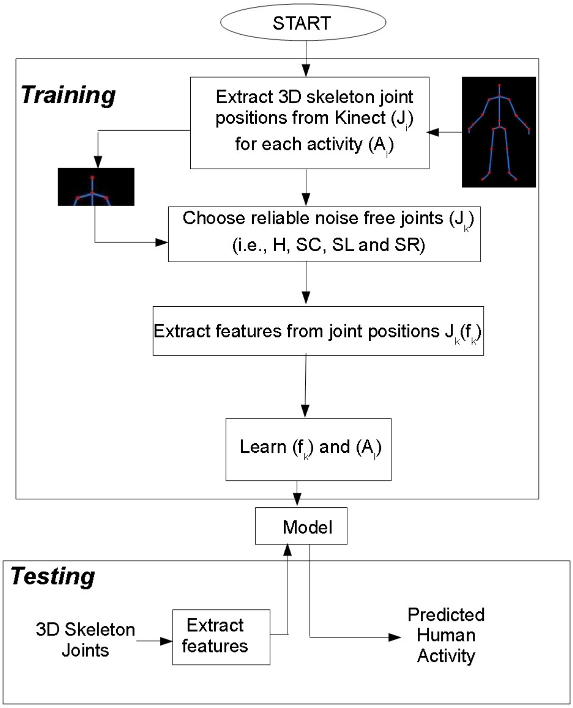

START
Training
Extract 3D skeleton joint positions from Kinect (J_i) for each activity (A_i)
Choose reliable noise free joints (J_k) (i.e., H, SC, SL and SR)
Extract features from joint positions J_k(f_k)
Learn (f_k) and (A_i)
Model
Testing
3D Skeleton Joints
Extract features
Predicted Human Activity

Fig. 3. Architecture of activity recognition system(H:Head, SC: Shoulder Center, SL:Shoulder Left and SR:Shoulder Right

6 Experimental Results

We have tested all the activity recognition methods which are mentioned above on Data set #1 and Data set #2. The features extracted from above mentioned methods are used as input for SVM and activities as labels to SVM. In this work, we used 5 fold cross validation and then the average performance is computed. The performance accuracy of the different methods mentioned above is given in Table 1. Columns 1 and 2 of Table 1 indicates the data sets used for evaluation of the methods (rows of Table 1). The values in Table 1 indicate the average performance accuracy of different methods. The performance of activity recognition using PCA features observed to be poor. The poor performance is mainly due to dependency on the distance and their absolute values. The performance of PCA based approach is improved by normalizing in between -1 and 1 and thereby making it uniform and independence of distance. Normalization reduce the intr-class variation under different test sets. The drastic improvement in the performance of activity recognition for without and with normalization of PCA features can be observed from the rows 1 and 2 of Table 1. From Table 1, it is observed that the performance accuracy of the method using only the mean values of the X, Y, and Z co-ordinates of skeleton data extracted for every 30 frames performed better compared to other individual methods. It is observed that for data set #2, the average performance of human activities using difference between maximum and minimum values seems to be poor compared to data set #1. This is mainly due to more randomness of data present in data set #2. But the combination of features such as mean values, and difference of maximum and minimum values outperformed compared to individual features for both data sets. The performance accuracy of human activities of different methods for data sets #1 and #2 is also plotted in Fig. 4.

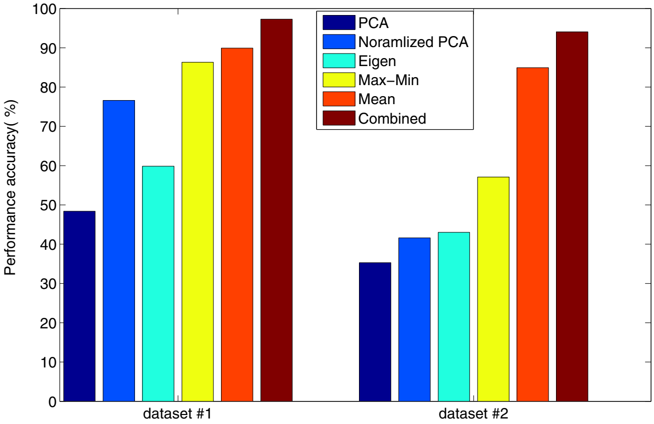


Fig. 4. Performance plot of human activities of different methods

Table 1. Performance of the different methods for identification of activities)

Sl. No	Features	**Average Classification Performance (%)**	
		Our dataset	**MSR 3Ddaily activity dataset**
1	PCA	48.40	35.30
2	Normalized PCA	76.60	41.62
3	EigenJoints	59.86	43.03
4	Maximum-Minimum(Range)	86.32	57.10
5	Mean	89.93	84.95
6	Combination 4 and 5 (Range and Mean)	97.29	94.06

7 Summary and Conclusions

In this paper we have presented a skeleton joint based method for activity recognition where we have used four major joint points that reduces the possibility of error due to noise as well as it reduces the over all time complexity of the system. We have also explored two different features for recognizing the activity. We have compared our features against eigen joint based approach and find that our method out performs that. Our system can recognize these basic activities up to 97.29% accuracy which is much better than the state of the art. Our system is not working in some sequences as our system is not using any noise cleaning method. We are working on incorporating any suitable de-noising method. We are now working to integrate it with our live system so that it can be deployed in homes to monitor the activities of the elderly people.

References

1. Cohen, J.E.: Human population: The next half century. Science 302(5648), 1172–1175 (2003)
2. `http://www.who.int/ageing/events/idop_rationale/en/`
3. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) 43(3) (April 2011)
4. Tapia, E.M.: Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors., Master degree Thesis, MIT (2003)
5. Cheng, H., Liu, Z., Zhao, Y., Ye, G.: Real world activity summary for senior home monitoring. In: IEEE International Conference on Multimedia and Expo (ICME), July 11-15, pp. 1,4 (2011)
6. Mikic, I.: Human Body Model Acquisition and Tracking Using Voxel Data. International Journal of Computer Vision 53(3), 199–223 (2003)
7. The teardown. Engineering Technology 6(3), 94-95 (April 2011)
8. Lepri, B., Salah, A.A., Pianesi, F., Pentland, A.S.: Human Behavior Understanding for Inducing Behavioral Change: Application Perspectives. In: Wichert, R., Van Laerhoven, K., Gelissen, J. (eds.) AmI 2011. CCIS, vol. 277, pp. 252–263. Springer, Heidelberg (2012)
9. Machida, E., Meifen, C., Murao, T., Hashimoto, H.: Human motion tracking of mobile robot with Kinect 3D sensor. In: Proceedings of SICE Annual Conference (SICE), Akita university, Akita, Japan, August 20-23, pp. 2207–2211 (2012)
10. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-21, pp. 1290–1297 (2012)
11. Le, T., Nguyen, M., Nguyen, T.: Human posture recognition using human skeleton provided by Kinect. In: Proceedings of International Conference on Computing, Management and Telecommunications (ComManTel), January 21-24, pp. 340–345 (2013)
12. Yang, X., Tian, Y.: EigenJoints-based Action Recognition Using Naive-Bayes-Nearest-Neighbor. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 16-21, pp. 14–19 (2012)
13. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Sequence of the Most Informative Joints (SMIJ): A New Representation for Human Skeletal Action Recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 16-21, pp. 8–13 (2012)
14. Xia, L., Chen, C.-C., Aggarwal, J.K.: View Invariant Human Action Recognition Using Histograms of 3D Joints. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 16-21, pp. 20–27 (2012)
15. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (2001)

Multi-sensor Based Gestures Recognition with a Smart Finger Ring

Mehran Roshandel[1], Aarti Munjal[2], Peyman Moghadam[3],
Shahin Tajik[1], and Hamed Ketabdar[4]

[1] Deutsche Telekom Innovations Laboratories, Ernst-Reuter Platz, 10587 Berlin Germany
{mehran.roshandel,shahin.tajik}@telekom.de
[2] Department of Biostatistics and Informatics, University of Colorado Denver,
13001 E 17th Pl Aurora CO 80045
aarti.munjal@ucdenver.edu
[3] Autonomous Systems, CSIRO Computational Informatics, 1 Technology Court,
Pullenvale, QLD 4069
peyman.moghadam@csiro.au
[4] Quality and Usability Lab, TU Berlin Deutsche Telekom Innovation Laboratories
Ernst-Reuter-Platz 7, 10587 Berlin Germany
Hamed.Ketabdar@clickandbuy.com

Abstract. Recently several optical and non-optical sensors based gesture recognition techniques have been developed to interact with computing devices. However, these techniques mostly suffer from problems such as occlusion and noise. In this work, we present Pingu, a multi-sensor based framework that is capable of recognizing simple, sharp, and tiny gestures without the problems mentioned above. Pingu has been calibrated in the form of a wearable finger ring, capable of interacting even when the device is not in the vicinity of the user. An advanced set of sensors, wireless connectivity, and feedback facilities enable Pingu for a wide range of potential applications, from novel gestures to social computing. In this paper, we present our results based on experiments conducted to explore Pingu's use as a general gestural interaction device. Our analysis, based on simple machine learning algorithms, shows that simple and sharp gestures performed by a finger can be detected with a high accuracy, thereby, stablishing Pingu as a wearable ring to control a smart environment effectively.

Keywords: Human Computer Interaction (HCI), Touch less gestural interaction, Wearable device, Finger ring.

1 Introduction

Gesture recognition is one of the important fields in Human Computer Interaction (HCI) as it enables users to interact with their computing devices more easily and naturally. In addition, use of multiple sensors for gestural recognition can extend the ability of accepting 3D inputs, which is not supported by conventional input devices.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 316–324, 2014.

Several applications, such as sign-language recognition, physical activity monitoring, and social interactions, can be developed based on simple gestures made by humans.

Several gesture recognition techniques based on optical and non-optical sensors were developed. Optical-based gesture recognition methods use optical sensors such as cameras [1, 2] or infrared (IR) sensors [3], to capture the movements of fingers and interpret them to commands. While the optical-based methods have problems such as occlusion, the non-optical methods try to use magnetometer [4, 5], accelerometer [7, 8], and proximity sensors [9] to overcome the limitations of optical gestural recognition techniques. However they have also their weaknesses, such as working in a limited space near the user device (MagiTact [4], MagiThings [17] and MagiSign [16]), accepting only 1D input (Nenya [5]), or in general they may be not socially acceptable.

In this work, we used our framework Pingu [6], a wearable and small finger ring that can interact with other electronic devices more naturally. While gestures made by any part of the body can be used for interacting with a computing device, previous research based on experiments conducted by Card et al. [10] shows that the information entropy of a finger-based interaction is much larger than the interaction based on any other human body parts. For instance, the interaction with the arm and wrist have the information rates of 11.5 and 25 bits/s respectively, while the information rate of finger is 40 bits/s. Therefore, Pingu is calibrated in the form of a finger ring with the following features:

1. It is composed of an advanced set of sensors (gyroscope, accelerometer and magnetometer).
2. It is equipped with wireless connectivity that makes it suitable for use in ubiquitous human-computer or human-human interaction.
3. The tiny size in the form of a finger ring makes Pingu wearable and, thus, socially acceptable.

Pingu is also capable of accepting 3D inputs from different gestures which are performed either in air or on surfaces such as a user's palm, top of the table. These gestures can be used in developing several interesting applications, including remote controlling or signature recognition. In this work, we analyze Pingu for recognizing a range of simple gestures. The main contribution of our work is to present results based on a multi-sensor interaction framework and effective classification algorithms. Our analysis shows that these generic gestures can be recognized with high accuracy.

The rest of this paper is organized as follows. In section 2 we review the related gesture recognition's solutions. Then in section 3 we explain the architecture of Pingu and its hardware. Experimentation and feature extraction via Pingu is discussed in section 4 and the results of gesture classification via different machine-learning algorithms are shown in the section 5. Finally we conclude the paper in the section 6.

2 Related Works

There are different gesture recognition approaches which have been developed in recent years and can be categorized into two groups: optical and non-optical gestural recognition techniques.

In optical-based gestural recognition approaches, optical sensors like cameras (e.g. SixthSense [1] and Gesture Pendant [2]) or infrared (IR) sensors (e.g. SideSight [3]) are the essential components to recognize the movements of fingertips and hands to interpret them to different commands. Although these approaches perform gesture recognition in some applications accurately, they do not support applications that are required to work with no direct line of sight (occlusion problem). Furthermore, optical data is sensitive to illumination conditions and, therefore, can only be used in certain circumstances. Finally, the user should wear additional cap or pendant which may be obtrusive and/or socially unacceptable.

On the other hand, non-optical gestural recognition methods use sensors such as magnetometer (e.g., MagiTact [4] and Nenya [5]), accelerometer [7, 8 and 11], and proximity sensors (e.g., Gesture Watch [9]) to mitigate the problems of optical-based methods.

Although proximity sensor solves the illumination problems, it still has the occlusion problem, as the gestures should be captured in the line-of-sight of sensors. Other methods based on accelerometer [7, 8, 11] do not have the occlusion and illumination problems, but since the acceleration data is very sensitive to noise, complementary sensors should be used. Techniques based on magnetometer send interaction commands when the magnetic field around the computing device is deformed. The advantage of this method is that there is no occlusion and illumination problem like previous approaches.

The gesture recognition techniques can also be categorized into types of wearable devices which they are embedded in. In some techniques, user should wear additional gloves such as Acceleration Sensing Glove [11] to interact with the computing device. The disadvantage of working with gloves is that they can be socially unacceptable or obtrusive. Other techniques like SixthSense [1] or Gesture Pendant [2] which require users to wear additional hat and pendant respectively, suffer from the same problems.

One possible solution is to develop the gestural recognizer as a ring or wristwatch, which may be socially more acceptable. Pinchwatch [12] is one of these systems which use a wristwatch for finger gesture recognition with the help of a camera. Users invoke functions by pinching and entering parameters by performing sliding and dialing motions. However, again this suffers from the problem of line of sight. Our previous work, MagiTact [4], involves interacting with a computing device equipped with an embedded compass (magnetic) sensor via a magnet placed on a finger. Coarse gestures made with the magnet affect the magnetic field around the device and, thus, used for gestural interaction. Although this approach has no occlusion problems, interaction is still limited to the immediate 3D space around the device.

More recently Nenya [5] a magnetically-tracked finger ring is developed which includes a permanent magnet in the form of a finger ring and worn-watch wireless tracking bracelet. While magnetometer is used to track the ring's position, a Bluetooth radio allows the bracelet to send ring input to the user's devices. Nenya supports only 1D input in comparison to Pingu which supports 3D input. Furthermore, it consists of two accessories in contrast to Pingu which includes all sensors and radio in only one ring. Magic Ring [13] is another finger-worn device which is developed for using static finger gestures and it uses accelerometer data to detect different gestures. Magic Ring is tested with six different finger gestures with doing some predefined task.

In our approach we classify nine finger gestures with four machine learning algorithms to derive the accuracy of gesture recognition.

3 Design

Using multiple sensors provides rich information related to motion and angular position of the device and, thus, results in a high accuracy in gesture recognition. Pingu is equipped with four sensors: accelerometer, magnetometer, gyroscope, and proximity sensor. A tri-axel accelerometer is used to detect the orientation and motion of the device along three axes x, y, and z. A tri-axel gyroscope detects the angular rate of movement of the ring along x, y, and z axis. Using gyroscope in addition to accelerometer provides six degree of freedom and can be used to detect the 3D trajectories of the ring. Additionally, the deformation of magnetic fields is useful in recognizing the coarse gestures made around the device. Moreover, Pingu has proximity-sensing plates installed, which allow sensing the proximity of other fingers. Figure 1 shows the prototype of Pingu and Table 1 lists the configuration of the sensors and radio which are used in the design of Pingu.

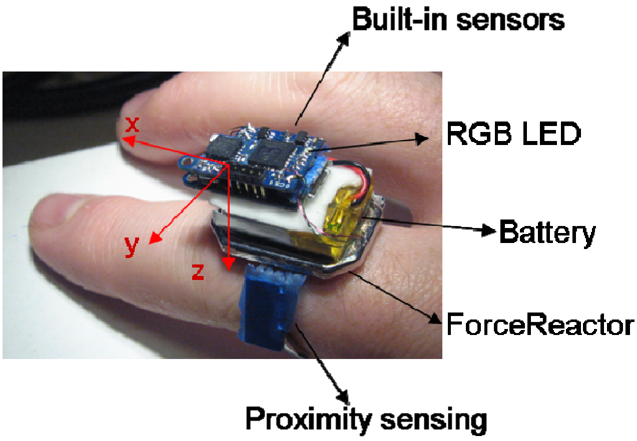


Fig. 1. Pingu, our multi-sensor framework, for interaction with a smart environment

Table 1. Sensors used in the design of Pingu and their specifications

Sensor	Description
Accelerometer	[-8g, 8g]
Magnetometer	[-2gauss, 2gauss]
Gyroscope	[-2000deg/s, 2000deg/s]
Bluetooth	Up to 2m

4 Experiment

In this work, we evaluate Pingu for general gestural interaction. For this purpose, we have defined a set of nine general gestures shown in Figure 2. As shown, the gestures are highly general in nature and can be used to control smart environments. For example, gesture 1 and 2 can substitute the volume control buttons on a remote control and gesture 5 and 6 can change music tracks forward and backward respectively. To evaluate Pingu for general gestural interaction in smart environments, we perform all the nine gestures in three ways, as follows:

1. General Gestures in the air, in which a gesture is performed in the air.
2. General Gestures on the table, in which a gesture is performed on the top of the table.
3. General Gestures on the palm, in which a gesture is performed on user's palm.

Setting the three medium of air, palm and desk provides a variety of surfaces for gesturing. In this way, the methodology can be tested under more variable yet practical scenarios. Both palm and desk are surfaces which are commonly available for users during the gesturing process. The air medium also provides the fantasy of writing in air for the user, when the two other mediums are not available.

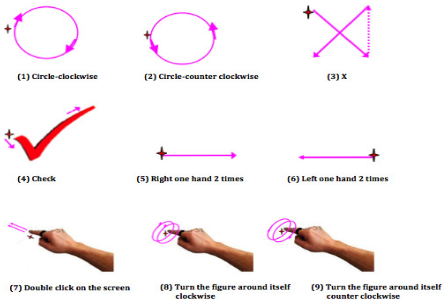

(1) Circle-clockwise
(2) Circle-counter clockwise
(3) X
(4) Check
(5) Right one hand 2 times
(6) Left one hand 2 times
(7) Double click on the screen
(8) Turn the figure around itself clockwise
(9) Turn the figure around itself counter clockwise

Fig. 2. A set of nine general gestures used in this work

Our results are based on a dataset collected from 24 users.

Table 2. User statistic Table

Total Users	Male	Female	Right Handed	Left Handed
24	10	14	20	4

Feature ExtractionEach of the nine gestures shown in Figure 2 is performed 15 times per user. The sensor readings specific to each of the nine gestures are then captured via a Java desktop application developed for Mac OS. To evaluate the interaction made by Pingu, we classify gestures based on the sensor readings collected for each gesture. In particular, we adopt the following approach:

We mix the data collected from all the 24 users and cross-validate. For this purpose, we form a feature vector containing data specific to each sensor. For example, a feature vector obtained from the accelerometer used in Pingu contains the following:

1. Mean of the linear acceleration along x, y, and z axis (3 features),
2. Variance of the linear acceleration along x, y, and z axis (3 features),
3. Mean of the Euclidian norm of the linear acceleration along x, y, and z axis (1 feature),
4. Variance of the Euclidian norm of the linear acceleration along x, y, and z axis (1 feature),
5. Standard Deviation of the linear acceleration along x, y, and z axis (3 features),
6. Piecewise correlation between linear acceleration along x, y, and z axis (3 features), and
7. Frequency features along x, y, and z axis (3 features).

As shown in Figure 2, mean and variance of the sensor readings obtained for gestures 1 and 2 may not be able to differentiate between these two gestures. Therefore, we have included the piece-wise correlation and the frequency features specific to each sensor. The feature vector for the angular rate movement of the ring (i.e., from gyroscope) is obtained in a similar manner. Feature vector for each sensor, therefore, contains 17 elements. Since multiple windows provide more detailed information in gesture classification, our results are based on 4 windows. Feature vectors obtained from each window are concatenated to form a new feature vector of 68 (=17×4) features. To further validate that Pingu is effective in gestural interaction in a smart environment, we do not use the magnetometer readings in this analysis.

5 Gesture Classification

The feature vectors obtained for each of the three experiments are then used as an input to a classification algorithm for gesture classification. Specifically, we have classified the gestures with four classifiers: (a) Decision Tree (DT), a decision tool that uses graphs and model of decisions to derive the outcomes and consequences, (b) Multi-Layer Perceptron (MLP), a feed forward artificial neural network that models the relationship of inputs and outputs to find the patterns, (c) Naïve Bayes (NB), a probabilistic classifier that uses Bayes' theorem with strong independence assumptions, and (d) Support Vector Machines (SVM), a set of hyperplanes in high dimensional space for using classification and regression . Our analysis is based on the implementation of these classifiers in the Weka machine learning toolkit [14, 15]. Tables 2-4 list the classification results obtained for all the three experiments. As shown in Tables 2 and 3, MLP classifier outperforms all the other three classifiers for gestures performed in air and on table, with more than 97% and 93% accuracy, respectively. Table 4 shows that SVM has better results with more than 77% accuracy. To illustrate further how accurate different gestures can be distinguishable, confusion matrices obtained from MLP classifier for General Gestures performed in air are shown in Table 5. As shown, all nine gestures are most of the times distinguishable, but more specifically the confusion matrices indicate that gesture 1 and gesture 2 are somewhat more difficult to classify. Similarly, we note that gesture 8 and gesture 9 are classified with lower accuracy due to the inherent similarity in performing these gestures. On the other hand, gesture 4 is the easiest to be classified, as it's easily distinguishable s from other gestures. These results show that generally, the gesture recognition by Pingu is trustworthy.

Table 3. Gesture Classification Results for General Gestures in the air

Algorithm	Accuracy
MLP	**97.879%**
DT	87.043%
NB	57.389%
SVM	96.443%

Table 4. Gesture Classification Results for General Gestures on the table

Algorithm	Accuracy
MLP	**93.689%**
DT	73.080%
NB	49.238%
SVM	83.991%

Table 5. Gesture Classification Results for General Gestures on the palm

Algorithm	Accuracy
MLP	71.130%
DT	71.160%
NB	33.415%
SVM	77.541%

Table 6. Confusion matrix obtained from MLP for the results shown in Table 3

Gesture	1	2	3	4	5	6	7	8	9
1	333	3	0	0	0	1	0	3	0
2	3	334	2	0	0	0	0	0	1
3	2	1	336	0	1	0	0	0	0
4	1	1	0	343	0	0	1	1	0
5	0	0	0	1	338	2	0	0	0
6	1	0	0	0	3	324	0	0	0
7	0	1	0	0	0	0	340	2	0
8	7	0	1	0	0	0	0	320	12
9	0	0	0	0	0	0	0	14	332

6 Conclusions and Future Work

In this work, we have presented our results for gestural recognition using a multi-sensor based framework called Pingu. Our results are based on a set of nine pre-defined general gestures that can be used to interact in a smart environment. Pingu is a socially wearable, small finger ring that is equipped with multiple sensors to provide rich information about the general gestures made by a user. Our analysis is based on a large dataset of 24 users. We have shown that with simple classification algorithms, different gestures can be distinguished from each other with high accuracy. Therefore, we can trust Pingu to be involved in many interesting applications such as remote controlling, signature recognition, physical activity analysis and sign-language recognition.

References

1. Mistry, P., Maes, P.: SixthSense: A wearable gestural interface. In: ACM SIGGRAPH ASIA 2009 Sketches. ACM (2009)
2. Starner, T., et al.: The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In: The Fourth International Symposium on Wearable Computers. IEEE (2000)
3. Butler, A., Izadi, S., Hodges, S.: SideSight: Multi- “touch” interaction around small devices. In: Proc. UIST, pp. 201–204 (2008)
4. Ketabdar, H., Yüksel, K.A., Roshandel, M.: MagiTact: Interaction with mobile devices based on compass (magnetic) sensor. In: Proceedings of the 15th International Conference on Intelligent User Interfaces. ACM (2010)
5. Ashbrook, D., Baudisch, P., White, S.: Nenya: Subtle and eyes-free mobile input with a magnetically-tracked finger ring. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems. ACM (2011)
6. Ketabdar, H., Moghadam, P., Roshandel, M.: Pingu: A new miniature wearable device for ubiquitous computing environments. In: 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS). IEEE (2012)
7. Wu, J., Pan, G., Zhang, D., Qi, G., Li, S.: Gesture recognition with a 3-d accelerometer. In: Zhang, D., Portmann, M., Tan, A.-H., Indulska, J. (eds.) UIC 2009. LNCS, vol. 5585, pp. 25–38. Springer, Heidelberg (2009)
8. Fukumoto, M., Tonomura, Y.: “Body coupled FingerRing”: wireless wearable keyboard. In: Proceedings of the SIGCHI Conference on Human Factors in Computing systems. ACM (1997)
9. Kim, J., et al.: The gesture watch: A wireless contact-free gesture based wrist interface. In: 2007 11th IEEE International Symposium on Wearable Computers. IEEE (2007)
10. Card, S.K., Mackinlay, J.D., Robertson, G.G.: A morphological analysis of the design space of input devices. ACM Trans. Inf. Syst. 9(2), 99–122 (1991)
11. Perng, J.K., Fisher, B., Hollar, S., Pister, K.S.J.: Acceleration sensing glove (ASG). In: The Third International Symposium on Wearable Computers (ISWC 1999), pp. 178–180 (1999)
12. Loclair, C., Gustafson, S., Baudisch, P.: PinchWatch: A wearable device for one-handed microinteractions. In: Proc. MobileHCI (2010)

13. Jing, L., et al.: Magic Ring: A Finger-worn device for multiple appliances control using static finger gestures. Sensors 12(5), 5775–5790 (2012)
14. Weka3: Data Mining Software in Java, `http://www.cs.waikato.ac.nz/ml/weka/`
15. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2011)
16. Ketabdar, H., Moghadam, P., Naderi, B., Roshandel, M.: Magnetic signatures in air for mobile devices. In: Mobile HCI 2012, pp. 185–188 (2012)
17. Ketabdar, H., Abolhassani, A.H., Roshandel, M.: MagiThings: Gestural Interaction with Mobile Devices Based on Using Embedded Compass (Magnetic Field) Sensor. IJMHCI 5(3), 23–41 (2013)

View-Invariant Human Detection from RGB-D Data of Kinect Using Continuous Hidden Markov Model

Sangheeta Roy and Tanushyam Chattopadhyay

TCS, Innovation Labs,
Kolkata, India
{roy.sangheeta,t.chattopadhyay}@tcs.com

Abstract. In this paper authors have presented a method to detect human from a Kinect captured Gray-Depth (G-D) using Continuous Hidden Markov models (C-HMMs). In our proposed approach, we initially generate multiple gray scale images from a single gray scale image/ video frame based on their depth connectivity. Thus, we initially segment the G image using depth information and then relevant components were extracted. These components were further filtered out and features were extracted from the candidate components only. Here a robust feature named Local gradients histogram(LGH) is used to detect human from G-D video. We have evaluated our system against the data set published by LIRIS in ICPR 2012 and on our own data set captured in our lab. We have observed that our proposed method can detect human from this data-set with a 94.25% accuracy.

1 Introduction

Human activity detection for indoor and outdoor surveillance has a major research interest since last two decades. Activity detection is very effective in human based application like video indexing and retrieval, intelligent human machine interaction, video surveillance, health care, driver assistance, automatic activity detection and predicting person behavior. Some of such applications can be found in literature like [1] (in office), [2], [3], [4] (in retail stores), and in [2] for elderly people monitoring. A significant survey on human activity recognition [22] concludes that the availability of depth information can improve the recognition accuracy. So, the onset of Kinect, a Microsoft gaming platform, with the capability to capture depth along with the color information creates a new area of research on the problem of human activity recognition because of the availability of the depth information along with the color value. On the other hand the results of the HARL competition organized in ICPR 2012 [6] shows that the human activity recognition accuracy increases with the increase in human localization accuracy. Therefore, the human localization approaches on RGB videos need to be modified with the availability of depth information. Traditional back ground modeling based methods didn't work on the RGB-D data

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 325–336, 2014.

when the camera is not static and the lighting condition varies over time. The video frames, on which we have tested our system, contains human leaning over wall, one person occludes another person partially which makes the task of human detection more difficult as color based segmentation didn't work, too. One such example image is shown in Figure 1. In this paper we have concentrated on the problem of human detection from Kinect captured videos by combining RGB and depth information. We have proposed a system that can detect the presence of human in such a G-D video using machine learning technique, namely C-HMM. We have compared our method against other methods like on [6] dataset as well as on our own data set.

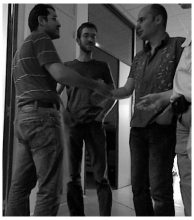


Fig. 1. Apparently touching objects in 2D projection plane

2 Related Literature Survey

There have been a large number of methods dealing with recognition of human activity in color image. The recognition problem is very difficult because of large variation involved in human appearances and views. From the literature review it can be seen that, face detection algorithm is often applied [8], [9] for recognizing human in image and video. They addressed the solution in this field either by feature based or image based approach. Bottom-Up analysis is done utilizing feature based approach. Window scanning technique is employed in later case. Muhammad and Atif et al. [10] combines color and motion information to detect face and hence, human. In [11], human detection is achieved by integrating the cascade-of-rejectors concept with the Histogram of Oriented Gradients (HoG) of variable size blocks. They used an AdaBoost training algorithm to learn a cascade of rejecters to eliminate the non human image patches. A new learning method for human detection is proposed in [12] which is based on weak classifier, built from L1-norm minimization learning scheme (LML). The augmentation of edge-based features, texture measures and color information have been used by Schwartz et al. [13]. They handle this high dimensionality resulting from the combination of features, using dimensionality reduction technique Partial Least Squares (PLS). There are several papers [16], [21] that addressed human detecting problem based on body-part. Bhaskar and Jordi [21] presents a technique for view invariant human detection using body-part(head, leg, arm etc) detectors. Human detection is proposed by probabilistic body part assembly in [16]. First, different body parts are detected by Adaboost and after that detected parts are assembled using RANSAC. Mohan et al. [17] have used hierarchical classification architecture using SVM. They also use different components detector at primary level. These method perform poor due to failure of detector in handling of variability of body parts. In [14] authors have presented a graphical model based approach for estimating poses of upper-body parts by fusing depth and RGB color data based on Haar cascade. This method works well for detecting upper-body human pose but not for full human shape. In addition to this, the primary focus of the above methods is identification of front view of human but not view-invariant. Lu et al. [15] uses 2-D head contour model and a 3-D head surface model to detect head of human. Next, segmentation scheme is used to extract the whole contours of human based on head point. The performance of the method extremely depends on the accurate head detection and it uses only depth information not both depth and color. Therefore, from the above discussion, it can be concluded that there are methods to improve human recognition rate but these methods concentrate on color and edge information of images and a little variation in view point but not on color, depth and large view where we can expect much more challenges compared to exiting state of the art methods. Again, the literature also suggests that the segmentation of the input prior to recognition can lead to higher recognition rate [18]. But most of the methods do not take care about proper segmentation. Therefore, improving human detection in Kinect through segmentation, irrespective of view point and background complexity is challenging. In this paper, we present a method of human detection in images captured by Kinect and performs recognition of human using HMM instead of detecting individual body parts. HMM is popular and found robust in printed and handwritten text recognition. It motivates us to use HMM in human recognition. To the best of our knowledge, there is no work on human detection using HMM. This work is motivated by our preliminary work reported in [23]. Our goal is to classify observed feature sequences into human or nonhuman category utilizing depth and color information. Hence, this paper presents view-invariance human detection method using HMM approach. The rest of the paper is organized as follows. The proposed method is described in Section 3 which includes depth based segmentation method, noise cleaning and view-invariant human detection method using HMMs. Section 4 presents the experimental results. Finally, conclusion is drawn in Section 5. In this paper human detection and human localization phrases were some times used interchangeably.

3 Proposed Method

Our proposed method initially attempts to localize the human from the video frame to reduce the time complexity of the method. Kinect based systems have a major limitation that it can work on indoor environment only as the depth sensor uses Infra Red (IR) signal. So the captured videos contain some wall, floor and ceiling parts. We remove noisy components from the candidate regions by removing the floor/ceiling from the image if required. Next we extract the LGH features and train them to classify into human and non-human class. Finally we use C-HMM classifier to separate out human and non human.

3.1 Segmentation Using Depth Connected Operator

Kinect provides two sets of values namely the G image and the D information for each video frame. Any G image is a 2-Dimensional projection of the 3-Dimensional objects located at different distances from the sensor. So it is not possible to segment them on their depth unless we have the depth information exclusively. So we have used the depth information to segment each video frame into number of layers so that each segment contains the pixels those are connected over depth. 8 neighbor connected component analysis is a common method of image processing but we have used depth connectivity instead. The method of depth connectivity is described in details in [23]. The method of is based on the concept of of connected operator for sets as described in [7]. The concept of connected operator on set says that an operator ψ can be said to be a connected operator if the symmetrical difference $P \Delta \psi(D)$is exclusively composed of connected components of D or its compliment D^C. Here we use depth information obtained from the depth sensor of Kinect as the connected operator ψ applying over the gray scale image pixel set G. As per the definition of partition space as stated in [7] this ψ operator partitions the space G into two disjoint subsets G_i and G_j such that $G_i \bigcap G_j = \emptyset \forall i \neq j$. Our proposed method of such segmentation is described below:

- For each video frame/image construct a set (P) by concatenating the gray and the depth information. So P is a two tuple set containing gray value (g) of the pixel and its depth (d) information from the sensor. So $P = (g_i, d_i)$
- Every pixel who are connected over depth are mark with the same depth map label. We have used a threshold to check the depth connectivity.
- For each depth map label we create one image with keeping the gray value of those pixels as it was and mark the rest of the pixels as black.

In Figure 2 we have shown one such example video frame/image and its four out of five partitions. In this image the backgrounds are marked as black. This image shows that one man is standing in one partition and rest two are residing in an another partition. The main advantage of this method over the simple depth quantization is that human are not separated into two different segments in our approach when the human is between the depth separation.

3.2 Floor and Ceiling Removal

The depth connectivity based segmentation generates multiple images form a input video frame those are connected over depth. Each of these images represents the objects connected over depth by their corresponding gray scale value. The height and width of the images are same as that of height and width of the original video frame/image. The backgrounds are marked as black and an additional tag is added to mark them as background. We have implemented the additional back ground tag information by using a Boolean Flag which is set for background and FALSE for foreground. We shall refer each of these images as a partition of the original video frame. Connected component analysis is used

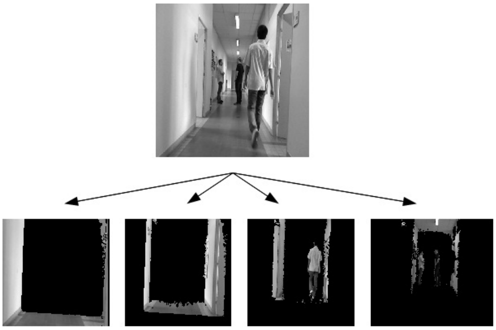


Fig. 2. Objects at different depth from the camera

to mark the different components in each of these partitions. The outcome of connected component analysis shows that some components containing human include some non-human objects through the floor/ceiling. In such cases the component width is almost equal to the width of the input video frame/image. So we formulate a vertical pixel projection based method to eliminate the floor region and thus separate out the human component from the rest part for such components. Here is the description of that proposed method:

- If the width of the component is greater than 75% of the width of input video frame/image execute the following steps. We have used this value as a heuristic obtained from our experiments.
- Count the number of pixels (cnt_i) in a column i for which the background flag is FALSE
- Run a K-Means clustering with K=2 on $cnt_i \; \forall i \in 0, H$ where H is the height of the image
- The two cluster will represent the columns with higher number of FG pixels (C_1) and lower number of FG pixels (C_2)
- Make the flag for all the pixels from the column those are residing in C_2 as TRUE

This method is explained using the example images. Figure 3. a shows one such partition which contains both human and non human objects. Now the corresponding vertical projection is shown in Figure 3. b. Finally the outcome of our proposed method is shown in Figure 3. c which shows that the connected component is now divided into multiple components. We run connected component analysis as described above on each of these partitions after being modified by the above method. Finally we get some segments containing either human or non human objects. Some outcomes of this method are shown in Figure 4. a and Figure 4. b. Finally some non human objects are removed by effacing of noisy component as described in [23].

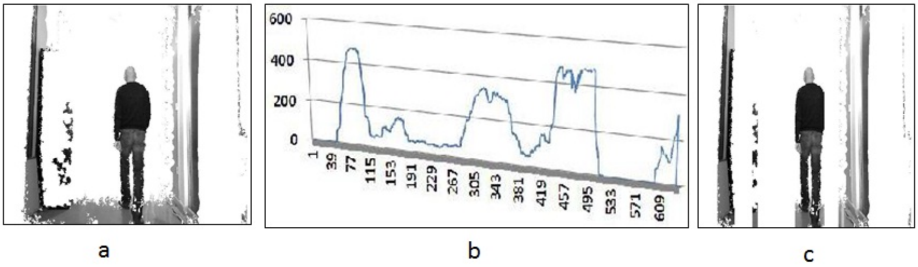


Fig. 3. a) One partition b) Histogram of FG pixels in each column c) Image after floor estimation and removal

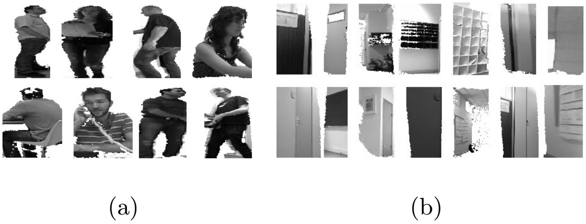


Fig. 4. a) Some segments containing human b) Some segments containing non human

3.3 LGH Feature

Sliding Window is a common technique for many signal processing applications like speech and character recognition. We have used a rectangular sliding window of l pixel width to collect the features of a component. It is shifted from left to right across the normalized gray level segmented image to generate feature vector sequences at each shift position. Adjacent image windows overlap in the vertical direction. This results in a vast amount of features for each frame. The frames are normalized to a pre-defined height before the feature extraction stage. Figure 5 illustrates an example of the sliding window feature extraction process. This feature extraction approach is based on the calculation of the local gradient histogram [24]. Each sub-image is sub-divided into 4 * 4 blocks and from all pixels in each block a histogram of gradient orientations is calculated. Here, we considered 8 orientations. Therefore, the final feature vector concatenation of the 16 histograms results in vector containing 128 features.

- For the entire image the horizontal and vertical motion components V_x and V_y are determined and a gradient magnitude (m) is computed for each pixel.
- The field vector $\overrightarrow{V}$ is sliced up in an L bin histogram.

- Each bin specifies a particular octant in the angular radian space. Here we consider 8 bins (360°/45°) of angular information.
- The concatenation of the 16 histograms of 8 bins provides a 128-dimensional feature vector for each frame.

Let $\overrightarrow{V}=(V_x, V_y)$ and histogram $H = h(1), h(2), ..., h(8)$. The histogram is constructed by quantizing $\theta(x,y) = tan^{-1}\frac{V_y}{V_x}$ and adding up $m(x,y) = \sqrt{(V_x + V_y)}$ to the bin indicated by quantized θ. In mathematical definition,

$$h(i) = \begin{cases} \sum_{x,y} m(x,y) & when\ \theta \in ith\ octant \\ 0 & otherwise \end{cases} \tag{1}$$

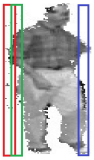


Fig. 5. Path of overlapping sliding window (shown in different colors)

3.4 HMMs Based Recognition

HMMs have been proven to be a powerful stochastic approach and found robust in speech and text (printed and hand written) recognition. It is a special type of dynamic Bayesian networks. We have used HMM in our application because of its ability to cope with variable-length observation sequences obtained from images. Generally, HMM follows the first-order Markov assumption where each state S_t at time t depends only on the state S_{t-1} at time $t-1$. It contains a fixed number of hidden states. HMM is characterized by 3 matrices: state transition probability matrix A, symbol output probability matrix B, initial state probability matrix π. The parameter A, B and π are determined during learning process. The image is represented as a sequence of feature vectors $X = x_1, x_2, ..., x_T$ also known as sequence of frames. In HMMs, the likelihood of emitting a frame x_t in state i is modelled using a GMM. For a model λ, if O is an observation sequence $O = (O_1, O_2, .., O_T)$ which is assumed to have been generated by a state sequence $Q = (Q_1, Q_2, ., Q_T)$, of length T. We calculate the observations probability or likelihood as follows:

$$P(O, Q|\lambda) = \sum_{Q} \pi_{q1} b_{q1}(O_1) \prod_{T} a_{qT-1} qT b_{qT}(O_T) \tag{2}$$

where π_{q1} is initial probability of state 1, is transition probability from state i to state j and is output probability of state i. The observation likelihoods are computed from a Gaussian Mixture Model (GMM).

$$b_j(x) = \sum_{k=1}^{M_j} c_{jk} \mathcal{N}(x, \mu_{jk}, \Sigma_{jk}) \tag{3}$$

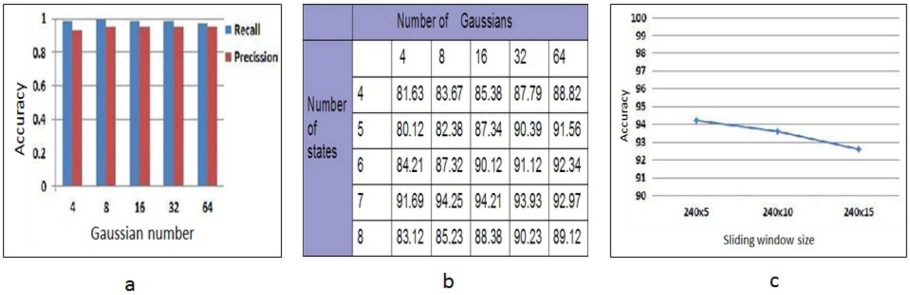
where, M_j is the number of Gaussians assigned to j. and $\mathcal{N}(x, \mu, \sigma)$ denotes a Gaussian with mean μ and covariance matrix σ and c_{jk} is the weight coefficient of the Gaussian component k of state j. Next, the Viterbi decoding searches the subsequence of an observation that matches best to a given HMM. For a classifier of C categories, we choose the model which best matches the observation from C HMMs $\lambda_m = A_m, B_m, \pi_m$, where $m = 1, ..., C$, and $\sum_{m=1}^{c} \lambda_m = 1$. This means when a unknown sequence of unknown category is given, we calculate $P(\lambda_i|O)$ for each HMM $\lambda_m$and select λ_c^* , where

$$c^* = argmax_m P(\lambda_m|O) \tag{4}$$

An HMM should be learned for each class. For our application 2 HMMs have been used to model human and non human. The 128 dimensional LGH features extracted from each sliding window of image were used to represent sequence of local feature vectors. The extracted feature of each window is arranged row-wise to form complete vector set. The task of the learning algorithm is to find the best set of state transitions and observation probabilities. The Baum-Welch recursive algorithm is used to obtain the final parameters of HMMs. For classifying an observed symbol sequence O, classifier choose the model whose likelihood is highest as the recognition result. The recognition is performed using the Viterbi algorithm.

4 Result and Discussion

It is well known fact that when the training sample size is small, the recognition rate becomes low. The performance of any recognition system depends not only size number but also the well variation of samples used in training phase as these both are very crucial to estimate HMM parameters. To construct a robust recognition system, appropriate training patterns are important. This means training pattern should capture the maximum test pattern variation. In this experiment training and test samples were completely segregated to make this evaluation more reliable. We have developed a working prototype of human detection using x86 PC system. C/C++ and OpenCV library were used for segmentation and feature extraction on a windows environment. We have used the popular HTK toolkit for HMMs training and evaluation [20]. Our data set includes G frames of the Kinect module are encoded as lossy JPEG images and D frames of the Kinect module are encoded in lossy 16bit JPEG2000 images with a compression factor of 20, resulting in 30KB per frame. We validated our proposed system on two data sets. One of these data sets is our own. This data set includes the Kinect captured RGB-D images recorded at our lab which includes more than 30 videos each having more than 1000 frames. The other data set used in our experiment is the [6] data set published by LIRIS for Human Activity Recognition and Localization (HARL) in ICPR 2012. This database contains 107 training and 69 test videos with gray and depth information, captured in Kinect sensor in indoor scenes. These videos contain different views of human with a vast range of poses like standing, walking, siting while performing action. These are taken under different illumination condition with camera movement and scale variation by different person against complex background. We have used 35,434 segments during training phase. Among them 14,867 segments are positive and 20,567 segments are negatives. Positive consists of human containing image. On the other hand, representation of negative images are floor, ceiling and wall component (shown in Figure 4. a, 4. b). We have applied our method for all those test images for human detection. For human detection, a set of human images is used in the training of HMM. The images in the training set represent different views of different persons taken from segmentation results generated by previous step. Once the models have been trained for human and non human, the Viterbi algorithm was applied to find the most likely state sequence and its likelihood in the recognition process. The observation sequences for a image are formed from image window or block that are extracted by scanning the image from left-to-right. The observation vectors consist of 128 features. In training set, we have measured the performance of system using a 10-fold cross validation. Recall, describes how many object have been correctly detected, with respect to the total number of objects in the dataset and Precision evaluate how many detected objects are matched with respect to the total number of detected objects. We define recall (R) as $R = \frac{c}{c+m}$ and precision(P) as $P = \frac{c}{c+fp}$ where c indicates the correct recognition, m means misses and fp is the false positive. If we can't detect a human we define this error as miss as our intention of the research is to localize the human. On the other hand if our proposed method detects one component as human though it is actually a non human one, we define it as false positive. We have observed that our P is always less than R. The main reason behind that error comes from the lack in training of different instances of non human objects. In Figure 6. a we have shown the recall and precision against Gaussian number. We have evaluated our algorithm performance in terms of *variation of Gaussian number, variation of states and different sliding window size.* Experiments show that learned HMM classifiers have good performance for detecting human. Results show that most of the humans have been correctly recognized. The HMM system was tested with different number of states and Gaussian numbers. In Figure 6. b, the recognition accuracies are given in terms of both of these. Next, we inspected that increasing the number of states up to 8 states improves the performance of the HMM recognizer, but a larger number and small number states decreases its accuracy. The best average accuracy was obtained for an HMM with 7 states and a mixed output probability of 16 Gaussians, 94.25% with Local gradient Histogram(LGH) on unseen samples. We have shown the effect of sliding window size on recognition accuracy in Figure 6. c. We observed that the increase in sliding window size after a limit reduces the recognition accuracy. The advantage of the sliding window based-HMM is that the detection of human is very robust. Some qualitative images detected from our approach are shown in Figure 7. a and c. Our proposed method can detect the human in case of improper segmentation and even when the human is blended while sitting and partially occluded by other object and human as shown in Figure 7. b. We have compared the performance of our proposed



	Number of Gaussians					
Number of states		4	8	16	32	64
	4	81.63	83.67	85.38	87.79	88.82
	5	80.12	82.38	87.34	90.39	91.56
	6	84.21	87.32	90.12	91.12	92.34
	7	91.69	94.25	94.21	93.93	92.97
	8	83.12	85.23	88.38	90.23	89.12

Fig. 6. a)Recall and precision accuracy on testing set b) Recognition accuracy for different number of states and Gaussians on testing set c) Recognition accuracy vs Sliding window size

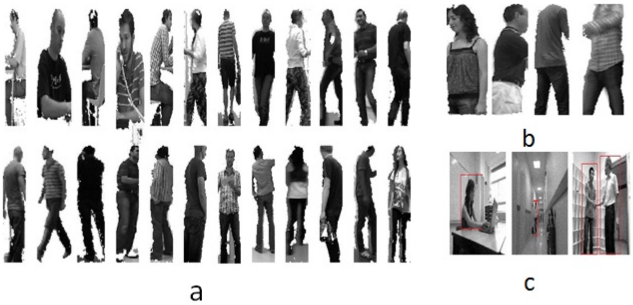


Fig. 7. a) Example of view-invariant human recognition generated by proposed method b) Recognition of partial segmented human part c) Detection results of the proposed method

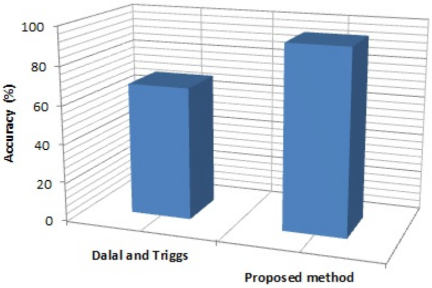


Fig. 8. Comparative result

method against a state of the art method described in Figure 8. We observe that our algorithm outperforms the [19] method. [19] works only when the human is in upright position but in the real world human can be found in any orientation.

5 Conclusions

In this paper we have presented a method that combines color and depth information in the pre-processing phase to localize the candidate segments containing human being and thus the proposed method overcomes the limitations of 2D based methods. The use of depth with robust machine learning framework makes the system robust against variations in viewpoint. So the performance of the proposed system outperforms the state of the art methods. We have shown that using locally normalized histogram of gradient orientations features descriptors in a overlapping window with HMM framework gives very good results for person detection. These results show that our method is promising to recognize human for numerous applications such as video indexing and retrieval, intelligent human machine interaction, video surveillance, health care, driver assistance, automatic activity detection and predicting person behavior. Performance of our proposed method partly depends on the candidate human localization and all errors are mostly coming from the failure in proper localization. This method can precisely identify whether the candidate segment contains human being or not and thus if the area of the candidate region is much bigger than the actual human region, the accuracy falls. So we are currently working to find a better localization method.

References

1. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGBD Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. In: Proc. ISER 2010 (2010)
2. Trinh, H., Fan, Q., Pankanti, S., et al.: Detecting Human Activities in Retail Surveillance Using Hierarchical Finite State Machine. In: Proc. ICASSP 2011, pp. 1337–1340 (2011)
3. Trinh, H., Fan, Q., Gabbur, P., Pankanti, S.: Hand tracking by binary quadratic programming and its application to retail activity recognition. In: Proc. CVPR 2012, pp. 1902–1909 (2012)
4. Gabbur, P., Pankanti, S., Fan, Q., Trinh, H.: A pattern discovery approach to retail fraud detection. In: Proc. KDD 2011, pp. 307–315 (2011)
5. Sinha, A., Chattopadhyay, T., Mallik, A.: Segmentation of Kinect Captured Images using Grid Based 3D Connected Component Labeling. In: Proc. VISAPP 2013, pp. 327–332 (2013)
6. Wolf, C., Mille, J., Lombardi, L.E., Celiktutan, O., Jiu, M., Baccouche, M., Dellandrea, E., Bichot, C.-E., Garcia, C., Sankur, B.: The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition, Technical Report RR-LIRIS-2012-004. In: Proc. ICPR 2012 (2012)

7. Salembier, P., Serra, J.: Flat Zone Filtering, Connected Operator, and Filters by Reconstruction. Proc. IEEE Transactions on Image Processing 1995, 1153–1160 (1995)
8. Jin, R., Hauptmann, A.G.: Learning to Identify Video Shots With People Based on Face Detection. In: Proc. ICME 2003, pp. 6–9 (2003)
9. Low&, B.K., Hjelmas, E.: Face Detection: A Survey, Computer Vision and Image Understanding 2001 (2001)
10. Khan, M.U.G., Saeed, A.: Human Detecion in Videos. Journal of Theoretical and Applied Information Technology (2009)
11. Zhu, Q., Yeh, M., Cheng, K., Avidan, S.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: Proc. CVPR 2006, pp. 1491–1498 (2006)
12. Xu, R., Zhang, B., Ye, Q., Jiao, J.: Cascaded L1-norm Minimization Learning (CLML) classifier for human detection. In: Proc. CVPR 2010, pp. 89–96 (2010)
13. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: Proc. ICCV 2009, pp. 24–31 (2009)
14. Jain, H.P., Subramanian, A., Das, S., Mittal, A.: Real-time upper-body human pose estimation using a depth camera. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2011. LNCS, vol. 6930, pp. 227–238. Springer, Heidelberg (2011)
15. Xia, L., Chen, C., Aggarwal, J.K.: Human Detection Using Depth Information by Kinect. In: Proc. CVPRW 2011, pp. 15–22 (2011)
16. Micilotta, A., Ong, E., Bowden, R.: Detection and tracking of humans by probabilistic body part assembly. In: Proc. British Machine Vision Conference 2005, pp. 429–438 (2005)
17. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. Proc. IEEE Transaction on Pattern Analysis and Machine Intelligence 2001, 349–361 (2001)
18. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation. Proc. IEEE Transaction on Pattern Analysis and Machine Intelligence 2009, 1685–1699 (2009)
19. Dalal, N., Triggs, B., Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
20. Young, S.J., Jansen, J., Odell, J.J., Ollason, D., Woodland, P.C.: The HTK Hidden Markov Model Toolkit Book. Entropic Cambridge Research Laboratory (1995)
21. Chakraborty, B., Rudovic, O.N., Gonzlez, J.: View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In: Proc. FG 2008, pp. 1–6 (2008)
22. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) 2011 43(3) (2011)
23. Chattopadhyay, T., Roy, S.: Human Localization at Home Using Kinect. In: Proc. HomeSys, UbiComp (Adjunct Publication) 2013, pp. 821–828 (2013)
24. Jos, A., Serrano, R., Perronnin, F.: Handwritten word-spotting using hidden Markov models and universal vocabularies. In: Proc. Pattern Recognition 2009, pp. 2106–2116 (2009)

A Survey of Datasets for Human Gesture Recognition

Simon Ruffieux[1], Denis Lalanne[2], Elena Mugellini[1], and Omar Abou Khaled[1]

[1] University of Applied Sciences and Arts of Western Switzerland, Fribourg
{Simon.Ruffieux,Elena.Mugellini,Omar.AbouKhaled}@Hefr.ch
[2] University of Fribourg
Denis.Lalanne@unifr.ch

Abstract. This paper presents a survey on datasets created for the field of gesture recognition. The main characteristics of the datasets are presented on two tables to provide researchers a clear and rapid access to the information. This paper also provides a comprehensive description of the datasets and discusses their general strengths and limitations. Guidelines for creation and selection of datasets for gesture recognition are proposed. This survey should be a key-access point for researchers looking to create or use datasets in the field of human gesture recognition.

Keywords: human-computer interaction, gesture recognition, datasets, survey.

1 Introduction

The fields of human activity, action and gesture recognition gained more and more attention these last years, notably due to the numerous affordable sensors commercially released. In recent years, more and more datasets have been created by researchers in order to develop, train, optimize and evaluate algorithms; several of them have been made publicly available to developers and researchers. Several articles have already addressed the topic of datasets for the general field of human activity and action recognition [1,2] and a couple of websites already list publicly available datasets [3,4]. However the topic of datasets for the specific field of gesture recognition has not been addressed yet. Gesture recognition is defined as a subset of human action and activity recognition and generally requires its own specific datasets for the development of algorithms. The devices and sensors employed are often similar in both fields however they are generally used with different setups in gesture recognition: the sensors tend to be closer to the user in order to augment the granularity and the users are generally aware of the presence and position of the sensor thus interacting towards it. Therefore, most datasets acquired for activity and action recognition cannot be directly used for gesture recognition; the same asset is, in most cases, also applicable with algorithms.

The goal of the present survey is two-fold: provide an overview and a discussion about the available datasets and provide brief guidelines to help researchers when selecting or creating datasets.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 337–348, 2014.

This work takes place in the context of the FEOGARM project [5]. The goal of FEOGARM is to provide a comprehensive framework for facilitating gesture evaluation and recognition methods. A dataset for gesture recognition has been publicly released in this context [6].

2 Related Works

Several surveys have already addressed topics related to datasets although most of them have mostly considered the field of human action and activity recognition; only short sub-sections were addressing the gesture recognition domain. A recent and informative survey addressed the topic of datasets for activity recognition but explicitly omitted datasets focusing solely on gesture recognition in order to narrow the survey [3]. Another survey addressed the methods, systems and evaluation metrics for vision based human-activity recognition to detect abnormal behaviors in videos streams, a subset of activity recognition called surveillance systems [7]. Large surveys of the activity recognition domain are also available [8,9], resuming the taxonomies, techniques, challenges and listing the datasets for full-body activity recognition. In [10], a survey of the datasets for action recognition are presented, a domain at the frontier between activity and gestures. In [11], the datasets available for pose estimation and tracking are listed and discussed; the need for common standards in the domain is strongly highlighted. These surveys provide a good overview of the human activity and action recognition field, although they do not address directly gesture recognition.

The surveys that specifically addressed the field of gesture recognition have mostly considered three perspectives: the topic of gesture recognition in general [12,13], the specific topic of hand gestures for human-computer interaction [14,15] and the topic of sign language [16]. None of these surveys focused on the specific topic of the existing datasets for gesture recognition. A few research papers have addressed topics such as modeling, building and using datasets in the context of gesture recognition. In [17], they presented a framework based on databases for gesture recognition. They developed an ASL and hand shape real-time recognition systems based on comparisons with examples of images stored in their databases. The developed method demonstrated the ability to search a gesture database fast enough for real-time gesture recognition applications. However the low accuracy rate of the recognition system was not satisfactory and required some additional work. In [18], they discussed and highlighted some important modeling considerations when creating a database for hand gesture recognition in the context of natural interfaces. They identified the required assumptions to create an effective database: naturality of the gesture set, size of the set, a precise analysis of the potential effects of the recording conditions and a precise description of the acquisition process. They also stated the importance of recording the data with multiple sensors as a way to achieve independence from the acquisition conditions; they notably promote motion capture systems and video cameras. Finally, in [19], they study the impact of the semiotic modalities such as text, images or videos, which are used to instruct the subjects, on the quality of the performed gestures. They also illustrate the importance to balance correctness and coverage properties of a gesture dataset in order to obtain the best recognition performances with machine learning algorithms. The study demonstrated that video instructions promote correctness while texts and images together are best for coverage; the latter also giving a strong sense of freedom to the subjects. Gesture datasets are also slowly moving away from research and spread to the commercial market; for example, ARB Labs [20] has recently started a company based on a gesture dataset and the related acquisition software.

3 Survey

This section presents the main datasets that have been employed or developed for the field of gesture recognition these last years. The datasets are presented through two chronologically ordered tables: Table 1contains the general information and a short description for each datasets. Then Table 2 resumes the main technical characteristics and categorizes the datasets according to the three main types of ground truth annotations. Note that older datasets have been omitted due to the important changes in data quality and on the types of sensors employed. This survey has also been limited to datasets containing gestures mostly involving hand(s) and arm(s) motion.

The Table 1 provides an overview of the 15 reviewed datasets. The table presents the *name* or acronym of the datasets and their reference paper. The number of *citations* for the reference papers, which have been retrieved from Google Scholar the 03.02.2014. Two of the papers have more than one hundred citations. The *placement of the sensor(s)* indicates if the sensor was placed in the environment or on the user. The sensors and their placement are rather constant amongst reviewed datasets. Most datasets rely on a single video camera at a fixed location in the environment. Only a couple of datasets used alternative setups such as multiple video cameras or a combination of environmental and wearable sensors. Only two datasets are based on environmental and wearable data. The ChAirGest dataset uses a combination of RGB-D camera fixed in the environment and inertial motion units (IMU) located on the arm of the user. The 6DMG dataset uses a combination of hand-held controller and optical tracker to obtain both the motion of the hand and its position in the space. Such setup enables the comparison or the fusion of both approaches on common material. *The quality of information* depicts the amount of documents, description and information which have been provided with a dataset. Such documentation can be very important to understand and use a dataset. Large variations can be observed between datasets. The *types of gestures* distinguish the gesture vocabularies present in the datasets. Datasets are either taking their vocabulary from existing ones such as sign language [21], cultural signs [22] or military gestures [23] or creating original vocabularies. Numerous datasets uses their own vocabulary of gestures which are thought for specific applications or domains such as gaming [24] or human-computer interaction [6]. These vocabularies usually rely on iconic gestures which imbue a correspondence between the gesture and the reference, symbolic gestures which are highly lexicalized and metaphoric gestures which correspond to an abstract representation.

Table 1. This table provides a general description of the most recent datasets for human gesture recognition. Notation for; for the information: from poor (1) to very good (5); and finally for the availability: Public, Public on Request or Not Yet.

Name	Year	Citations (03.02.2014)	Sensors placement	Information	Types of gestures	Purposes & Description	Availability
3DIG [25]	'13	1	Environment	2	Iconic	**Recognition** of iconic gestures where subjects were free to perform their own gesture to depict each object	P
ASL Dataset [21]	'13	-	Environment	5	Sign language	American sign **recognition**. Evaluation of hands **detection** & **tracking.** Acquisition still on-going.	NY
CGD2013 [22] **ChaLearn Dataset**	'13	2	Environment	5	Metaphoric	Multimodal gesture **recognition** of cultural Italian gestures accompanying speech. **Challenge**-related dataset	P
ChAirGest [6]	'13	1	Env. & wear.	5	Iconic & metaphoric	Gesture **spotting** & **recognition** from multimodal data in the context of close HCI. **Challenge**-related dataset	PR
SKIG [26]	'13	5	Environment	3	Iconic & metaphoric	Improve gesture recognition from RGB-D data, notably with different illuminations. Hand gesture **recognition** seen from above	P
6DMG [27]	'12	2	Env. & wear.	5	Iconic & metaphoric	Explore gesture **recognition** from implicit & explicit data. Subjects performed the gestures with a Wiimote in their right hand	P
MSRC-12 [19]	'12	21	Environment	5	Iconic & metaphoric	Gesture **recognition** from the skeleton data. Study the motion variation across users with skeleton data	P
G3D [24]	'12	7	Environment	5	Iconic	Gaming actions and gestures **recognition** & **spotting.** Specifically designed to improve gaming without controller	PR
MSRGesture3D [28]	'12	17	Environment	3	Sign language	Sign language **recognition** from hand depth data. Only the segmented hand sections of the images are provided	P
CGD2011 [29] **ChaLearn Dataset**	'11	17	Environment	5	Iconic & metaphoric	Improve one-shot learning for **recognition**. **Challenge**-related dataset. The competition had a large success.	P
NATOPS Aircraft Handling Signals Database [30]	'11	26	Environment	5	Metaphoric & symbolic (Real vocabulary)	Body-and-hand **tracking** & gesture **recognition** requiring both body and hand information to distinguish gestures	PR
NTU Dataset [31]	'11	68	Environment	2	Metaphoric & symbolic poses	Hand pose & shape **recognition** in cluttered conditions. Only contains static images, no motion.	P
Keck Gesture Dataset [23]	'09	153	Environment	4	Metaphoric & symbolic (Real vocabulary)	Military gestures performed with perturbations in the background. Designed to evaluate gesture **recognition** and **spotting** in harsh conditions.	P
ASLLVD [32]	'08	17	Environment	4	Sign language	A reference database in automatic sign language **recognition** and **spotting** with data captured from several viewpoints.	PR
CHGD [33] **(Cambridge Hand Gesture Dataset)**	'07	136	Environment	4	Metaphoric	Hand segmentation & gesture **recognition** in varying illuminations conditions. It only contains sequences of images.	P

Although most datasets span on multiple gestures types, some dataset focus on a specific type. For example the approach of 3DIG dataset focusing on iconic gestures is interesting: the subjects were free to perform the gesture of their choice to depict a specific object; the classification goal being to recognize the depicted object. Such approach generates large variations within a class which complexifies the recognition. Then the *purposes* and a short *description* of the datasets are provided to better characterize each dataset. Finally the last column shows the current *availability* of each dataset. In this survey, all the presented datasets are available online either publicly or on request, except one which was not yet available. Generally datasets are available on request due to image rights of the recorded subjects; researchers have to sign an End-User License Agreement (EULA) to obtain a dataset. This EULA ensures that researchers will preserve the data of the subjects. Only one of the reviewed datasets is available commercially and has not been listed in the tables due to the lack of information about it [20]. Note that for some datasets, notably the ones used in challenges, only around 75% percent of the instances are publicly available, the remaining is kept private to safely evaluate the performances of the algorithms developed by the challengers.

The Table 2 resumes the main technical characteristics of the reviewed datasets. It resumes the *body-parts* that are involved in the gestures to recognize. The reviewed datasets are quite heterogeneous in that respect, spanning from single hand to full-body. For example, the gestures from the CGD2011 dataset could be recognized only by having the information from the two hands and arms. The *sensor view-point* indicates the position(s) of the video sensor(s), when applicable, with respect to the subjects. Most datasets use a front-view, with the sensor in front of the subject. However a couple of datasets use a top-view, with the sensor above the user and facing downward, which greatly simplifies hand recognition from the images. A few other datasets use different approaches: a trade-off between top and front view for the ChAirGest dataset which uses a sensor inclined at 45° or multiple simultaneous view-points for the ASSLVD dataset. A single dataset contains a moving camera in order to evaluate algorithms in difficult conditions. The *subject stance* corresponds to the position of the user during the recording. For most datasets, subjects were standing in front of the camera, although in a few datasets, subjects were sitting on a chair which implies interaction with whole or part of the upper-body. The Keck Gesture Dataset is the only reviewed dataset containing subjects who are moving during the interaction; a very challenging recognition task. Finally the more classical characteristics: the *number of subjects* who are available in the data, the number of distinct classes (gestures) and the total number of instances. In general, the more subjects, classes and instances, the better. However, it is usually important to have a high ratio between the number of instances and the number of classes to properly train machine learning algorithms. Then the *sensors* used to acquire the data are described. The Kinect-based dataset have not all recorded each of the streams from the sensor; Kinect being a multimodal sensor, it provides color and depth stream, the approximate position of the subject's body-parts through a skeleton representation and the sound. Non video-based sensors include inertial motion units (similar to motion sensors embedded in phones, smartwatches and smart-bands), Optical tracker or Vicon system for motion capture or a Wiimote+ controller from Nintendo. The next column indicates the *resolution f*or the sensors based on videos. An increase of the resolution through the years is clearly observable. Higher resolution implies more information in the image but also more processing time when processing an image. The *frequency* is indicated for all mentioned sensors, when applicable. Similarly to the resolution, a higher frequency means more information but increases processing time and data storage size; many algorithm implementations artificially down sample the frequency for real-time applications. However, a high frequency is important in order to capture all the information during rapid movements. Finally the *size* of the datasets in Gigabytes (GB) usually results from the previous choices and can largely vary across datasets.

Table 2. This table provides technical information about the most recent datasets for human gesture recognition. Notation for *body-parts*: Full-Body, Upper-Body, Hand and Arm; for the *sensor-view*: Front-View, Top-View, Lateral-View and Moving-View; for the *user stance*: Standing, Sitting and Moving; for the *Kinect sensor*: Color, Depth, Skeleton and Sound.

Name	Body-parts	Sensor view	Subject stance	Subjects	Classes	Instances	Sensors	Resolution	Frequency [Hz]	Size [GB]	Ground truth		
											Label	Temporal	Spatial
3DIG [25]	HA	FV	St.	29	20	1739	Kinect^{CDS}	640x480	30	85[2]	X		
ASL Dataset [21]	UB	FV	St.	2	1300+	1300+	Kinect^{CDS}	640x480	25	?	X	X	X[1]
CGD2013 [22] **ChaLearn Dataset**	UB	FV	St.	27	20	13000	Kinect^{CDSSo}	640x480	20	27[2]	X	X	
ChAirGest [6]	HA	FV 45°	Si.	10	10	1200	Kinect^{CDS} 4 IMU	640x480 -	30 50	1000 3[2]	X	X	
SKIG [26]	H	TV	Si.	6	10	1080	Kinect^{CD}	320x240	10	1.2[2]	X		
6DMG [27]	H	-	-	28	20	5600	Wiimote+ Optical tracker	-	60	0.02	X	X	X
MSRC-12 [19]	FB	FV	St.	30	12	6244	Kinect^{S}	-	30	0.2	X	X[3]	
G3D [24]	FB	FV	St.	10	20	600	Kinect^{CDS}	640x480	30	47[2]	X	X	
MSRGesture3D [28]	H	FV	-	10	12	336	Kinect^{D}	130x130	20	0.03	X		
CGD2011 [29] **ChaLearn Dataset**	2HA	FV	St.	20	30	50'000	Kinect^{CD}	320x240	10	30 5[2]	X	X	X[1]
NATOPS Aircraft Handling Signals Database [30]	UB	FV	St.	20	24	9600	Stereo Cam. Vicon[1]	320x240	20	19	X	X	X[1]
NTU Dataset [31]	H	FV	Si.	10	10	1000	Kinect^{CD}	640x480	-	0.1	X		
Keck Gesture Dataset [23]	2HA	FV MV	St. Mo	3	14	294	Color Cam.	640x480	15	0.15	X	X	
ASLLVD [32]	UB	3FV LV	St.	6	2700	3300	4 Color Cameras	640x480	60	1.6[2]	X	X	
CHGD [33] **(Cambridge Hand Gesture Dataset)**	H	TV	Si.	2	9	900	Color camera	320x240	?	1	X		

When working with video, many datasets offer a couple of data qualities: raw or compressed/encoded qualities. Encoding video dramatically reduces the size of the data with only a partial loss of information but a large gain in download, loading and processing times. The *types of ground truth* present in the datasets have strong impli-

[1] Only for part of the data.

[2] The data has been encoded or compressed.

[3] Only the start event of gestures has been temporally labeled.

cations on the type of algorithms that may be trained and evaluated. Therefore this information has been used as a way to categorize the datasets. In this work, datasets are grouped in three non-exclusive incremental categories: recognition, spotting and tracking. This categorization allows the definition of the potential usage(s) of the dataset. Gesture labels are normally always provided because they allow recognition algorithms to be trained and evaluated. Spotting algorithms require temporal segmentation which corresponds to annotate the time at which gestures occur. Finally tracking algorithms require the labeling of the positions of the body-parts of interest in all frames, also called spatial segmentation. Temporal and spatial segmentation may involve several levels of accuracy. Temporal segmentation can be provided as an ordered list of appearance of the gestures or as accurate start and stop timestamps. Similarly, spatial segmentation can be provided as an approximate position of body-parts using bounding boxes or as an accurate position in the 2d/3d space. Bounding boxes are generally used for body-parts detection and segmentation while accurate positions are used to evaluate tracking algorithms. This categorization appears on both tables; it is represented in Table 1 by the bolded terms in the description of the main purposes of the datasets and can be inferred from the three types of ground truth shown in Table 2. Temporal segmentation is provided for most of the datasets; although several datasets only provide the gesture ordering. Spatial segmentation is rarely provided and when provided, it is generally only for a small percentage of the data. The 6DMG dataset provides an accurate spatial segmentation which has been acquired using an optical tracker. This approach is generally not considered valid when acquired concurrently with video streams due to visual artifacts on the images resulting from markers attached to the subject.

4 Discussion

The number of datasets released in the domain of gesture recognition has largely increased these last years, simultaneously with the regain of interest for human gesture recognition. The transition from color cameras and stereo cameras to single multimodal sensors capable of providing color and depth images and body-joint position is clearly visible in Table 2. Although the number of citations may seem a good indication of the popularity of a dataset most of the reviewed papers introducing a dataset are focused on novel recognition algorithms rather than the dataset itself. This tends to bias the number of citations about the dataset itself. Most of the reviewed datasets have been developed to explore one or several specific contexts; general-interest datasets have currently not been explored. These contexts can concern the type of gestures involved: gaming, iconic, metaphoric, deictic or sign language; the types of algorithms that can be applied: static or dynamic gesture recognition, one-shot learning, spotting, body-part segmentation or tracking or the type of input data: implicit or explicit, depth, color, body-joint position or acceleration data. The type of ground truth available for each dataset is related to the intended algorithm(s) and on the available "man-power" dedicated to manage the dataset. Indeed, ground truthing of datasets remains problematic. In theory and practice, a dataset is considered better if it contains more annotations. However, the ground truthing task is generally performed manually by one or more expert annotators and may consume a lot of time and/or money depending on the amount of data to annotate and the precision level of the desired annotations. Some automatic, semi-automatic and crowd-sourced systems and methods are being explored to solve this problem; however first results tend to show problems in accuracy [34]. Notably, temporal segmentation and spatial segmentation of body-parts can be particularly costly to provide. Note that accurate spatial segmentation can be provided automatically using expensive and cumbersome motion capture systems at the cost of visual artifacts in video streams. The Skeleton data from the Kinect has been used and considered as a marker-free tracking system in a research paper based on MSRC-12. Although this can be valid for an approximate study of motion [19], the problems of accuracy and lost-of-tracking should not be neglected when evaluating tracking algorithms.

Another interesting and surprising information than can be observed from the reviewed datasets is the limited number of multi-sensors datasets; only two datasets contains multiple sensors: 6DMG contains inertial and motion capture data thus providing both implicit and explicit data. Similarly, the ChAirGest contains data from two popular sensors (Kinect and IMU). Although having multiple sensors may require more development on the acquisition software and complexify the acquisition procedure, the added value to the dataset can be worth it and may lead to innovative research directions [35]. The Kinect sensor is a multimodal device in itself as it provides image, depth, approximate body-joints positions and sound which greatly reduces problems of synchronization between sensors. Additionally, comparison methods for the performance of algorithms based on multimodal data must be carefully designed and defined. A discussable example is the ChaLearn 2013 challenge, which was relying on all modalities provided by a Kinect sensor. The best results of the challenge have been obtained by algorithms relying mostly on speech although the task was to recognize the gestures [22]. Even if this is not incorrect, it illustrates the importance of producing well designed vocabularies, datasets and tasks in order to prevent such shortcuts. Multimodal datasets also provide a way to prove quantitatively that some technologies, sensors, data or algorithms may be better suited for recognition than others depending on the conditions. Multimodal datasets for gesture recognition enable researchers to perform quantitative comparisons of modalities and combination of modalities on common data.

Most of the reviewed datasets have been first developed for internal projects and then released publicly. However datasets specifically and carefully designed for benchmarking and comparisons purposes gain more and more interest in the research community. This interest promotes challenges and workshops organized around datasets. Indeed, challenges provide a few advantages such as ensuring that participants can compare their results with a guarantee of validity and fairness and incentive for researchers to compete on similar data and goals.

5 Guidelines

This section contains the guidelines that have been developed to help researchers during the task of selecting or creating datasets.

Selecting a dataset that fits perfectly your needs is not a trivial task and often implies several considerations. Two approaches are distinguished in this paper: researcher and developer .A researcher usually needs a dataset for the evaluation of a new algorithm in order to prove its validity and performances compared to others. The developer usually needs data to provide a rapid and constant solution for testing and optimizing his platform and existing algorithms during the development phase, before starting the tests in real conditions. The following guidelines have been devised for researchers desiring to find a dataset suiting their needs.

- **Task**: The first selection depends on the task of the intended algorithm (recognition, spotting or tracking). Note that adding the missing ground truth information to a dataset might be feasible in certain cases and would probably be welcomed by any dataset author.
- **Requirements of algorithm**: an algorithm implementation generally relies on specific data and features which may be related to certain types of sensors or data types (body joint, depth information, acceleration, etc.).
- **Situation and interaction setup:** the interaction setup such as the position of video-based (front-view, top-view, etc.) and user conditions (standing, sitting, moving, etc.) must be clearly defined.
- **Types of gestures:** some gestures vocabularies may not be suited for all algorithms. Subtle gestures involving limited motion of hands and finger might yield problems for an algorithm initially intended for full-body gesture recognition.
- **Classes and instances:** a dataset with more classes is usually more interesting at the condition that it has enough instances of each class to train and validate the your algorithms. A dataset with many classes and very few instances is generally not usable for most machine learning algorithms.
- **Practical tests:** Researchers should download, when possible, small portion of the selected datasets and then visualize and test the data to take their final decision.

Once the selection finished, researchers should try to take advantage of all the potential of the dataset. When multiple recording conditions are available, the performances of the algorithm for each available condition should be evaluated. Specific evaluation metrics are often imposed, specifically in challenges; researchers should take this into account during the optimization of their performances. Similarly, challenges generally impose specific recognition task(s), if a developed algorithm does not fit exactly the task; researchers should not hesitate to contact the organizers as some alternative solutions can often be found.

Creating a dataset is also a complex task which involves many hours of work. The researcher creating a dataset should always keep in mind the possibility of releasing the dataset publicly at the end of his work. Indeed the time spent to record a dataset may quickly become very long and the dataset could be valuable to other researchers. The following brief guidelines should give an insight of the main tasks when creating a dataset.

- **Careful design**: The initial design of the dataset is very important. All the desired characteristics and recording conditions should be well defined and thought before starting the implementation. The dataset should aim for novelties compared to existing datasets as previously outlined in this paper.

- **Software development**: Several frameworks provide tools to record simple datasets with standard sensors. For more complex scenarios, specific development is usually required. Several frameworks accept the addition of custom plugins.
- **Acquisition methodology**: the acquisition methodology should be accurately defined simultaneously with the software development. A well-defined methodology simplifies the acquisition process. Consider automatizing all the possible processes such as gathering of subjects data, labeling of conditions or ground truthing.
- **Acquisition**: The acquisition data is a time-consuming process. Before starting real acquisition with subjects, the setup should have been thoroughly tested several times in real conditions to ensure the validity of the final recordings. When possible, acquire the data with the highest possible quality and then convert it to lower quality for public release.
- **Annotation and Verification**: Once the dataset has been recorded, perform manual or automatic annotation and verifications on the data to ensure absence of errors. Finally apply a few well-known algorithms on the dataset before release it in order to provide a baseline to researchers.
- **Documentation**: A good documentation and description of the dataset is important for a public release of the dataset. The acquisition setup and the data should be precisely described.

6 Conclusions

This paper filled a void in the literature by providing a survey of the available datasets for the field of gesture recognition, a sub-domain of human actions and activity recognition. The survey provided a comprehensive description of the main publicly available datasets, exhibiting their characteristics, potential usage and highlighting their strengths and weaknesses through two tables. The categorization of the datasets provided a clear distinction between them. This distinction has been based on the usability of the datasets for the different algorithms involved in the gesture recognition. The survey and discussion also highlighted the current design space of the existing datasets and hinted at potential perspectives and challenges for the future datasets such as multimodal and multi-sensors approaches, automatic ground truthing methods and common standards. The discussion outlined the evolution of gesture recognition datasets and highlighted the importance of the presented characteristics through examples. The lack of documentation and information has also been highlighted as a major problem in most reviewed datasets. Finally, brief guidelines have been provided on the main notions and facts researchers should keep in mind when selecting or creating datasets for research.

Acknowledgements. This research has been supported by the HASLER foundation within the framework of “Living in Smart Environment” project.

References

1. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: 2011 Int. Conf. Comput. Vis., pp. 2556–2563 (2011)
2. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. 117, 633–659 (2013)
3. Computer Vision Index Dataset: `http://riemenschneider.hayko.at/vision/dataset/index.php` (accessed: October 23, 2013)
4. CV Datasets on the web: `http://www.cvpapers.com/datasets.html` (accessed: October 1, 2014)
5. Ruffieux, S., Mugellini, E., Lalanne, D., Khaled, O.A.: FEOGARM : A Framework to Evaluate and Optimize Gesture Acquisition and Recognition Methods. In: Work. Robust Mach. Learn. Tech. Hum. Act. Recognition; Syst. Man Cybern., Anchorage (2011)
6. Ruffieux, S., Lalanne, D., Mugellini, E.: ChAirGest: A Challenge for Multimodal Mid-Air Gesture Recognition for Close HCI. In: Proc. 15th ACM Int. Conf. Multimodal Interact. - ICMI 2013, pp. 483–488. ACM Press, Sydney (2013)
7. Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., Qiu, Y.: Exploring techniques for vision based human activity recognition: methods, systems, and evaluation. Sensors (Basel) 13, 1635–1650 (2013)
8. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. 28, 976–990 (2010)
9. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis. ACM Comput. Surv. 43, 1–43 (2011)
10. Ahad, S., Tan, M.A.R., Kim, J., Ishikawa, H.: Action dataset—A survey. In: 2011 Proc., SICE Annu. Conf. (SICE), pp. 1650–1655 (2011)
11. Andriluka, M., Sigal, L., Black, M.J.: Benchmark Datasets for Pose Estimation and Tracking. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) Vis. Anal. Humans. Springer, London (2011)
12. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev. 37, 311–324 (2007)
13. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. Commun. ACM. 54, 60 (2011)
14. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Trans. on Pattern Anal. Mach. Intell. 19, 677–695 (1997)
15. Hasan, H., Abdul-Kareem, S.: Human–computer interaction using vision-based hand gesture recognition systems: A survey. Neural Comput. Appl. (2013)
16. Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., Ney, H.: Benchmark Databases for Video-Based Automatic Sign Language Recognition. In: Int. Conf. Lang. Resour. Eval., Marrakech, Morocco, pp. 1–6 (2008)
17. Athitsos, V., Wang, H., Stefan, A.: A database-based framework for gesture recognition. Pers. Ubiquitous Comput. 14, 511–526 (2010)
18. Glomb, P., Romaszewski, M., Opozda, S., Sochan, A.: Choosing and Modeling Hand Gesture Database for Natural User Interface. In: Proc. 9th Int. Gesture Work., Athens, Greece, pp. 72–75 (2011)
19. Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI 2012, p. 1737 (2012)

20. ARB Labs: `http://www.arblabs.com/` (accessed: October 23, 2013)
21. Conly, C., Doliotis, P., Jangyodsuk, P., Alonzo, R., Athitsos, V.: Toward a 3D Body Part Detection Video Dataset and Hand Tracking Benchmark Categories and Subject Descriptors. In: Pervasive Technol. Relat. to Assist. Environ. (2013)
22. Escalera, S., Sminchisescu, C., Bowden, R., Sclaroff, S., Gonzàlez, J., Baró, X., et al.: ChaLearn multi-modal gesture recognition 2013. In: Proc. 15th ACM Int. Conf. Multimodal Interact. - ICMI 2013, pp. 365–368. ACM Press, New York (2013)
23. Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th Int. Conf. Comput. Vis., pp. 444–451. IEEE (2009)
24. Bloom, V., Makris, D., Argyriou, V.: G3D: A gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 7–12 (2012)
25. Sadeghipour, A., Morency, L., Kopp, S.: Gesture-based Object Recognition using Histograms of Guiding Strokes. In: Procedings Br. Mach. Vis. Conf. 2012, British Machine Vision Association, pp. 44.1–44.11 (2012)
26. Liu, L., Shao, L.: Learning Discriminative Representations from RGB-D Video Data. In: Proc. Int. Jt. Conf. Artif. Intell. (2013)
27. Chen, M., AlRegib, G.: A new 6d motion gesture database and the benchmark results of feature-based statistical recognition. Emerg. Signal Process, 131–134 (2012)
28. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: Signal Process. Conf., pp. 1975–1979 (2012)
29. Guyon, I., Athitsos, V., Jangyodsuk, P., Hamner, B., Escalante, H.J.: ChaLearn Gesture Challenge: Design and First Results. In: IEEE Conf. Comput. Vis. Pattern Recognit. Work., pp. 1–6. IEEE (2012)
30. Song, Y., Demirdjian, D., Davis, R.: Tracking Body and Hands for Gesture Recognition: NATOPS Aircraft Handling Signals Database. In: Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., pp. 500–506. IEEE, Santa Barbara (2011)
31. Ren, Z., Meng, J., Yuan, J., Zhang, Z.: Robust hand gesture recognition with kinect sensor. In: Proc. 19th ACM Int. Conf. Multimed. - MM 2011, p. 759. ACM Press, New York (2011)
32. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Thangali, A.: The American Sign Language Lexicon Video Dataset. In: 2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work, pp. 1–8. IEEE (2008)
33. Kim, T.-K., Wong, S.-F., Cipolla, R.: Tensor Canonical Correlation Analysis for Action Classification. In: 2007 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–8 (2007)
34. Ruffieux, S., Lalanne, D., Mugellini, E., Abou Khaled, O.: Gesture Recognition Corpora and Tools: A Scripted Ground Truthing Method, Publ. Submitt. to J. Comput. Vis. Image Underst. (2014)
35. Banos, O., Calatroni, A., Damas, M., Pomares, H., Rojas, I., Sagha, H., et al.: Kinect=IMU? Learning MIMO Signal Mappings to Automatically Translate Activity Recognition Systems across Sensor Modalities. In: IEEE 2012 16th Int. Symp. Wearable Comput., pp. 92–99 (2012)

Speech, Natural Language and Conversational Interfaces

Accessing Cause-Result Relation and Diplomatic Information in Ancient “Journalistic” Texts with Universal Words

Christina Alexandris

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

Abstract. For the International Public, ancient historical and “journalistic” texts, such the “Peloponnesian War” of the Ancient Greek historian Thucydides, may allow an insight for the understanding of current international political and economic relations. The present approach targets to facilitate the accessibility of such texts for non-experts in the International Public, with no knowledge of the ancient language concerned, especially journalists, translators and students. The possibility of directly accessing text content and viewing features, as close as possible to the original text is attempted to be achieved here, using predefined sublanguage-specific keywords and Universal Words.

Keywords: Ancient Greek, keyword ontology, Universal Words, International Public, online Machine Translation.

1 Introduction and User Requirements

For the International Public, ancient historical and “journalistic” texts, such the “Peloponnesian War” of the Ancient Greek historian Thucydides, may allow an insight for the understanding of current international and national political affairs and international political and economic relations. However, the content of original ancient historical and “journalistic” texts is typically accessible to scholars and other categories of experts, requiring specialized knowledge and command of the ancient language concerned. The present approach targets to facilitate the accessibility of such texts for non-experts in the International Public, especially journalists, translators and students, taking into account basic problems clustered around User Requirements [10]. The possibility of directly accessing the content of these texts and to view features as close as possible to the original text is attempted to be achieved here. Queries based on conventional information extraction strategies may be successful in retrieving concrete information within the “Who/What-When-Where-(How)” framework [4] [6], such as names, events (for instance, battles or treaties) and places. However, diplomacy, especially international politics and bilateral relations of the past, mentalities and attitudes of politicians and military men of the past, as well as reactions of citizens to policies in the past constitute complex information that is difficult to be automatically retrieved (Problem 1). An additional problem is that some

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 351–361, 2014.

information may be partially or wholly omitted in one target language but may be successfully retained and transferred in another target language, mostly due to the linguistic parameters of the language concerned, but also due to the individual style of the translator (Problem 2). This is of special importance in cases where languages from diverse language families are concerned, such as English, Chinese and Ancient Greek. An additional problem of the texts of Thucydides, as well as other Ancient Greek texts, is the extensive use of pronouns and other forms of anaphora and context-dependent expressions, which pose difficulties for the direct access to information with conventional information extraction methods (Problem 3).

Specifically, the basic issue to be addressed here is the possibility to access complex information in the Ancient Text related to diplomacy and to compare it to passages from online journalistic texts (1) and to directly find out respective passages in the original texts along with a translation in English (2) as well as a second type of translation containing structures close to the original text, minimizing language-specific interference and parameters of translations (3). The latter possibility (3) provides a closer look to the content and structure of the original text and is less dependent on language-specific parameters interfering in the English translation.

The present approach concerns the integration of expert knowledge within a System-controlled framework for the detection of information concerning diplomacy, especially cause and result relations contained in the online Ancient Text. The module presented here is designed to make use of already-existing tools and mechanisms, the construction of a database and interface with low computational cost, combined with expert knowledge and sublanguage – specific parameters. For the handling of topics related to complex information such as "Diplomacy", expert knowledge and sublanguage – specific parameters are put to use to constitute a framework replacing conventional information extraction methods and statistically-based approaches [1].

2 Design and User Interaction

2.1 Overview and Design

The proposed approach is designed to work within a framework of a partially implemented interface and database, intending to respond to queries regarding diplomatic and political problems, their resolution, correct or bad decisions, mistakes and socio-cultural phenomena related to politics. The proposed interface and database is specially constructed in respect to the sublanguage of the texts of Thucydides "Peloponnesian War" and is linked to the available translations in English and in formal Modern Greek or "Katharevousa", a "compromise" between Ancient Greek and the Modern Greek used in literature and official documents, especially before the 1980's. In the present approach, the "Katharevousa" translation plays the role of the so-called "Buffer" translation. Specifically, these translations are the English translation by Welsh writer Richard Crawley (1840-1893) [11] [13] and the "Buffer" translation, namely the translation in Katharevousa Greek by Eleftherios Venizelos [9] [12] converted in English by Google Translate.

The approach proposed here is based on a set of ontologies interacting with the two corpora, namely the English and Katharevousa Greek translation. These ontologies are sublanguage-based and aided with the additional use of “Universal Words” [7] [8] [14], used within the UNL framework of the United Nations Research project (The UNDL Foundation). The use of “Universal Words” reinforces the access to the multiple keyword search for the International Public, due to the fact that the “Universal Words” are based on a strictly language-independent structure, enabling the processing of languages as diverse as, for example, Chinese, Arabic, Hindi, Japanese, Russian, German, English, French, Portuguese and Greek. The use of “Universal Words” is proposed as an option for the International Users to enable the use of the sublanguage-based ontology and interface in their own native language.

The proposed and partially implemented module may be characterized by a minimal requirement of tools. Specifically, it makes use of the following online tools: (1) the text corpora of the Portal of the Ancient Greek language, of the Institute of the Greek language in Thessaloniki [12], (2) the available online translations of Richard Crawley (English) in websites such as the Internet Classics Archive of the Massachusetts Institute of Technology [13], (3) the online Machine Translation System: Google Translate, as well as (4) “Universal Words” [14].

The present approach concerns a combination of two search mechanisms: the “Buffer” translation and the multiple keyword ontology [1]. The first search mechanism is based on the alignment of the original Ancient Text and the English and Katharevousa Greek (“Buffer”) translations, where numbered text passages (average length: 5 – 10 lines) act as pointers to text content.

The second search mechanism is based on the use of multiple keywords derived from both types of corpora, the English translation and the Katharevousa Greek (“Buffer”). The keyword types are English, chosen according to the features of the sublanguage of politics and diplomacy. Specifically, the keywords include proper nouns (including names of persons and places), sublanguage-related terms related to “Facts” (for example “battle”, “treaty”) as well as sublanguage-related expressions with specific features related to the notion of “Diplomacy”. The searched elements also include a group of specified conjunctions constituting pointers to causal relations, such as “because”, “due” and sentence-initial “for”, as well as conjunctions describing their connection to them, such as “and” and “or”.

The search mechanisms presented here are specially engineered for complex types of queries in the sublanguage of politics and diplomacy, since search in respect to simple queries such as basic facts may be restricted only a (English) translation. For complex queries in respect to information related to “Diplomacy”, the search is performed on both translations (corpora). In particular, the searched elements are detected at text passage level. When the passage containing the searched elements is detected, the User is presented with the (numbered) passage of the English translation and the Katharevousa “Buffer” translation, with the numbered corresponding passage of the Ancient Text. In particular, the search for the keywords and the respective pointers to Cause-Result relations is designed to be based on maximally four elements (sublanguage-specific keywords and UWs) (Figure 1), combined with a Cause-Result conjunction. The multiple words may be retrieved by available software or programmed with the use of the "grep" function [2]. Since the formulation of these concepts may not be easy to access, the dialog box of the interface assists the User to choose the keywords from a list, corresponding to a multiple keyword ontology presented here. The keyword database concerns the multiple keywords derived from the two corpora, the translated English text [11] and the processed Katharevousa Greek "Buffer" translation translation [9]. The sets of keywords are derived from the study of the sublanguage of Thucydides text. The maximal size of the database constructed on ontological principles is designed to comprise about 300 entries.

```
                    FACT /DIPLOMACY (max. 2 keywords/UWs)
SEARCH :            conj (CAUSE) {+ "and"/"or"}
(5 – 10 lines)      FACT /DIPLOMACY (max. 2 keywords/UWs)
```

Fig. 1. Search mechanism with multiple keywords

2.2 User Interaction

User interaction may be described in two basic steps and an optional step. Specifically, the User is presented with the following features linked to the (a) one or more online journalistic texts obtained from the websites of the news networks: (b) the original Ancient Text, (c) the translations of the original Ancient Text and (d) an interface with queries presenting possible choices to the User. In the first step, the User reads the online journalistic text or texts (a), for instance, in English. The User selects the group of words related to the online journalistic text and types them in the interface of the module (d). Search is performed in respect to passages in the texts containing the search words in the two translations. The words from the User's query are matched to the respective word-group of a keyword database, a multiple keyword ontology constructed with keywords from the English translation and the "Buffer" translation. If there is no match, the module becomes interactive and presents all alternative options (System: "Please select from the following list of words, which best describes your query"). For the International Users who wish to use the sublanguage-based ontology and interface in their own native language, the Universal Words presented in the interface may be used as a stepping stone to access the multiple keywords related to "Diplomacy" and connected to Cause-Result relations and respective passages in the translated and original Ancient Texts.

In the second step, the module presents the passages of the Ancient Text (b) appearing next to the online journalistic text(s). The Ancient Text is presented with the respective English translations (c), namely the available online English [13] translation and the English conversion of the "Buffer" translation of Katharevousa

Greek, which may contain additional elements not present in the available online English translation, allowing an approximate evaluation of the content from both translations or possible comparison to translations in a third language, other than English. In the optional third step, the User may view the original and the partially edited Greek translation in Katharevousa Greek.

3 Search Mechanisms

3.1 Corpora Alignment

The first search mechanism concerns an approach including the so-called "Buffer" translation in Katharevousa Greek by the prominent Greek statesman and political leader Eleftherios Venizelos (1864-1936), published in 1940 in the University of Oxford, after his death [9]. The translation is very close to the original Ancient Greek text, however, it explicitly presents most of the information implied by pronouns and other forms of anaphora and context-dependent expressions in the original Ancient Greek text. Thus, the translation by Eleftherios Venizelos [9], provided online by the Portal of the Ancient Greek language, of the Institute of the Greek language in Thessaloniki, Greece [12], is the corpus on which the multiple keyword-based database and link to the English translation is based. It should be stressed that due to the fact that the available translation in Katharevousa Greek minimizes (but does not eliminate) the extensive use of pronouns, other forms of anaphora and context-dependent expressions, it facilitates the direct access to the text content with the use of the sublanguage-specific keywords related to "Diplomacy". We note that in the "Buffer" translation, more causal relations are visible with pointers such as "due" (Figure 2), which might not be available in an original English translation.

For the quick and efficient access to the translation, the translation in English [11] was numbered and matched to the pre-existing number of each corresponding passage in the "Buffer" [9] translation and the linked Ancient Text. The same number is given to the corresponding English translation. This task was a manually performed process requiring a command of English and knowledge of the Katharevousa Greek language. It should be stressed that in the Portal of the Ancient Greek language, the available translation by Eleftherios Venizelos presents marked passages with the same number as the corresponding passages in the original Ancient Text.The aligned English Text and the "Buffer" translation concern the first search mechanism, where search is performed at a numbered text passage level. For the possibility to be processed by Google Translate, the "Buffer" translation was submitted to minimal necessary processing, namely, a partial editing with a simple "replace" function [1] as well as a pre-translation/default correction of selected words related to the "Diplomacy" word group presented below. The editor also starts a new segment with the negation "den" ("δεν"), due to observed translation errors by Google Translate.

English Translation: (Crawley, 1903)
[6.24.1] With this Nicias concluded, thinking that he should either disgust the Athenians by the magnitude of the undertaking, or, if obliged to sail on the expedition, would thus do so in the safest way possible.
Google-Translate (Katharevousa text)
[6.24.1] That said Nicias, thinking that due to the number of necessary supplies will either prevent the Athenians from the campaign or, if forced to go to war would sail with the utmost safety.
Minimally preprocessed Katharevousa text
[6.24.1] Τούτο είπε ο Νικίας, νομίζοντας, ότι λόγω του πλήθους των αναγκαίων εφοδίων είτε θα απέτρεπε τους Αθηναίους από την εκστρατεία, είτε, εάν αναγκάζεται να εκστρατεύση θα έπλεε με την μεγαλύτερη δυνατή ασφάλεια.
Katharevousa Translation:
[6.24.1] Ταύτα είπεν ο Νικίας, νομίζων, ότι δια του πλήθους των αναγκαίων εφοδίων ή θα απέτρεπε τους Αθηναίους από την εκστρατείαν, ή, εάν ηναγκάζετο να εκστρατεύση, θα εξέπλεε με την μεγαλητέραν δυνατήν ασφάλειαν.
Original Ancient Greek Text:
[6.24.1] Ὁ μὲν Νικίας τοσαῦτα εἶπε νομίζων τοὺς Ἀθηναίους τῷ πλήθει τῶν πραγμάτων ἢ ἀποτρέψειν ἤ, εἰ ἀναγκάζοιτο στρατεύεσθαι, μάλιστ’ <ἂν> οὕτως ἀσφαλῶς ἐκπλεῦσαι·

Fig. 2. Aligned Ancient Greek Text and Translations

3.2 Multiple Keyword Ontology

The multiple keyword ontology (max. size 300 words) may be divided into two basic categories: (A) Facts and (B) Diplomacy. The category of Facts contains a small group of sublanguage-related keywords related to events such as the concept “war”, “battle”, “event” ,”incident”, “treaty”, “ally” “side” and “speech” (approximately 40 words) as well as an open list of proper names, easily accessed with the conventional search using names of people, places and dates [1]. The “Facts” and “Diplomacy” word group are both connected to the Universal Word Framework, whose use is proposed connect sublanguage-specific concepts from both ontologies to each other. The search performed includes the relation of two or more of these concepts in respect to the Facts (“Subject”) word group. Two additional word groups are signalized, not belonging to the keyword ontology, namely (a) a set of function words expressing “Cause-Result” relations (conjunctions and adverbials) connecting the “Facts” and/or “Diplomacy” words to each other and (b) a set of words expressing quantity and quality, such as “many”, “few”, “good”, “bad”, “large” and “small”.

To access and to cover the most commonly occurring types of information related to the subject field of “Diplomacy and other forms of explicitation [3] [5] in the Ancient Text, the category of “Diplomacy” is designed to contain keywords clustered around the concept (Category) (1) state, (2) action and (3) result.The concept of “state” (Category “state”) contains singular words or expressions such as “neutrality” or “disadvantage”.The concept of “actions” (Category “actions”) contains expressions such as “response”-”reaction”-”answer” or “accept”, “accept”, “reject”, and “follow”. The concept of “result” (Category “result”) contains expressions such as “gain”-”benefit”-”profit” or “loss”. The set of verbs contained in this word group are the verbs related to the concepts of feelings and perception “believed”, “hoped”, “saw” and “feared” (Figure 3). The multiple keyword ontology is enriched with a set of Universal Words, connected to the multiple keywords derived from the two corpora, the translated English text [11] [13] and the processed Greek translation [9]. The Universal Words provide an additional and more language-independent access point to the texts for the International Public. Universal Words may coincide with the multiple “Facts” and “Diplomacy” keywords from the sublanguage (Figure 3).

UW Example 1

(i) [ακολουθεί] ("follow")

UW Example 2

(i) [πιστεύει] ("believe")
(ii) [βλέπει] ("see")
(iii) [ελπίζει] ("hope")

UW Example 3

(i) [δέχεται] ("accept")
(ii) [απόρριψη] ("rejection")

UW Example 4

(i) [αποτέλεσμα] ("result"), [αντίδραση] ("reaction")
(ii) [απαντούν] ("answer") [απάντηση] ("answer")

UW Example 5

(i) [κερδίζουν] ("gain")
(ii) [επωφελούνται]("benefit)
(iii) [κέρδους] ("profit")
(iv) [καλό] ("good")

Fig. 3. Greek entries for Universal Words

Relation to UWs 1.1

i. [συνθήκη](“treaty”): agreement
ii. [συμφωνία](“agreement”): promise

Relation to UWs 1.2

i. [ομιλία](“speech”):activity
ii. [ομιλία](speech): information
iii. [γλώσσα] (“speech”): talk
iv. [γλώσσα](“speech”):word

Relation to UWs 1.3

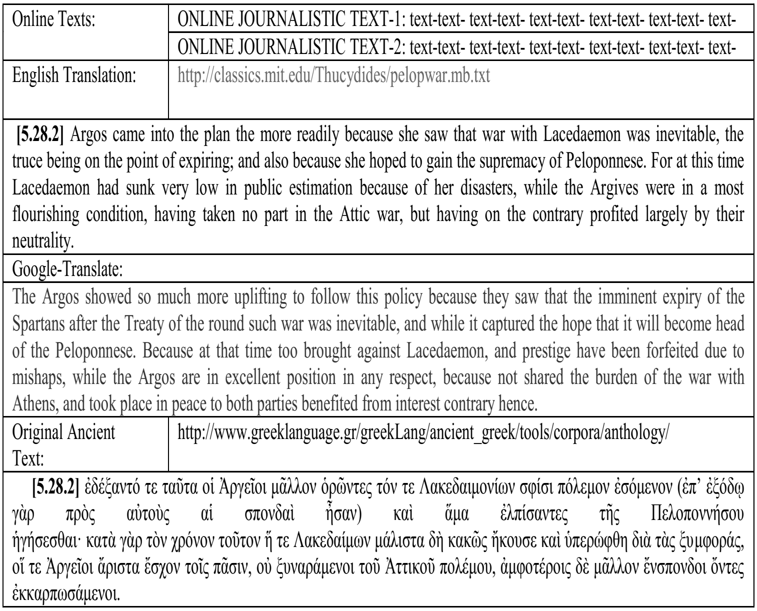
i. [πλευρά](“side”): aspect
ii. [πλευρά](“side”): attitude
iii. [πλευρά](“side”):opinion
iv. [πλευρά](“side”): place
v. [πλευρά](“side”):position

Fig. 4. Relation of Universal Words with Facts and Diplomacy Word Groups

The Facts and Diplomacy word groups are compatible with the Universal Word Framework, whose use is proposed to connect sublanguage-specific concepts from both ontologies to each other. Therefore, there is a connection between a general (or universal) ontology, in this case, the Universal Word framework, and a sublanguage-specific ontology. Universal Words are designed to be compatible with a number of languages, aiming to provide a language-independent analytical framework. This framework is especially helpful in languages with a remarkable polysemy of commonly occurring concepts, such as Ancient Greek, since it allows a concept in the form of a Universal Word to be directly connected to the ontology designed for the sublanguage. The Universal Word framework concerns concepts already connected to each other, such as "treaty" classified as "agreement", also related to the concept "promise", the concept of "speech" connected to "activity" and "information", as well as the concept of "side" connected to "opinion", "attitude" and "aspect", "place" and "position" (Figure 4). Examples of such connected concepts within the sublanguage-specific "Diplomacy" word group are the relations of such as "believe" connected to "hope", "reaction" connected to "response", "benefit" connected to "useful" and "loss" connected to "disadvantage" (Figure 3). Other examples within the Universal Word framework are "fear" classified as "feeling", "answer" classified as "information" and "follow" containing the concept "watch" in its encoding.

4 Examples of Accessing Cause-Result Relations

In the following examples of accessing Cause-Result relations we note that the online journalistic texts accessed from the international news networks are not presented here, to avoid any connection to sensitive political issues in international affairs. The User may wish to acquire information in respect to various queries, for example, the concepts "benefit from neutrality" or "hope to become leader" or "change sides". In the "multiple ontology" from both corpora (translations) and the UWs, there is a direct match to the User's query, for example, in passage 5.28.2 (Figure 5) in respect to the words "neutrality" to the Category "state" and "benefit" related to the Category "result".

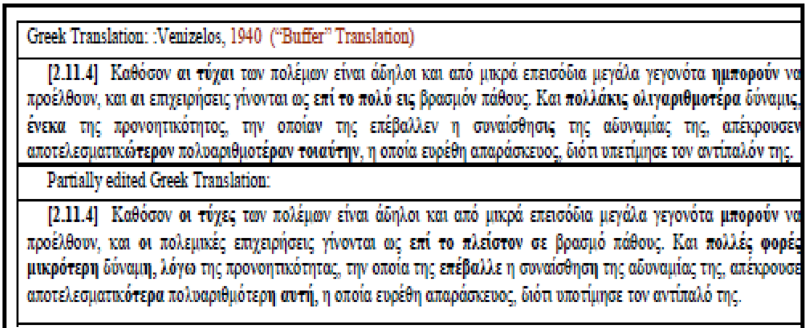


<table>
<tr><td rowspan="2">Online Texts:</td><td>ONLINE JOURNALISTIC TEXT-1: text-text- text-text- text-text- text-text- text-text- text-</td></tr>
<tr><td>ONLINE JOURNALISTIC TEXT-2: text-text- text-text- text-text- text-text- text-text- text-</td></tr>
<tr><td>English Translation:</td><td>http://classics.mit.edu/Thucydides/pelopwar.mb.txt</td></tr>
<tr><td colspan="2"><b>[5.28.2]</b> Argos came into the plan the more readily because she saw that war with Lacedaemon was inevitable, the truce being on the point of expiring; and also because she hoped to gain the supremacy of Peloponnese. For at this time Lacedaemon had sunk very low in public estimation because of her disasters, while the Argives were in a most flourishing condition, having taken no part in the Attic war, but having on the contrary profited largely by their neutrality.</td></tr>
<tr><td colspan="2">Google-Translate:</td></tr>
<tr><td colspan="2">The Argos showed so much more uplifting to follow this policy because they saw that the imminent expiry of the Spartans after the Treaty of the round such war was inevitable, and while it captured the hope that it will become head of the Peloponnese. Because at that time too brought against Lacedaemon, and prestige have been forfeited due to mishaps, while the Argos are in excellent position in any respect, because not shared the burden of the war with Athens, and took place in peace to both parties benefited from interest contrary hence.</td></tr>
<tr><td>Original Ancient Text:</td><td>http://www.greeklanguage.gr/greekLang/ancient_greek/tools/corpora/anthology/</td></tr>
<tr><td colspan="2"><b>[5.28.2]</b> ἐδέξαντό τε ταῦτα οἱ Ἀργεῖοι μᾶλλον ὁρῶντες τόν τε Λακεδαιμονίων σφίσι πόλεμον ἐσόμενον (ἐπ' ἐξόδῳ γὰρ πρὸς αὐτοὺς αἱ σπονδαὶ ἦσαν) καὶ ἅμα ἐλπίσαντες τῆς Πελοποννήσου ἡγήσεσθαι· κατὰ γὰρ τὸν χρόνον τοῦτον ἥ τε Λακεδαίμων μάλιστα δὴ κακῶς ἤκουσε καὶ ὑπερώφθη διὰ τὰς ξυμφοράς, οἵ τε Ἀργεῖοι ἄριστα ἔσχον τοῖς πᾶσιν, οὐ ξυναράμενοι τοῦ Ἀττικοῦ πολέμου, ἀμφοτέροις δὲ μᾶλλον ἔνσπονδοι ὄντες ἐκκαρπωσάμενοι.</td></tr>
</table>

Fig. 5. Interface with retrieved passage and translations from query

A related word to the Category "result" is the word "interest". The concept "hope to become leader" is matched to the verb "hope" and the sublanguage-related keyword "leader". A related word is the word "supremacy" (Category "state"). The words "condition", "position" and "prestige" in the passage accessed are also included in the Category "state". To indicate Cause-Result relations, the "for" and "because" causal conjunctions are signalized, appearing in passages of both translations. However, certain User queries cannot be directly matched to the keywords. In the present case, there is no direct match to the concept "change sides", since this type of expression was not typical of Thucydides. In this case, the search process becomes more interactive. The System proposes the concept matched to the keyword "side" in the Category "state" of the keyword ontology. The System also presents the list with the closest matching multiple keywords. The User chooses the word "follow" from the Category "action" and the sublanguage-related keyword "side". However, no match is found. A match in the same passage 5.28.2 is achieved with keywords "follow" and "policy" (Figure 5).

In Figure 6, the User may wish to acquire information in respect to the query "outcome of war is unpredictable", "small unrests may lead to major outbreaks" or "failure due to underestimating smaller opponent" or "lack of preparation". The concepts are matched, for example, in the passage 2.11.4 (Figure 6) in respect to the keyword "war" and the concepts "outcome" ("fortune", "unpredictable" in Category



Greek Translation: :Venizelos, 1940 ("Buffer" Translation)

[2.11.4] Καθόσον αι **τύχαι** των πολέμων είναι άδηλοι και από μικρά επεισόδια μεγάλα γεγονότα **ημπορούν** να προέλθουν, και αι επιχειρήσεις γίνονται ως **επί το πολύ εις** βρασμόν πάθους. Και **πολλάκις ολιγαριθμοτέρα** δύναμις, **ένεκα** της προνοητικότητος, την οποίαν της επέβαλλεν η συναίσθησις της αδυναμίας της, απέκρουσεν αποτελεσματικ**ώτερον** πολυαριθμοτέραν **τοιαύτην**, η οποία ευρέθη απαράσκευος, διότι υπετίμησε τον αντίπαλόν της.

Partially edited Greek Translation:

[2.11.4] Καθόσον **οι τύχες** των πολέμων είναι άδηλοι και από μικρά επεισόδια μεγάλα γεγονότα **μπορούν** να προέλθουν, και οι πολεμικές επιχειρήσεις γίνονται ως **επί το πλείστον σε** βρασμό πάθους. Και **πολλές φορές μικρότερη** δύναμη, **λόγω** της προνοητικότητας, την οποία της **επέβαλλε** η συναίσθηση της αδυναμίας της, απέκρουσε αποτελεσματικ**ότερα** πολυαριθμότερη **αυτή**, η οποία ευρέθη απαράσκευος, διότι υποτίμησε τον αντίπαλό της.

"The course of war cannot be foreseen, and its attacks are generally dictated by the impulse of the moment; and where overweening self-confidence has despised preparation, a wise apprehension often been able to make head against superior numbers". [English (Crawley) Translation]

"**As** the fortunes of war are hidden and small events big events can be generated, and military operations are mostly in a boiling passion. ***And*** many times less power, because of anticipation, of which impose a sense of failure, effectively repelled more numerous, which was found unprepared **because** they underestimated her opponent" ["Buffer" Translation- Google Translate]

Fig. 6. Retrieved passage and translations from query

"result"), and the verbs "predict" and "forsee" retrieved from both corpora. The second query is matched to the sublanguage-related term "unrest" related to the keyword "event", as well as the words "small" and "major" related to quantity and quality. For the third query, there is an additional match in passage 2.11.4 to the words "confidence" and "unprepared" (Category "state"), the word "failure" (Category "Result"), as well the verb "underestimate". The Cause-Result relations are signalized by the causal "as" and the "because" conjunctions, appearing in passages of the "Buffer" translation, connected to each other with the "and" conjunction.

5 Conclusions and Further Research

Although the proposed module may not provide an in-depth analysis of the Thucydides text, it intends to capture the most commonly occurring categories of diplomatic information and to provide access related to most types of information related to Cause-Result relations. Expert knowledge of the translations provided by R. Crawley (a writer), and E. Venizelos (a politician) is provided, as well as knowledge of Katharevousa Greek texts, containing elements close to Ancient Greek. Thus, complex information such as diplomacy is handled both by expert knowledge and sublanguage – specific parameters, replacing statistically-based approaches and allowing the use of already-existing tools, a database and interface with low computational cost. Further research and full implementation by User groups may provide more upgraded versions of the present design and evaluation results for further development. We note that the tool could have an even simpler configuration if online Machine Translation of Katharevousa Greek were available. Furthermore, these specifications may be adapted to the needs of Ancient Texts in other languages.

References

1. Alexandris, C.: User Interface Design for the Interactive Use of Online Journalistic Texts and Ancient 'Journalistic' Texts for the International Public. In: Proceedings of the International Conference on Applied and Theoretical Systems Research (ATISR) 2012, Taipei, Taiwan (2012)
2. Bambenek, J., Klus, A.: Grep Pocket Reference. O'Reilly, Sebastopol (2009)
3. De Silva, R.: Explicitation in Media Translation: English Translation of Japanese Journalistic Texts. In: Proceedings of the 1st Portuguese Translation Conference, Caparica, Portugal (2006)
4. Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., Piao, S.: The METER Corpus: A Corpus for Analysing Journalistic Text Reuse. In: Proceedings of the Corpus Linguistics 2001 Conference, March 29-April 2, Lancaster University, United Kingdom (2001)
5. Hatim, B.: Communication Across Cultures: Translation Theory and Contrastive Text Linguistics. University of Exeter, Exeter (1997)
6. 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)

7. Uchida, H., Zhu, M., Della Senta, T.: Universal Networking Language. The UNDL Foundation, Tokyo (2005)
8. Uchida, H., Zhu, M., Della Senta, T.: The UNL, A Gift for Millennium. The United Nations University, Institute of Advanced Studies UNU/IAS, Tokyo, Japan (1999)
9. Venizelos, E.: Thoukudidou Istoriai: Kata Metaphrasin Eleutheriou Benizelou. Thucydides' History: translated by Eleftherios Venizelos, ed. D. Caclamanos, 2 vols. Oxford University Press, Oxford (1940)
10. Wiegers, K.E.: Software Requirements. Microsoft Press, Redmond (2005)
11. Thucydides' Peloponnesian War, translated by Richard Crawley. J.M. Dent and Co., London (1903)
12. Centre for the Greek language: Portal for the Greek Language: `http://www.greeklanguage.gr`, E.Venizelos Translation [1940] 1960. *Θουκυδίδου στορίαι*. I–II. 2edition, Athens: The Estia Bookstore (1st Edition: Oxford University Press) (in Greek)
13. Internet Classics Archive of the Massachusetts Institute of Technology: `http://classics.mit.edu/Thucydides/pelopwar.mb.txt`
14. The Universal Networking Language: `http://www.undl.org`

Human Factors in the Design of Arabic-Language Interfaces in Assistive Technologies for Learning Difficulties

Sahar Alkhashrami[1], Huda Alghamdi[2], and Areej Al-Wabil[3]

[1] Department of Special Education, College of Education, King Saud University, Saudi Arabia
salkhashrami@ksu.edu.sa
[2] Disability Service Center, King Saud University, Riyadh, Saudi Arabia
hudalghamdi@ksu.edu.sa
[3] Software Engineering Department, College of Computer and Information Sciences,
Software and Knowledge Engineering Research Group, King Saud University, Saudi Arabia
aalwabil@ksu.edu.sa

Abstract. This paper reports on insights gained from collaborations between multi-disciplinary research teams and practitioners in a Disability Service Center in King Saud University (KSU) in Saudi Arabia. Projects were conducted in the context of designing, developing and evaluating different assistive technologies in the university's Software and Knowledge Engineering Research Group. In these projects, methodological considerations have been reported for effectively involving domain specialists in research and development projects for assistive technologies. Subject Matter Experts (SMEs) are often involved in the technology design cycles of these projects in various roles (e.g. design partners, design informants, testers). This paper highlights the human factors relevant for the design and evaluation of interactive systems for SpLDs that were synthesized from these collaborative contexts. We also shed light on issues to consider in the design partnerships between researchers and practitioners for requirements engineering and user acceptance testing phases of system development. Implications for the design and development of systems for SpLDs in other languages and cultural contexts are discussed.

Keywords: SpLD, Learning Difficulty, Dyslexia, Brain-Computer Interaction, BCI, Usability, User Experience, Disability, Attention Deficit Disorder, ADHD, Augmentative and Alternative Communication, AAC, Arabic Interfaces.

1 Introduction

Despite growing awareness of usability and accessibility issues for designing interactive systems for users with Specific Learning Difficulties (SpLDs), designers still face challenges when creating such systems and evaluating the users' experience (UX) with target user populations. One major stumbling block is a lack of understanding about how to effectively gain insights into the users' needs in the contexts-of-use of systems designed for screening, computerized assessment, cognitive training, and

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 362–369, 2014.

learning. In recent years, different methodological considerations have been reported for involving domain experts in research and development projects of assistive technologies in various contexts [1], [5], [10]. Subject Matter Experts (SMEs) are often involved in the technology design cycles in various roles ranging from the role of design partners such as in the dyslexia screening programs described in [2] and the augmentative and alternative communication (AAC) system described in [7]; design informants such as in [3-5], [8-10] and [17]; and participants in usability evaluations and User Acceptance Testing (UAT) as reported in the systems described in [6] and [14].

Several multi-disciplinary projects were conducted in the context of designing, developing and evaluating assistive technologies for people with disabilities in collaborations between a Disability Service Center in King Saud University (KSU) and a multidisciplinary research group, the Software and Knowledge Engineering Research Group [20]. This paper highlights the human factors relevant for the design and evaluation of interactive systems for SpLDs that were synthesized from these collaborative contexts. Emphasis in the joint activities between the disability service center and the research teams is often on interface design considerations for our target user populations, interaction modalities for input and output that match the needs of users with SpLDs, cultural and language considerations for designing Arabic interfaces.

This paper is organized as follows: Section 2 describes the human factors that are relevant to the context of designing interactive systems for users with Specific Learning Difficulties. Section 3 describes the methodological considerations for involving subject matter experts, practitioners, and users in the design cycles of such interactive systems. We also shed light on issues to consider in the design partnership between researchers and practitioners for requirements' engineering and UAT phases of system development. Section 4 concludes with synthesis of our insights from these projects and lines of future work in multidisciplinary partnerships between researchers and practitioners involved in the research, design and development of assistive technologies.

2 Human Factors in Systems Designed for SpLDs

Individuals with Specific Learning Difficulties (SpLDs) can demonstrate a wide range of cognitive and behavioral abilities on a spectrum of difficulty levels. Moreover, there can be considerable variability within different cognitive capabilities of individuals with a specific difficulty. For example, dyslexics are a heterogeneous group and no two dyslexics are alike; a child with dyslexia can be both good at sequencing and weak in phonological processing of written language. Intelligent interactive systems offer a viable mechanism to provide a personalized user experience (UX) and adaptive modes of interaction to support the multitude of individual needs in people with SpLDs. SpLDs offer some specific design challenges such as the need for configurable controls to account for individual differences in target users, interaction modalities, and types of multimedia feedback that match users' abilities.

The development of interactive systems for supporting individuals with SpLDs has progressed along with the success of frameworks for integrating SMEs, practitioners, and users in the User Centered Design Cycles (UCD) of these systems [1], [9-10], [16-17], [19]. User modeling is essential in requirements engineering phases of assistive technologies to understand the perceptual, emotional, physical, and cognitive capabilities of users [21]. Modeling of users is important to identify functional and non-functional requirements, estimate behavior of users, and simulate scenarios of usage in testing phases. Insights from SMEs and representative samples of real users aid in developing accurate user models for specific target user populations.

Software designers and system developers can refer to user models in comparing design alternatives, input modalities, navigation structures, and multimedia presentations of content. As noted by Simpson in [21], user modeling is not intended to eliminate the need to conduct usability evaluations with real users, but it has been shown to effectively reduce the cost and complexity of the design process and accelerate the development and deployment process. Projects described in [8-13] have utilized user modeling for accelerating the software development process. Furthermore, personas have been used in [5] to model users in early phases of the design process for an auditory discrimination software program and these personas were used later in the project (i.e. in testing phases) to guide the UAT sessions with real users who had SpLDs in local school contexts. User models also guided the design and development of projects described in [2-4], [6], and [8] and deployed versions of these systems were tested with real users in collaboration with KSU's DSC. Moreover, heuristic evaluations were conducted with practitioners in the DSC center in iterative cycles of development for the systems described in [2] and [3] with low-fidelity and high-fidelity prototypes of assistive technologies. Figure 1 shows screenshots of these systems that use gaze-based and brainwave interaction methods which need UAT sessions to examine the UX, usability and subjective satisfaction with these emerging technologies.

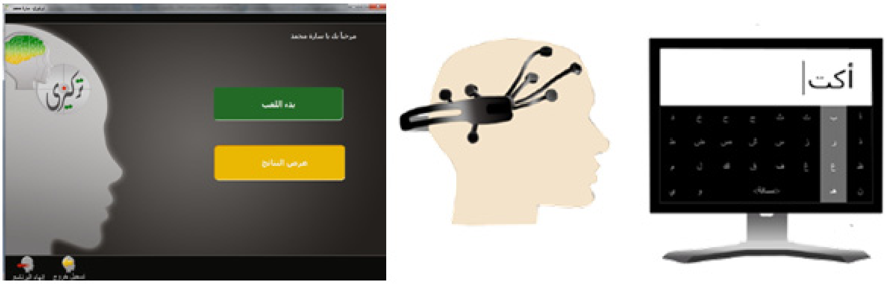


Fig. 1. Interaction modalities of gaze and brain-computer interfaces

A summary of human factors in systems designed with Arabic interfaces, that are characterized with bi-directional interfaces which have right to left text rendering and left-to-right numeric presentation, are listed in Table 1. Projects in which these features were examined in collaboration with SMEs, practitioners, and users are also listed. The human factors that were particularly relevant for these contexts emerged either by the system analysts and designers or were highlighted by the practitioners in heuristic review sessions and focus group meetings.

Table 1. Human Factors in Interactive Systems for People with SpLDs

Human Factor Design Issue	Systems in which issues were examined
Dynamics of pointing and selection	
Default cursor positions	[2], [3] bi-directional interfaces in Arabic
Touch-screen design considerations for children and elderly users	[7], [17], [19]
Psychomotor movements in gaze-based interactive systems	[2], [3], [8], [16]
Psychomotor movements in Brain-Computer Interfaces (BCI)	[4], [18]
Sensitivity in selection modalities	Touch [7] Dwell time in Gaze [2-3], [8] Brainwaves [4], [18]
Text entry design consideration	
Size and resolution of keys	[3-4], [7-8], [17], [19]
Interchangeable layouts of navigation	[9-10] familiarity with existing non-Arabic systems was considered
Prediction	Frequency of use for Arabic letters [4], [6]
Visual design of interfaces	
Cognitive abilities and individual differences	[3], [5], [8-9]
Personalization and gender-specific design	[7], [10], [15] cultural contexts of gender-segregated learning and personalized avatars in interfaces
Configurable Text	Readability of Arabic text in [5] ,[9]
Perception and Interaction	
Multimedia adaptation	Language considerations [9-10], cultural considerations [15-17]
Embedded Arabic speech engines	[7] [10] insights from practitioners on perceived spoken phrases in Arabic

Iterative design cycles have facilitated incorporating these design recommendations in line with the design approach described in [1] and [10]. Different usability protocols [e.g. 14 and 18] have been applied to assess the efficiency, effectiveness, and subjective satisfaction of users in their interaction with such systems.

3 Methodological Considerations in the Interaction Design Process for Assistive Technologies

The use of UCD, ISO 9241 [22], in the design of assistive technologies has been gaining popularity in a variety of systems' development scenarios [2-10]. However, the involvement of practitioners, users, and domain experts in roles such as design-partners and design-informants may not be the optimal if their integration does not take into account the planned activities for different phases of systems' development [1], [10]. For example, in phases of requirements engineering, system analysts need to effectively elicit insights into the needs of target user populations from users and SMEs as described in UCD activities of [2-3] and [5]. Careful planning of UAT phases is needed in collaboration with practitioners and SMEs so that usability engineers can effectively assess the system with representative samples of real users in performance-based evaluations such as sessions described in [6-10] or in heuristic evaluations with SMEs as conducted in KSU's DSC for the projects described in [2] for dyslexia, [3] for attention deficit disorders, and [5] for auditory discrimination therapy.

It is also important to note that limited resources were available that document benchmarks and best practices of collaborations between disability service entities from one side, and assistive technology research and development (R&D) entities in academic and industry contexts from the other side of partnerships in our local context. To address this issue, documentation and reporting of the collaborative projects was conducted with in-depth analysis of the UCD methods and the type of contributions from members involved in both the disability service center and the SKERG research group [20].

Specific activities include briefing and debriefing sessions in which the research teams would conduct walkthrough of the system with members of the disability service center to ensure that sessions are designed to meet the UCD objectives of evaluating design concepts or functionality from different perspectives such as in the design of [2] and [4]. Cultural context was very important in early stages of design and development. For this reason, projects often involve a survey of existing technologies, gaps in addressing the requirements of the local user population, and a critique of the functionality for similar systems designed for non-Arabic-speaking users or designed for different cultural contexts. The survey would highlight design opportunities for adaptation and activities would elicit a critique of alternative design proposals from SMEs and users in brainstorming sessions or in task-based assessment sessions with a specific focus on strengths and weaknesses of users with SpLDs.

4 Conclusion

In this paper, we presented an overview of human factors and methodological considerations for the contexts of assistive technology designed for people with SpLDs. UCD cycles can only be effective if users and SMEs are involved in key phases of the software development cycles in which their contribution is directly related to the functionality being considered and/or examined for users with SpLDs. The level and mode of involvement (e.g. design partners, design informants, testers) also need to be considered in relation to the complexity of the system and constraints of the system development project.

Several successful collaborations, between the Disability Service Center of KSU and multidisciplinary teams in the SKERG research group [21], have demonstrated different approaches in considering partnerships between teams in R&D and practitioner contexts. These collaborations were established with the aim of eliciting insights into the user needs and efficiently evaluating the systems from the perspective of practitioners, domain experts, and real users. Key issues to consider in the design and development of systems for SpLDs that can be generalized to other contexts can be categorized into two areas; namely, human factors and methodological considerations in partnerships between researchers and practitioners for requirements' engineering and user acceptance testing phases of system development. For identifying human factors in interactive systems for users with SpLDs, language and cultural contexts need to be considered in the visual design, presentation, and mode of interactions. For methodological considerations, the level and type of involvement of practitioners, SMEs, and users needs to be determined based on user needs and established within the constraints of the software development project and organizations involved in the context of use. Collaborations need to consider examining the contrast between what has been developed in the scope of assistive technologies for SpLDs in other contexts and existing systems; and aim to identify design opportunities for adaptation, further development, and re-engineering to meet the target user population's requirements.

Future lines of research are planned to examine effective frameworks for collaborations in academic contexts and industry-oriented systems' development for assistive technology.

Acknowledgement. The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding the work described for Brain-Computer Interface projects and AAC through the research group project number RGP-VPP-157**.** Special thanks for all the students, researchers, and assistants involved in these projects; Najwa Alghamdi, Arwa Al-Rubaian, Bayan Alarifi, Ghada Al-Ofisan, Nora Al-Romi, Amani Musrea, Aljawharah Al-Abdullatif, Hadeel Al-Negheimish, Latifah Al-Mofeez, Nuha Al-Khalifa, Lama Al-Andas, Ashwag Alshathri, Arwa Al-Edaily, Arwa Al-Issa, Itisam Hazzaa, May Al-Humaimeedi, Lujain Al-Tamimi, and Bushra Al-Kadhi, Sara Drine, Sara Alkoblan, Arwa Alamoudi, Rawan Al-Abdulrahman, Modhi Almuzainy, and KSU's DSC staff and students.

References

1. Al-Abdulkarim, L., Al-Wabil, A., Al-Yahya, M., Al-Humaimeedy, A., Al-Khudair, S.: Methodological Considerations for Involving SpLD Practitioners in the Design of Interactive Learning Systems. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 1–4. Springer, Heidelberg (2010)
2. Al-Edaily, A., Al-Wabil, A., Al-Ohali, Y.: Dyslexia Explorer: A Screening System for Learning Difficulties in the Arabic Language Using Eye Tracking. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) SouthCHI 2013. LNCS, vol. 7946, pp. 831–834. Springer, Heidelberg (2013)
3. Al-Shathri, A., Al-Wabil, A., Al-Ohali, Y.: Eye-Controlled Games for Behavioral Therapy of Attention Deficit Disorders. In: Stephanidis, C. (ed.) HCII 2013, Part I. CCIS, vol. 373, pp. 574–578. Springer, Heidelberg (2013)

4. Al-Abdullatif, A., Al-Negheimish, H., Al-Mofeez, L., Al-Khalifa, N., Al-Andas, L., Al-Wabil, A.: Mind-Controlled Augmentative and Alternative Communication for People with Severe Motor Disabilities. In: Proceedings of the 9th International Conference on Innovations in Information Technology, UAE. IEEE CPS (2013)
5. Al-Wabil, A., Drine, S., Alkoblan, S., Alamoudi, A., Al-Abdulrahman, R., Almuzainy, M.: The Use of Personas in the Design of an Arabic Auditory Training System for Children. In: Proceedings of the 13th International Conference on Computers Helping People with Special Needs (ICCHP 2012) in the Universal Learning Design (ULD) track, Linz, Austria (July 2012)
6. Al-Wabil, A., Al-Issa, A., Hazzaa, I., Al-Humaimeedi, M., Al-Tamimi, L., Al-Kadhi, B.: Optimizing Gaze Typing for People with Severe Motor Disabilities: The iWriter Arabic Interface. In: Proceedings 14th International ACM SIGACCESS Conf. on Computers and Accessibility (ASSETS 2012), pp. 261–262. ACM, New York (2012)
7. Al-Arifi, B., Al-Rubaian, A., Al-Ofisan, G., Al-Romi, N., Al-Wabil, A.: Towards an Arabic Language Augmentative and Alternative Communication Application for Autism. In: Marcus, A. (ed.) DUXU 2013, Part II. LNCS, vol. 8013, pp. 333–341. Springer, Heidelberg (2013)
8. Al-Wabil, A., Alomar, A., Alhadlaq, K., Alrubayain, M., Alzakari, N., Alhamid, O.: Interactive Therapy of ADHD with Gaze-Based Games. In: Proceedings of the 10th Pacific Conference of Computer Human Interaction (APCHI) in Matsua, Japan. ACM SIGCHI (2012), http://apchi2012.org/catalog_poster/
9. Al-Harbi, O., Al-Arfaj, N., Al-Hathlool, L., Al-Ghofaily, M., Madani, D., Al-Wabil, A.: The Design and Development of an Online Multimedia Language Assistant for Web Users with Dyslexia. In: Proceedings of the 8th International Technology, Education and Development Conference, Valencia, Spain (2014) ISBN: 978-84-616-8411-3
10. Al-Wabil, A., Meldah, E., Al-Suwaidan, A., AlZahrani, A.: Designing Educational Games for Children with Specific Learning Difficulties: Insights from Involving Children and Practitioners. In: Proceedings of the Computing in the Fifth International Multi-Conference on Global Information Technology (ICCGI), pp. 195–198. IEEE CPS (2010), doi:10.1109/ICCGI.2010.43
11. AlGhamdi, N., AlOhali, Y.: Rannan: Computer Based Auditory Training For Arabic-speaking Children. In: Proceedings of the 2010 World Conference on Educational Multimedia, Hypermedia & Telecommunications EdMedia, Toronto, Canada, pp. 3225–3229. AACE (2010)
12. AlGhamdi, N., AlOhali, Y.: The Design and Development of 3D Auditory Environments for Computer-Based Aural Rehabilitation Programs. In: Proceedings of the Fifth International Multi-Conference on Computing in the Global Information Technology (ICCGI), pp. 209–213. IEEE CPS (2010), http://dx.doi.org/10.1109/ICCGI.2010.10
13. Al-Wabil, A., Al-Shabanat, H., Al-Sarrani, R., Al-Khonin, M.: Developing a Multimedia Environment to Aid in Vocalization for People on the Autism Spectrum: A User-Centered Design Approach. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A., et al. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 33–36. Springer, Heidelberg (2010)
14. Al-Wabil, A., Al-Husain, L., Al-Murshad, R., Al-Nafjan, A.: Applying the Retrospective Think-Aloud Protocol in Usability testing with Children: Seeing Through Children's Eyes. In: Proceedings of the 2010 User Science and Engineering Conference iUser, pp. 98–103. IEEE CPS (December 2010)
15. AlSuwaidan, A., AlZahrani, A., Meldah, E., AlNukhilan, H., AlIsmail, S.: Designing Software for Cognitive Training of Children with Learning Difficulties: The Memory Challenge Project. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2010, pp. 737–740. AACE, Chesapeake (2010)

16. Al-Omar, D., Al-Wabil, A., Fawzi, M.: Using Pupil Size Variation during Visual Emotional Stimulation in Measuring Affective States of Non-Communicative Individuals. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2013, Part II. LNCS, vol. 8010, pp. 253–258. Springer, Heidelberg (2013)
17. Al-Muhanna, H., Al-Wabil, R., Al-Mazrua, H., Al-Fadhel, N., Al-Wabil, A.: An Interactive Multimedia System for Monitoring the Progressive Decline of Memory in Alzheimer's Patients. In: Stephanidis, C. (ed.) Posters, Part II, HCII 2011. CCIS, vol. 174, pp. 381–385. Springer, Heidelberg (2011)
18. Alghamdi, N., Alhudhud, G., Alzamel, M., Al-Wabil, A.: Trials and Tribulations of Brain-Computer Interfaces. In: Proceedings of the Science and Information Conference the SAI, London, UK, pp. 212–217. IEEE CPS (2013)
19. Almazrou, H., Alzamel, N., Alwabil, R., Almuhanna, H., Alhinti, L., Al-Wabil, A.: Technology for Psychosocial Interventions for Individuals with Alzheimer's: Reminiscence Therapy and Monitoring Progressive Decline of Cognitive Abilities. In: Proceedings of the Science and Information Conference the SAI, London, UK, pp. 171–175. IEEE CPS (2013)
20. Software and Knowledge Engineering Research Group SKERG, `http://skerg.ksu.edu.sa`
21. Simpson, R.: Computer Access for People with Disabilities: A Human Factors Approach. CRC Press, Taylor and Francis (2013)
22. User Centered Design Process for Interactive Systems: ISO 9241 (Ergonomics of human-system interaction - Part 210), `http://www.iso.org/iso/home/store/catalogue_ics`

Design and Development of Speech Interaction: A Methodology

Nuno Almeida[1,2], Samuel Silva[1], and António Teixeira[1,2]

[1] Institute of Electronics and Telematics Engineering, University of Aveiro, Portugal
[2] Dep. of Electronics, Telecommunications and Informatics Engineering, University of Aveiro, Portugal

Abstract. Using speech in computer interaction is advantageous in many situation and more natural for the user. However, development of speech enabled applications presents, in general, a big challenge when designing the application, regarding the implementation of speech modalities and what the speech recognizer will understand.

In this paper we present the context of our work, describe the major challenges involved in using speech modalities, summarize our approach to speech interaction design and share experiences regarding our applications, their architecture and gathered insights.

In our approach we use a multimodal framework, responsible for the communication between modalities, and a generic speech modality allowing developers to quickly implement new speech enabled applications.

As part of our methodology, in order to inform development, we consider two different applications, one targeting smartphones and the other tablets or home computers. These adopt a multimodal architecture and provide different scenarios for testing the proposed speech modality.

Keywords: Speech, multimodal architecture, decoupled modalities.

1 Introduction

Speech is, in many situations, the easiest and most natural existing interface to deal with computers, not only for people with special needs, but for people in general [18]. The advantages of speech, as argued by Bernsen [6], are many: a) it is natural and so, people communicate as they normally do; b) it is fast (commonly 150–250 word per minute); c) it requires no visual attention; and d) it does not require the use of hands. Adding to these, one of the characteristics that distinguishes the auditory from the visual channel is its omni directionality, i.e., auditory information can be received from any direction and can also, to some extent, be transmitted in parallel with stimuli from other channels. Furthermore, auditory information, even though it is transient, has a slightly longer short-term storage than visual information which allows delayed processing [20].

Using speech for interaction requires the consideration of different components including speech recognition, text-to-speech, grammar management, a natural language generator and adaptability management, possibly considering multiple

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 370–381, 2014.

languages. Some components are inter-dependent and must communicate between them and with the application. One major challenge is to have a flexible design to enable communication and to support a loosely coupled and distributed architecture, allowing an easy integration with application and devices.

Furthermore, one of the most challenging aspects of speech interaction is dealing with users' expectations, as they often expect speech enabled systems to be capable of understanding much more commands than they actually do.

Using speech as an input/output modality should not be done lightly and the literature provides several guidelines [19,15]) that should be considered, covering when to use speech, what kind of tasks and data are best served by speech, how to combine speech with other modalities and how to address adaptability (e.g., to context). One important aspect to note, for example, is that speech should not be used alone, but as part of a multimodal approach, even though, sometimes, it might be the only useful modality for some users or contexts [21]. This integration with other modalities is also a challenging task [9].

Understanding the full potential of speech as an input/output modality, covering the different guidelines and desirable adaptability features, in different application scenarios, is a complex, multivariate problem which often translates in a considerable development effort.

To tackle these issues we argue that an effort should be made to propose an architecture based on which a generic speech modality, decoupled from any particular application context, can be developed. This generic modality should encapsulate dealing with most of the complexity described above and should provide easier deployment of speech enabled systems.

The work presented in this paper is part of that effort and presents the methodology being followed to design and develop a module that enables speech interaction in applications. This methodology is characterized by the following notable aspects:

– A multimodal framework is considered and implemented;
– The speech modality is first developed as a generic modality and then integrated with the multimodal framework;
– Different application prototypes are used as a testbed, to inform development.

This article is organized as follows: Section 2 briefly presents background and related work; Section 3 describes our work regarding the proposal of a generic multimodal architecture supporting the development of generic modalities focusing the particular case of a generic speech modality; Section 4 presents two prototype applications which are used as part of our design and development pipeline for testing; finally, Section 5 presents some conclusions and ideas for further work.

2 Background and Related Work

Our work is aligned with recent W3C recommendations [10] for multimodal frameworks. This provides the grounds on which modalities are built, such as the speech modality presented in this paper. Therefore, to provide context, we briefly present the overall aspects of the multimodal framework, based on w3C recommendations, followed by an overview of relevant work presented in the literature regarding the use of speech in multimodal scenarios.

2.1 W3C Multimodal Framework

The W3C Recommendation [10] defines the major components of a multimodal system and identifies standard markup languages used to support communication between the components and data modules. The architecture can be divided into four major components (illustrated in Fig. 1):

- **Interaction Manager (IM)** – manages the different modalities. It is similar to the Controller in a Model View Controller (MVC) paradigm;
- **Modality Components** – representing input/output modules;
- **Runtime Framework** – acts as a container for all others, providing communication capabilities;
- **Data Component** – stores the data model.

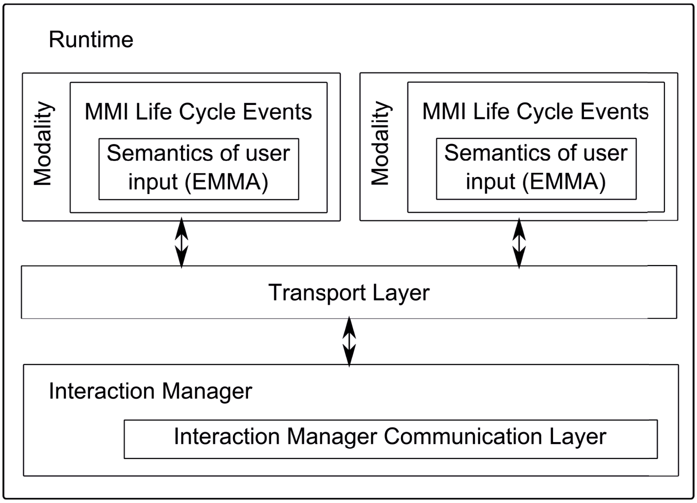

Runtime
Modality
MMI Life Cycle Events
Semantics of user input (EMMA)
Modality
MMI Life Cycle Events
Semantics of user input (EMMA)
Transport Layer
Interaction Manager
Interaction Manager Communication Layer

Fig. 1. The W3C Multimodal Architecture

Communication between Components (MMI Lifecycle Events). All communication is handled by MMI Lifecycle Events, a standard defined in the MMI Architecture. MMI Lifecycle events are messages exchanged between modalities and the Interaction manager, carrying the information of each event. Each message possesses common attributes. A request may possess attributes such as *context*, *source*, *target* or *requestID*. A response possesses attributes such as the *status*. Each MMI Life Cycle Event might also have the element *data* which is optional.

Standard Markup Language to Describe Events (EMMA). Extensible MultiModal Annotation markup language (EMMA) [4] is a standard language to describe events generated by different inputs, to be used within a multimodal system to exchange data information between inputs and multimodal components.

An EMMA document has three types of data:

- Instance data: Application-specific markup corresponding to input information;
- Data model: Constraints on structure and content of an instance;
- Metadata: Annotations associated with the data contained in the instance.

This language has a set of elements and attributes collected from the user's inputs, an *interpretation* element defines the event interpreted by the modality, with parameter such as *begin* and *end* time of the event, *confidence* of the recognition, *medium*, *mode* and recognized data.

SCXML. SCXML [5] is a markup language that defines a state chart machine and a data model. Its objective is to provide the application logics to the existing framework. The basic concepts of a state machine are states, transitions and events. When events occur, the machine tries to match the event to the transitions on the active state. If it matches, the target state is set as the new active state.

In SCXML, there are some extensions to a basic state machine. State machines can have executable content such as conditions, executable scripts, send messages to external entities or modalities and modify the data model. It also has two elements to execute content upon entering or exiting a state.

2.2 Speech for Interaction

Many recent applications using multimodal interaction explore the use of speech. It is one of the commonly present modalities in multimodal systems, appearing as part of the three most popular combinations mentioned by Bui et al. [11] for input: speech and lips movement, speech and gesture (including pen gesture, pointing gesture, human gesture) and speech, gesture, and facial expressions.

Popular combinations of output modalities, which include speech, are [11]: speech and graphics, speech and avatar and speech, text and graphics.

Adopting the definition of modality as "a way of exchanging information between humans [. . .] and machines, in some medium" [9], several "speech modalities" can be considered. In the Bernsen taxonomy three modalities are proposed, at atomic level: spoken discourse, spoken label-keywords and spoken notation [7].

The different Speech related modalities have different characteristics and, therefore, different suitabilities [7]. Spoken discourse is adequate for situated communication with the hearing and involving those who have the skills in interpreting and generating a particular language. It allows exchange of information when painstaking attention to detail is not required. If more complex data needs to be transmitted written language can be a better choice.

Spoken labels/keywords are suitable to convey small, isolated pieces of meaning as long as the context in which they are used helps reduce the inherent ambiguity. Bernsen et al. [8] refers the example of a user navigating a townscape. In that context, spoken words such as "house" or "door" are easily understood.

Spoken notation, might be a good option to convey information in the particular domain it refers too but, as it is often dynamic, it might be quite error prone or difficult to interpret by either human or machine [7] unless it is limited to particular contexts.

Speech is very resilient as a side channel, making it the ideal mode for "secondary task interfaces". These are interfaces for functions when the computational activity is not the primary task (ex: while driving) [13]. Furthermore, as discussed in Teixeira et al. [21], speech should not be used alone, it must be part of a multimodal input/output and, for some users or context of use (ex: mobile phone interaction with hands and eyes busy), will be the only useful modality.

The mTalk [17], developed by AT&T, Ford sync [1], Siri [2] and Xbox One [3] are well known examples of mutltimodal interaction that uses speech as a way to interact with the system, but those systems are commercial and closed solutions.

Mudra [16] and Manitou [14] are other examples of multimodal interaction frameworks that allow speech as a modality in the human-computer interaction. The first aims to process low-level streams and high level semantics and combine those events; the second aims for easy development of multimodal-enabled web applications.

3 Proposed Architecture for Speech Enabled Systems

Analysing existing work, it is important to note that most of the proposed solutions are very application oriented, i.e., the speech modality is developed tightly coupled with the envisaged application and device. As stressed before, we argue that this results in limited reuse of the developed modality, e.g., in a different application, yielding additional development costs and poses barriers, given the complexity of developing a speech modality, to its integration by third parties.

We propose a solution where modalities are decoupled and communicate with the applications through the multimodal framework enabling the reuse of modalities in other applications. Figure 2 illustrates one issue of current solutions and how it works for our proposed solution, namely, in the left we see that common scenarios use speech embedded as a part of the application and it is hard to reuse code to create new applications, on the other hand the desired scenario, on the right, has a speech modality decoupled from the application allowing the reuse of the modality in other applications.

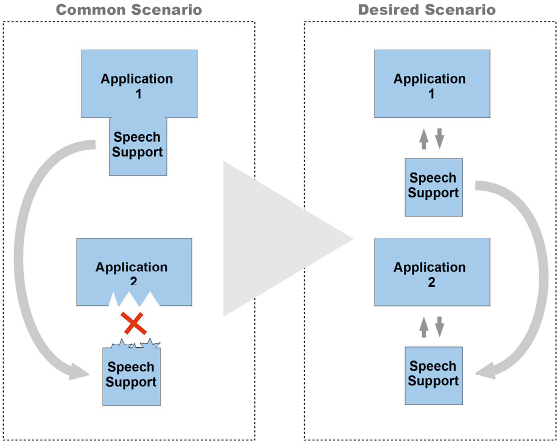


Fig. 2. Decoupled solution for the speech modality

3.1 Multimodal Framework

Our approach for speech enabled applications started by the development of a multimodal framework capable of managing different and generic modalities, supporting communication between modalities and the application.

The multimodal framework is directly based on the recommendations presented by the W3C, Multimodal Interaction (MMI) Architecture [10] and although it is focused on web scenarios, our goal is to extend it for interaction with mobile devices, tablets and AAL applications [23]. This choice is justified by the architecture's open standard nature and provides an answer to a significant part of the requirements presented, easing the creation and integration of new modules, as well as already existing tools.

Our multimodal framework has a main module, the Interaction Manager, which implements a state machine defined in SCXML that controls the flow of messages between modalities. To enable communication, the module implements an HTTP server listening to messages or requests sent by modalities, modalities only have to obey the message protocol in order to communicate with the system.

Therefore, having a standard for multimodal architecture helps application developers to avoid the unpractical situation of having to master each individual modality technology. This is particularly problematic as the number of technologies that can be used with multimodal interaction is increasing very fast. This standard architecture gives experts the possibility to develop standalone components [12] that can be used in a common way.

3.2 The Speech Modality

Considering the multimodal framework recommendations, modalities should be decoupled and communicate with the interaction manager with standard

MMI life cycle events, allowing other developers to focus on coding only the application.

Therefore, the proposed speech modality implements the communication languages described by the W3C architecture and communicates with the Interaction Manager which, in turn, communicates with the application sending the modalities' events.

The development of the speech modality starts with the creation of a generic modality supporting the different speech features required, considering both input and output. This modality is configured with a grammar, containing the possible sentences that the modality can recognize. We have created a tool that enables the translation of the grammars: by processing the grammar it generates all its possible sentences. Then, using translation services available on the web, each sentence is translated for the desired languages. Finally, the grammar is reassembled, creating a new grammar file for each language.

To support both mobile devices and desktop application, the modality has the capacity to process the recognition locally or remotely, enabling its use on mobile devices. When it is remotely, there is a local part of the modality to communicate with the remote part. Using this locally or remotely, does not affect how the framework is integrated. To accomplish this, services were created that process data and can be deployed in different locations (a device or a server).

Speech Recognition. The Asynchronous Speech Recognition (ASR) receives an audio stream with a spoken sentence, and the name of the grammar to be used to recognize the speech.

There are two kinds of grammars: GRXML, which is a W3C standard to specify the words or sentences to be recognized by the ASR, and ARPA, a statistical language model. The first type is more limited regarding the amount of sentences that can be recognized and is manually defined, but can return tags identifying the sentence's meaning. For ARPA, the creation of the grammar is automatic, since it is a statistical language model, but it requires large amounts of text in order to create the model, as well as the mechanisms to extract the meaning of the sentences.

Speech Synthesis. For this part of the service, called Text-to-Speech (TTS), the application sends a message with the information to be read to the user, the method to use to synthesize it to speech, using the Microsoft Speech platform (MSP), and the chosen voice. The service accepts other parameters such as speech volume and rate. The rate parameter defines the speed of the speech. Based on recent experiments in our group the default value chosen for the speed parameter makes the speech understandable for the elderly, and if the value increases, elderly people may have more difficulty in understanding it. The service returns an audio stream containing the spoken sentence.

3.3 Integration in the Multimodal Framework

In the second stage, the generic modality is integrated in a generic distributed multimodal framework, and dealt with as any other modality. Each modality follows the standard messaging specifications.

4 Application Prototypes

Finally, we have used the described multimodal architecture and speech modality to create two different applications, one targeting smartphones and other targeting home computers, with different use-case scenarios. These applications allow us to test and evaluate different aspects of our work informing further improvements to our proposed framework.

These applications, serving real application scenarios, are used as a test bed to improve our understanding of the different aspects involved, support brainstorming and inform development of future speech enabled applications.

Both applications use the Multimodal Framework and methodology previously discussed and each application targets a different device.

4.1 Newsreader

The application is a news reader developed for Windows 8, providing multimodal interaction for enhanced user experience and usability. It starts by loading some RSS news feeds from different sources depending on the users language and displaying the news to the user. At the same time, it processes the news contents to produce a list of headlines that it is used to configure a new grammar in the speech input modality.

An output modality called GUI, used as a part of the application, is continuously listening for messages coming from the Interaction Manager and it is responsible to update the interface of the application showing new content on the screen.

Figure 3 shows the modalities, states of the SCXML and the exchanged MMI Life Cycle events. Each modality, when it starts to run send a *NewContextRequest* to register in the Interaction Manager, it responds with a *NewContextResponse* informing if the registry was successful. After the speech modality recognizes the user sentence, it sends a *DoneNotification* with the event data to the Interaction Manager, which then sends a *StartResponse* to the GUI modality requesting some update in the user interface. The GUI modality replies with a *StartResponse* confirming the operation.

Different input modalities can be used to interact with the application. For instance, if the user wants to slide the container with the list of news, it can be done by any of the input modalities: via Kinect it is possible to swipe a hand to the left or right; Speech allows for actions to be active via words such as “left” or “right”; or Touch.

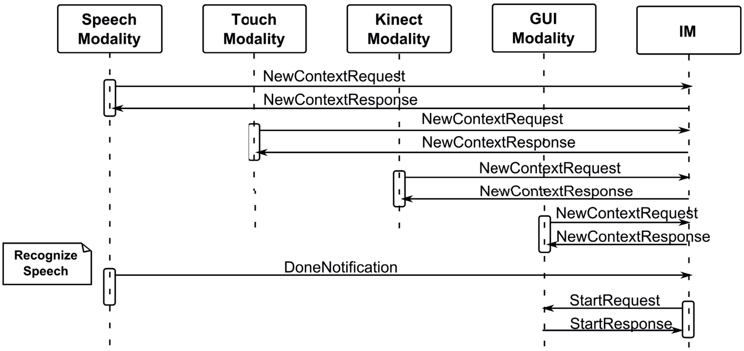


Fig. 3. Messages exchanged between the Interaction Manager (IM) and the modalities

In order for the user to read the entire body of the news, speech or touch can be used to select an article, by reading the headline or tapping the corresponding square.

Figure 4 presents an example of user interactions to read a particular article. The first screen shows the list of news by swiping the hand to the left or speak "left" the content slides to the left, it is shown in the screen in the upper right. Then the user says "Labours reputation at stake" to open the details of that article, as visible in the screenshot at the bottom left. Finally, the user says "go back" to return to the news list.

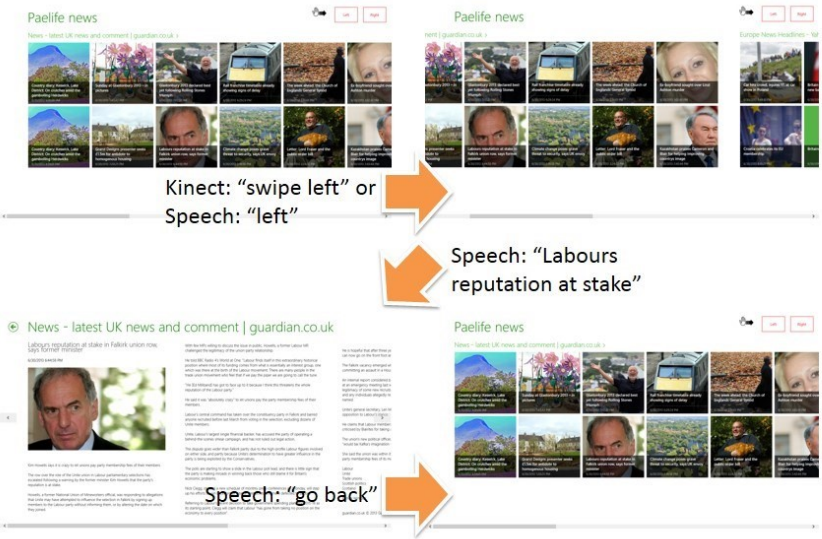
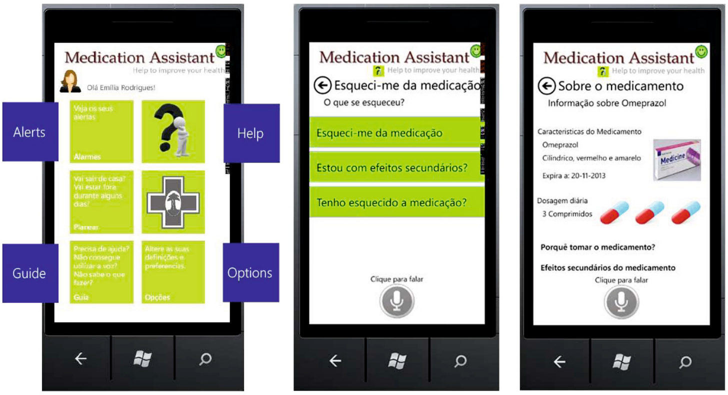


Fig. 4. Screens of the Newsreader application depicting some of the possible interactions

When an event occurs in the speech modality, the modality sends the event data to the IM to be processed. Upon processing it, the IM creates an action to be sent to an output modality, then the output modality presents that information to the user. Having a generic speech modality relieves the developers of having to handle with the recognizer, grammars, etc. In this scenario developers only have to inform about the sentences that can be recognized and a tag for which sentence.

4.2 Medication Assistant

This application, developed for Windows Phone, illustrated in Fig. 5, has two main functionalities: first, generating and showing medication intake alerts and, second, providing advice on how to proceed if the user misses a medication intake [24].



Medication Assistant
Help to improve your health
Olá Emília Rodrigues!
Veja os seus alertas
Alarmes
Vai sair de casa? Vai estar fora durante alguns dias?
Planear
Precisa de ajuda? Não consegue utilizar a voz? Não sabe o que fazer?
Guia
Altere as suas definições e preferencias.
Opções

Medication Assistant
Help to improve your health
Esqueci-me da medicação
O que se esqueceu?
Esqueci-me da medicação
Estou com efeitos secundários?
Tenho esquecido a medicação?
Clique para falar

Medication Assistant
Help to improve your health
Sobre o medicamento
Informação sobre Omeprazol
Caracteristicas do Medicamento
Omeprazol
Cilindrico, vermelho e amarelo
Expira a: 20-11-2013
Dosagem diária
3 Comprimidos
Porquê tomar o medicamento?
Efeitos secundários do medicamento
Clique para falar

Fig. 5. Graphical user interface of Medication Assistant depicting the main screen, advice on forgetting the intake of medicine and detailed information of medication

Moreover, the application provides additional information about the medications through multiple views making use of different representations (e.g. picture of the pills and the respective package, side effects, name, number of pills per day). The application implements two main use cases: "alert reading" and "missed medication intake". When the alert appears, the list of medications to take is displayed.

The user can interact with the application through speech or touch to obtain detailed information on each medication and in case he forgets to take the medication to inquire if he should take or not the medication. Speech can be used as a shortcut to go to specific views of the application, instead of having to select multiple options to select that view. In order to the system to give an efficient response it is necessary to provide relevant information to the system.

5 Conclusions

In this work, we propose a method to rapidly create new speech enabled application, by integrating the W3C multimodal framework and a generic speech modality in new application. To test our method we have developed two different application targeting different devices integrating the multimodal framework, serving as evidence of the increased ease of creating new and diversified application. Then, in a second stage, in which we are currently working on, this application allows us to define new requirements to enhance the generic modality.

Our method allows developers to easily implement an application with speech capabilities in multiple languages. Since the different modalities are decoupled from the application it is possible for the developers to focus only on the application features and design, and less concerns on the design of the interaction are required. Also, modalities can be extended to improve functionalities, to support other features, without the need to update the application. At time of writing the framework and modality is being explored for the development of Paelife Personal Assistan [22] and integrated multilingual support is being extended.

The decoupled nature of the interaction modalities and the existence of a standard multimodal framework pave the way to first attempts to consider multimodal design guidelines independently from the application, with the management of such aspects done at the multimodal framework level, e.g., regarding when to use speech, how to adapt the speech output considering the current context or how to use speech in parallel with other modalities.

Aknowledgments. The work presented is part of the COMPETE – Programa Operacional Factores de Competitividade and the European Union (FEDER) under projects AAL4ALL (www.aal4all.org): Part of the work presented was funded by FEDER, COMPETE and FCT in the context of AAL/0015/ 2009, IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEstC/ EEI/UI0127/2011) and project Cloud Thinking (QREN Mais Centro program, ref. CENTRO-07-ST24-FEDER-002031).

References

1. Ford sync, `http://www.ford.com/technology/sync/`
2. ios - siri, `http://www.apple.com/ios/siri/`
3. Xbox one, `http://www.xbox.com/en-GB/xbox-one/meet-xbox-one`
4. Baggia, P., Burnett, D.C., Carter, J., Dahl, D.A., McCobb, G., Raggett, D.: Emma: Extensible multimodal annotation markup language, `http://www.w3.org/TR/emma/`
5. Barnett, J., Akolkar, R., Auburn, R., Bodell, M., Burnett, D.C., Carter, J., McGlashan, S., Lager, T., Helbing, M., Hosn, R., Raman, T., Reifenrath, K., Rosenthal, N., Roxendal, J.: State Chart XML (SCXML): State Machine Notation for Control Abstraction, `http://www.w3.org/TR/scxml/`
6. Bernsen: Towards a tool for predicting speech functionality. Speech 23, 181–210 (1997)
7. Bernsen, N., Dybkjaer, L.: Multimodal Usability (2009)

8. Bernsen, N.O.: Multimodal usability: More on modalities (December 2012), http://www.multimodalusability.dk/
9. Bernsen, N.O.: Multimodality in language and speech systems – from theory to design support tool. In: Granstrm, B., House, D., Karlsson, I. (eds.) Multimodality in Language and Speech Systems, Text, Speech and Language Technology, vol. 19, pp. 93–148. Springer, Netherlands (2002)
10. Bodell, M., Dahl, D., Kliche, I., Larson, J., Porter, B.: Multimodal Architecture and Interfaces, W3C (2012), http://www.w3.org/TR/mmi-arch/
11. Bui, T.H.: Multimodal dialogue management - state of the art. Technical Report TR-CTIT-06-01, Centre for Telematics and Information Technology University of Twente, Enschede (January 2006)
12. Dahl, D.A.: The W3C multimodal architecture and interfaces standard. Journal on Multimodal User Interfaces (April 2013), http://link.springer.com/10.1007/s12193-013-0120-5
13. Deketelaere, S., Cavalcante, R., RasaminJanahary, J.F.: Oasis speech-based interaction module. Tech. rep. (2009)
14. Hak, R., Dolezal, J., Zeman, T.: Manitou: A multimodal interaction platform. In: 2012 5th Joint IFIP Wireless and Mobile Networking Conference (WMNC), pp. 60–63 (September 2012)
15. Hale, K.S., Reeves, L., Stanney, K.M.: Design of systems for improved human interaction (2011)
16. Hoste, L., Dumas, B., Signer, B.: Mudra: A unified multimodal interaction framework. In: Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, pp. 97–104. ACM, New York (2011)
17. Johnston, M., Fabbrizio, G.D., Urbanek, S.: mtalk - A multimodal browser for mobile services. In: INTERSPEECH, pp. 3261–3264. ISCA (2011)
18. Nass, C., Brave, S.: Wired for Speech: How Voice Activates and Advances the Human-computer Relationship. MIT Press (2007)
19. Sarter, N.: Multimodal information presentation in support of human-automation communication and coordination, vol. 2, pp. 13–35. Emerald Group Publishing Limited (2002)
20. Sarter, N.B.: Multimodal information presentation: Design guidance and research challenges. International Journal of Industrial Ergonomics 36(5), 439–445 (2006)
21. Teixeira, A., Braga, D., Coelho, L., Fonseca, J., Alvarelhão, J., Martins, I., Queirós, A., Rocha, N., Calado, A., Dias, M.: Speech as the basic interface for assistive technology. In: Proc. 2th International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion, DSAI (2009)
22. Teixeira, A., Hämäläinen, A., Avelar, J., Almeida, N., Németh, G., Fegyó, T., Zainkó, C., Csapó, T., Tóth, B., Oliveira, A., Dias, M.S.: Speech-centric multimodal interaction for easy-to-access online services – A personal life assistant for the elderly. In: Proc. DSAI 2013, Procedia Computer Science (November 2013)
23. Teixeira, A.J.S., Almeida, N., Pereira, C., Silva, M.O.: W3c mmi architecture as a basis for enhanced interaction for ambient assisted living. In: Get Smart: Smart Homes, Cars, Devices and the Web, W3C Workshop on Rich Multimodal Application Development. New York Metropolitan Area, US (July 2013)
24. Teixeira, A.J.S., Ferreira, F., Almeida, N., Rosa, A.F., Casimiro, J., Silva, S., Queirós, A., Oliveira, A.: Multimodality and adaptation for an enhanced mobile medication assistant for the elderly. In: Third Mobile Accessibility Workshop (MOBACC), CHI 2013 Extended Abstracts (April 2013)

Introducing Consciousnet : Internet Content as an Environment for Human-Machine Interaction

Vincenzo Catania[1], Davide Patti[1], and Mariagrazia Sciacca[2]

[1] DIEEI, University of Catania, Italy
{vincenzo.catania,davide.patti}@dieei.unict.it
[2] Dipartimento Scienze della Formazione, University of Catania, Italy

Abstract. In this work we introduce Consciousnet , an open source architecture aimed to provide a general purpose environment for experimenting with human-machine language interaction. The main idea is exploiting the distributed and unsupervised complexity of the Internet in order to get all the semantic/syntactic material needed to carry on a linguistic text based interaction. After describing the main elements of the architecture, the results of a set of Turing-inspired tests are shown to demonstrate how the unpredictability and generality of the environment can be used as a basis for designing tests and experiments involving both psychologists and AI scientists.

1 Introduction and Motivation

Language-based interaction between human and machines has always attracted several actors belonging to very heterogeneous fields, from computer architecture designers to cognitive science researchers, psychologists, language formalists, philosophers and sometimes also artists [6] [11] [12] [2]. While the final purpose behind each of these fields may be different, what all the approaches have in common is the intrinsic difficulty of dealing with a language-based interaction. Language is still probably the most hardly-reproducible behaviour of human entities, strictly related to the inner complexity of the way human intelligence represents and interacts with the environment. If several physical features of the human body have been mechanically replicated in the recent years [20] [4], the same success did not come for language-based interaction, where the hunt for a more human-like behaviour is still wide open.

In this work we present Consciousnet , an artificial intelligence environment for experimenting with human-computer linguistic interaction. The name itself comes from the contamination of the words *consciousness* and *net*, denoting the main idea which characterizes the environment: exploiting the chaotic, unsupervised knowledge of the Internet as a collective “consciousness” that can be stimulated by the user while interacting with an artificial entity.

Three fundamental requirements are behind the design philosophy adopted within Consciousnet :

Pure Text-Based Interaction: no need for any added “realism” based on multimodal techniques. This excludes speech synthesis/recognition, three-dimensional

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 382–393, 2014.

avatars, touch-based interactions or robotics. The idea was focusing on linguistic interaction instead of introducing distracting elements revealing/recalling the artificial nature of the interaction. For example, although speech synthesis is widely used to "humanize" the user experience, it makes the interaction more recognizable as non-human, while text-based output are more maskable.

Generality: not specialized or focused on a particular conceptual domain. So, while limited to a pure text-based interaction, the space of action of that interaction should have no limitation a priori. This, for example, differentiates Consciousnet from the field of expert systems and assistance or entertainment chat robots.

Unstructured Complexity: only simple and easily adaptable components should be used, with the aim to generate complexity from their interaction rather than forcing them to behave in a complex way. Thus we also avoided the usage of semantic/formal systems aimed to capture the "meaning" of the user input. In other words, the complexity is not encoded *inside* the functional model of Consciousnet , but is obtained by stimulating the intrinsic complexity of the Internet content (see Figure 1).

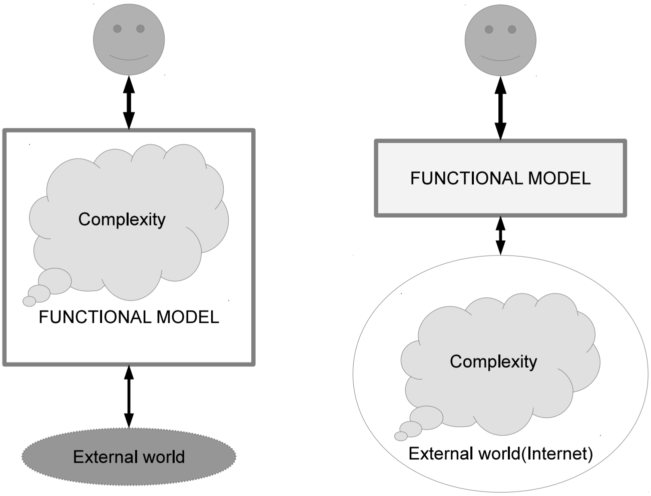


Fig. 1. Structured vs unstructured approach adopted in Consciousnet

2 Background and Contribution

From a computer science/software perspective, Consciousnet could be classified as a *chatbot*, that is a program which simulates an artificial intelligence capable to textually interact with users. This field, originated by the seminal work [19], was followed by many implementations over the years [3] [16] [17], basically differentiating each other by two main aspects: (i) the complexity of the parsing model applied to the user input and (ii) the knowledge base used to generate responses. In the last year some works also proposed to feed this knowledge base using Internet resources, e.g. discussion forums [9] [1].

The contribution we aim to introduce with `Consciousnet` is the attempt to remove any explicit knowledge base and use the Internet as an autonomous structure which provides the semantic and syntactic material that can be forged to create the interaction. In particular, two aspects we want to emphasize here:

Unsupervised linguistic space: no database of concepts or archive of responses to be maintained; the current status of the Internet itself determines the size and the content of the space into which `Consciousnet` moves. It should be pointed out that this affects both semantics and syntactical aspects of this space (e.g. slangs, abbreviations, common errors are part of this linguistic space)

Unpredictability: the same nature of Internet content, fluid, mutable, intrinsically chaotic and hardly controllable leads to an interesting degree of indeterminism in the behaviour of `Consciousnet` . Even knowing the user inputs in advance, it would be hard to predict the interaction development.

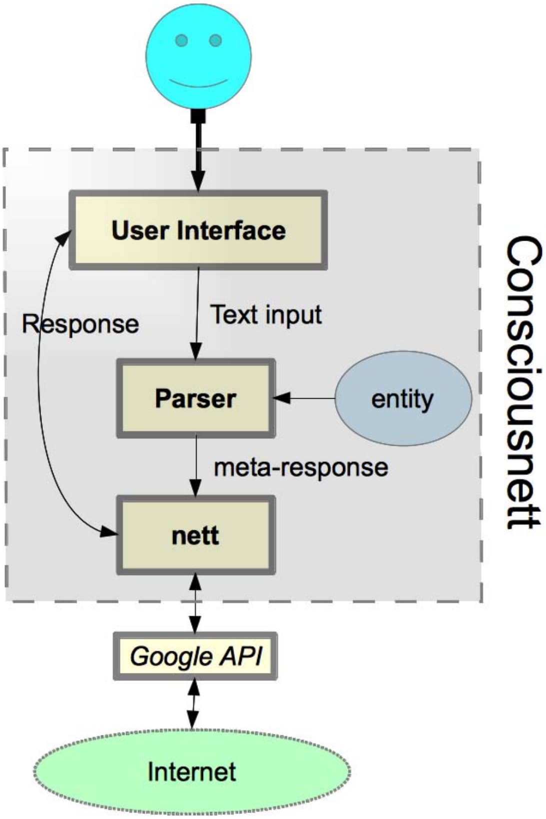


Fig. 2. `Consciousnet` Architecture components and data flow

3 `Consciousnet` Architecture

In this section we introduce the architecture of the `Consciousnet` environment and a detailed description of its elements. The source code of the whole environment is freely available at [15], together with the appropriate instruction for setting up the environment and the complete set of experiments carried out in the next Section.

3.1 Architecture Sketch

As shown in Figure 2, the text input is introduced by the user using an User Interface and then analysed by the *Parser* in order to produce a *meta-response* . The *meta-response* is not the final output of the entity, but some kind of "reaction" that will be used by the subsequent *Net* component to stimulate the Internet. The actual interface between *Net* and the Internet consists of a set Google APIs [10], freely available for non-commercial purposes. Once the *Net* component has processed the output of these API, the final response is returned to the user so that the interaction loop can repeat. As shown, a data structure called *Attitude* is used to generate the *meta-response* : *Attitude* represents the controllable part of the entity behaviour, in some way determining its "personality". In the following subsection we describe more in detail each component involved in the interaction loop.

3.2 User Interface

While the interaction is simply based on text input/output, a few tricks have been adopted in order to make the environment suitable for a more realistic experience. First, the entity response is not returned immediately after user input is entered, but following a time delay $T_{response\_delay} = T_{read} + T_{think}$, where both values are proportional to the number of chars of the text strings involved, with the aim of simulating the time required for reading the user input and then thinking a response. Further, text does not appears on the interface screen all at once, but as a flow of randomly intervalled chunks of chars, like happening in a live typing session. This last trick was not really necessary as the previous one, but adding some more human-like typing to the remote entity demonstrated to mitigate the artificiality intrinsic in the proportional delay $T_{response\_delay}$.

Note that all of these tricks could have been simply avoided using an hidden human counterpart in order to type the output generated by the artificial entity. However, having a self-contained environment which includes a modelization of controllable human-like typing/reading yields a further degree of flexibility of the environment as experimental platform, e.g. different delay values could be investigated to evaluate the effect of slower/faster interactions on language.

3.3 Parser

As next step, input is analysed by the *Parser* using the *Attitude* data structure in order to generate a *meta-response* . The *Attitude* consists in a hierarchy of *entry-point* , *decomposition pattern* and *meta-response* , as follows:

```
entry-point: X
   pattern: X1
       meta-response X1.1
       meta-response X1.2
       meta-response X1.3
   pattern: X2
       meta-response X2.1
       meta-response X2.2
       meta-response X2.3
```

The *entry-point* represents a sort of keyword that opens the understanding of the input. The idea is to think to the artificial entity like a person trying to understand some sentences in a foreign language. The first thing should be to capture as more words as possible that provide a meaningful interpretation key for the whole sentence. Of course, more than one these entry points could be found in each input, so a sort of priority mechanism has been chosen. Continuing the analogy with the foreign language, we can observe that the more abstract and general a word is, less is the semantic value useful as understanding entry point for the whole sentence. For example, denoting with "*" the not understood parts of a sentence, catching a pattern like "** me * **" does not help, since abstract words like "me", "you", "is","are" are very common and do not carry any particular semantic weight to characterize the meaning of the sentence. A less abstract entry-point, e.g. "food", could be more useful in that sense, since one could at least argue that the counterpart is talking about something to eat. Very low abstract entry-points, e.g. the name of a city or a car model, could give even more hints when trying to build a conversation in a foreign language, since you can have a more accurate understanding of which conceptual domains could be touched from the current stage of the dialogue. Table 1 show the class of entry points considered, ordered by abstraction level.

Once the *Parser* has used the *Attitude* data structure to find the *entry-point* with the highest priority (lowest abstraction), a set of decomposition pattern is

Table 1. entry-point list and abstraction levels

Level	Type of Element	Examples
0	Commonly found, generic syntactical elements, i.e. not useful for any restriction of the conceptual domains	I, me, are, you, sorry, yes, no
1	Verbs, nouns, elements denoting something less generic, such hypothesis, questions etc...	if, because, why, how, when, always
2	Terms introducing items that assume importance to the speaker	my X, your X
10	Terms introducing specific domains	Music, sport, love, school, food, money
10+	Very low level abstraction terms, referring to a very specific subject	Mozart, Golf, Berlin, spaghetti

considered in order to determine the constituting elements of the sentence and build an appropriate *meta-response* . As a practical example, let's consider the following snippet of *Attitude* . For sake of simplicity, it's a very basilar pattern with only a few entries:

```
entry-point: love
    pattern: * I love *
        meta-response: fans of (2)
        meta-response: (1) because loving (2)
    pattern: * when *
        meta-response: goto when
    pattern: * love *
        meta-response: (1) hate (2)
entry-point: when
    pattern: * when *
        meta-response: how often (2)
```

This example shows two *entry-point* : the first is associated to the word "love" and has three patterns. For each pattern, a set of *meta-response* is available. The input is compared against matching patterns, extracting some placeholders (e.g. (1) and (2) in the example), and then translated into a *meta-response* . The concept of *meta-response* is probably the most important in the architecture. As said, it is not the final output of the artificial entity, but an input that will be used by the *Net* component to generate the actual response. In this example, an user input like "In the morning I love cats" would match only the first pattern, mapping the placeholders (1) and (2) to "In the morning" and "cats" respectively. A randomly chosen item in the corresponding set of meta-responses (e.g. "fans of cats") will be then used by the subsequent *Net* component to access the Internet. Note how is possible to use one of the pattern to delegate the *meta-response* to a different *entry-point* ("when") if the corresponding keyword is present. The idea is to use an *entry-point* with lower priority if no better choice is available. The set of *entry-point* is read in order, so in the example above the input "when I love cats" will match the first pattern, while "you love me when I play golf" would match the second, linking then to *entry-point* "when" and generating the *meta-response* "how often I play golf".

3.4 Net

The *Net* component implements an interface to the Google CustomSearch APIs [10], used in conjunction with the *meta-response* to extract data from the Internet and generating the actual response of the artificial entity. This involves the following phases:

1. The *Net* component uses its own CustomSearch Engine object using the *meta-response* as main argument.
2. As result, an array data structure containing fragments of data extracted from the internet is returned.

3. At this point, *Net* extracts the *snippet* field of the returned structure. Indeed, our aim is not to deal with low level web code (e.g. HTML or javascript), but with already human-readable text.
4. The resulting blob of data is then processed using an regular expression system implemented in *Net* , performing some post processing tasks: discarding too long/short sentences, strange punctuation, text with not useful content (e.g. all numbers).
5. Finally, each of the filtered items is given a sort of "quality value", depending on some properties of the text, e.g. containing a first-person statement, having a question mark a last character and so on.

3.5 `Consciousnet` Tuning

Each of the components described is designed and implemented in order to work as a separate functional element. The degrees of freedom in the environment configuration are mainly encoded in the *Attitude* data structure. A lot of flexibility comes from tuning the "personality" of the entity in order to accomplish some specific tasks. We can distinguish between two different ways of tuning:

- *Extensively*: this consist in adding new *entry-point* items to the *Attitude* structure, expanding the conceptual domains that can be used as starting point for creating the *meta-response*
- *Intensively*: for a given *entry-point* , the number of decomposition patterns and *meta-response* could be customized in order to enrich the complexity of the behaviour mapped in each *entry-point* .

Users oriented toward an "expert-system" perspective would probably tune `Consciousnet` extensively, adding several *entry-point* elements focused on a particular domain, for example name of symptoms and diseases in an health-care application. Further, the complexity of *meta-response* could also be limited in order to get simpler and controllable text snippets; we cannot exclude interesting results in these ways of tuning `Consciousnet` , but they are apart from the original spirit of the environment and will not further discussed in this introductory work.

On the other side, different fields such as entertainment, language research, psychologists, could benefit from the flexibility of the environment and use the unpredictability and generality as an added value. In the interaction experiments between students and Prof. Gioio entity, presented later in this work, we used an *Attitude* including very heterogenous patterns, e.g. where `I love (1)` could result in the *meta-response* `Reasons why I hate (2)`, generating the not-so-friendly exchanges like:

```
You> I love studying psychology, hope will do good
Entity> I'd like to know from people studying psychology, if
they think
```

The choice of accessing external Internet complexity using appropriate *meta-response* instead of large databases avoids problems of scalability and maintainance that would arise when dealing with a not-specialized, general purpose system. For example, considering the snippet:

```
entry-point: play
   pattern:   I * play *
      meta-response:  people that (1) play (2)
      meta-response:  is the best for (2)
```

One of the experiments performed in the next Section, showed the exchange:

```
You> I used to play bass guitar
Entity> For distortion, definitely the Blackstar HT-Dual
```

which is really satisfying as interaction, even if not technical data about guitar tools has ever been introduced in Consciousnet . It has been sufficient to provide a good pattern to catch the placeholder "bass guitar", without really encoding anything strictly related to this particular instrument. In other words, the same patterns could give interesting results even when playing "drums" or "lego building blocks". As seen in Figure 1, a fundamental design choice is to leave complexity outside the environment enabling intelligent ways of letting it emerge from the network. As a further interesting consequence, if something that does not exists today will appear tomorrow, the collective consciousness of the network will be absorbing it, and then it will become something that Consciousnet could successfully use in its dialogues.

4 Example: Reversed Turing Test

In order to demonstrate the effectiveness of Consciousnet flexibility, this section shows how the environment has been used to carry out a set of Turing-test inspired experiments, specifically designed for this work. The idea was investigating some properties of the language used in two different sets of users: a set α, being aware of the artificial nature of the entity and a second set β, not being aware of. While the original Turing's test consists of an human entity which analyses language to guess whether the counterpart is artificial or not, the proposed experiment completely reverses the perspective: we explicitly make statement about the other entity nature, analysing how language is affected from this awareness (see Figure 3).

4.1 Experimental Setup

The two sets α and β consisted of 30 students, taken from a course on Fundamental of Informatics for Psychology, held at University of Catania in 2013 [14]. Students belonging to the set α were instructed as they were performing a text only connection with an american Professor, Doctor Paul Gioio [1], interested into

[1] Named after the italian nickname of one of the authors' son.

testing a new form of interaction to be used in future on his own students. Each student entered (one per time) in a isolated room, where a 10 minutes chatting session was performed. In order to maximize the chance of masquerading the artificial nature of the entity, students who already performed the session were moved into separated room. This avoided the change of influencing successive students with doubts or considerations on what happened during their interaction. Of course, interaction sessions belonging to set β did not require such expedients and they were explicitly told to perform a chat with an artificial intelligence entity.

4.2 Results

When all the students of set α ended their 10 minutes session, the experiment was revealed. Quantifying "how much" Prof. Gioio was considered as real is not in the purposes of the experiments and it would be really difficult to gather such a measure: the majority of the students declared themselves as "surprised" and a few of them told of having developed some suspects. In every case, we are not interested in what is their opinion *after* the interaction, but how they acted *during* the interaction. Thus we can safely assume that even more suspicious students have been interacting supposing an human counterpart, since the short time available and the particularity of the situation forced them to adopt a conservative behaviour.

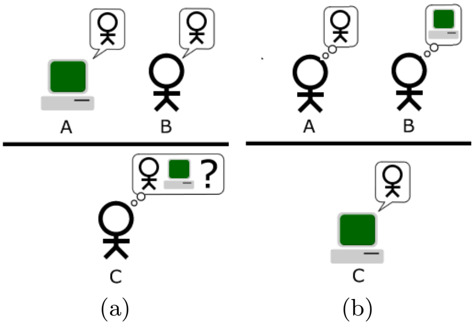


Fig. 3. Original Turing Test (a) and the reversed version (b) performed with `Consciousnet`

A total of 288 and 261 sentences have been collected from participants of sets α and β respectively, i.e. excluding those produced by the `Consciousnet` platform when interacting. A quantitative and qualitative analysis of the transcripts was conducted in order to investigate an impact in terms of semantic domains covered, syntax-oriented metrics, and statistically significative differences among the two sets α and β. Two kind of words have been removed from the collected sentence in order two produce less noisy results: the first set of words are those below a frequency threshold of 1% on the overall data. The second kind of words

Table 2. Most used words from Stanford POS analysis:α (top) and β (bottom)

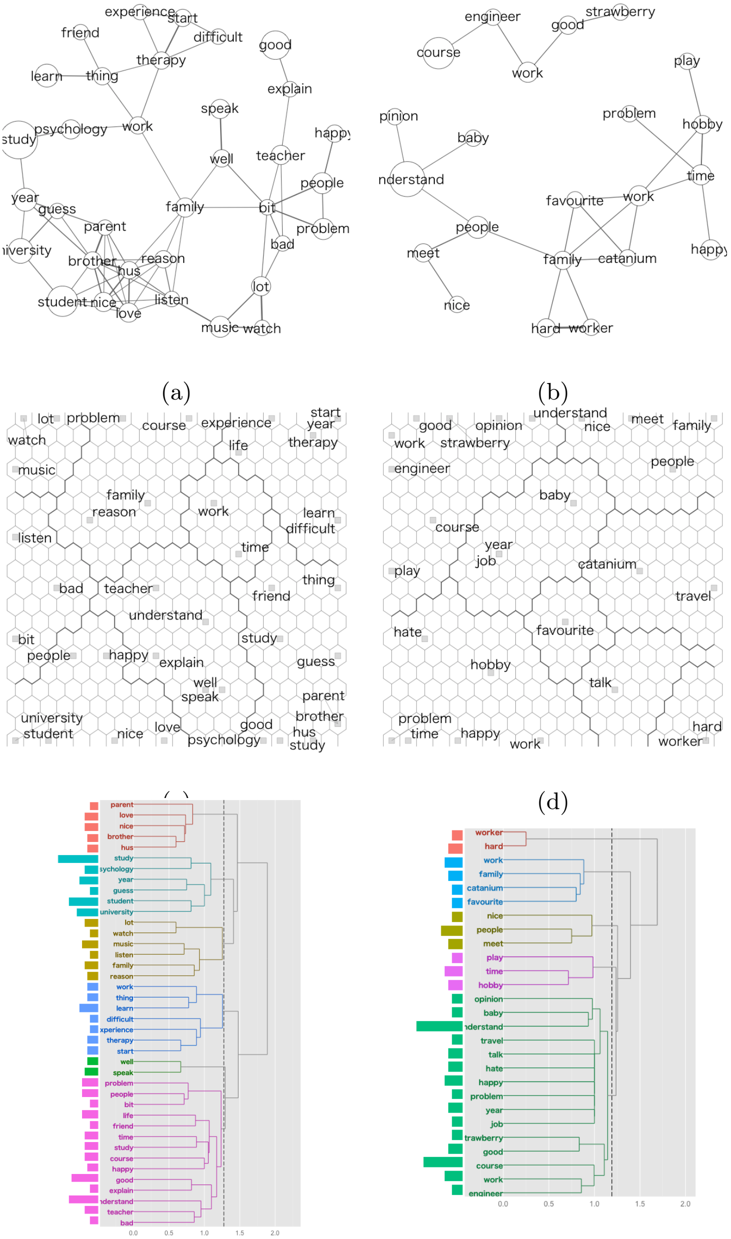
Noun		**Adj**		**Adv**		**Verb**	
study	15	good	10	well	4	understand	11
student	11	nice	5	exactly	2	learn	7
university	8	happy	4	probabily	2	love	5
year	7	bad	3	absolutely	1	speak	5
life	6	difficult	3	extremely	1	study	5
music	6	easier	2	frequently	1	start	4
people	6	favourite	2	good	1	work	4
Noun		**Adj**		**Adv**		**Verb**	
course	11	happy	5	hard	3	understand	13
people	6	good	4	well	2	work	5
time	5	hard	4	close	1	hate	4
work	5	favourite	3	dear	1	meet	4
family	4	nice	3	realy	1	talk	4
hobby	4	afraid	2	simply	1	play	3

excluded are those commonly referred as “stop words”, i.e. words that do not play any particular semantic role (for a complete list of the stop words adopted, see also [13]).

Table 2 shows frequency list of most used words, labeled with the Stanford POS Tagger [7] considering 4 categories (noun, adjectives, adverbs, verbs). Further, a more complex analysis was conducted using the *R* [5] statistical tool in conjuction with *KH coder* [8], an open source software for content analysis, text mining or corpus linguistics. Part of the results obtained are summarized in the Figure 4, including co-occurrence network, self-organizing maps and clustering using method Ward with a Jaccard distance [18]. While these preliminary results seems to show some interesting differences between the two sets, any detailed and meaningful interpretation of such data is beyond the scope of this introductory word, which has been explicitly focused on the architecture of the Consciousnet environment.

5 Conclusions

In this paper we introduced Consciousnet , an artificial intelligence environment exploiting the Internet to perform a general purpose, not-specialized text based interaction. A set of experiments have been carried out to demonstrate how the environment has been used to perform a reversed version of the Turing test. Future works will involve both the improvement of the Consciousnet network-based intelligence and the design of new experimental tests for the research in the human-machine interaction field.



(a) (b)

(c) (d)

(e) (f)

Fig. 4. (a,b) Co-occurrence Network, (c,d) self-organizing maps and (e,f) word clustering for sets α and β

References

1. Cao, Y., Yang, W.-Y., Lin, C.-Y., Yu, Y.: A structural support vector method for extracting contexts and answers of questions from online forums. Inf. Process. Manage. 47(6), 886–898 (2011)
2. Clarke, A.C.: 2001: A space odissey. Pearson Education (2001)
3. Colby, K.M.: Simulation of belief systems. In: Computer Models of Thought and Language, pp. 251–286 (1973)
4. Eaton, M.: An approach to the synthesis of humanoid robot dance using non-interactive evolutionary techniques. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3305–3309. IEEE (2013)
5. The R Foundation for Statistical Computing. The r project for statistical computing, http://www.r-project.org
6. Goldstein, I., Papert, S.: Artificial intelligence, language, and the study of knowledge. Cognitive Science 1(1), 84–123 (1977)
7. The Stanford Natural Language Processing Group. Stanford log-linear part-of-speech tagger, http://nlp.stanford.edu/downloads/tagger.shtml
8. Higuchi, K.: Kh coder, http://khc.sourceforge.net/en/
9. Huang, J., Zhou, M., Yang, D.: Extracting chatbot knowledge from online discussion forums. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI 2007, pp. 423–428. Morgan Kaufmann Publishers Inc., San Francisco (2007)
10. Google Inc. Google customsearch, https://developers.google.com/custom-search/
11. Isles, D.: Artificial intelligence as a possible tool for discovering laws of logic. Cognitive Science 2(4), 329–360 (1978)
12. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of artificial intelligence. Stanford University (1968)
13. Ranks nl webmaster tools. English stopwords, http://www.ranks.nl/resources/stopwords.html
14. University of Catania. Dipartimento di scienze della formazione, http://www.disfor.unict.it
15. Patti, D.: Network knowledge based ai entity, https://code.google.com/p/consciousnet/
16. Shawar, B.A., Atwell, E.S.: Using corpora in machine-learning chatbot systems. International Journal of Corpus Linguistics 10(4), 489–516 (2005)
17. Tarau, P., Figa, E.: Knowledge-based conversational agents and virtual storytelling. In: Proceedings of the 2004 ACM Symposium on Applied Computing, SAC 2004, pp. 39–44. ACM, New York (2004)
18. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244 (1963)
19. Weizenbaum, J.: Eliza a computer program for the study of natural language communication between man and machine. Commun. ACM 9(1), 36–45 (1966)
20. Yamane, K., Nakamura, Y.: Robot kinematics and dynamics for modeling the human body. In: Kaneko, M., Nakamura, Y. (eds.) Robotics Research. STAR, vol. 66, pp. 49–60. Springer, Heidelberg (2010)

Can User-Paced, Menu-free Spoken Language Interfaces Improve Dual Task Handling While Driving?

Alexander Eriksson[1, 3], Anders Lindström[2], Albert Seward[2]
Alexander Seward[2], and Katja Kircher[3]

[1] Linköping University, Department of Computer and Information Science, Linköping, Sweden
aleer821@me.com
[2] Veridict AB, Stockholm, Sweden
{anders.lindstrom,albert.seward,alexander.seward}@veridict.com
[3] Swedish National Road and Transport Research Institute, Linköping, Sweden
katja.kircher@vti.se

Abstract. The use of speech-based interaction over traditional means of interaction in secondary tasks may increase safety in demanding environments with high requirements on operator attention. Speech interfaces have suffered from issues similar to those of visual displays, as they often rely on a complex menu structure that corresponds to that of visual systems. Recent advances in speech technology allow the use of natural language, eliminating the need for menu structures and offering a tighter coupling between the intention to act and the completion of the action. Modern speech technology may not only make already existing types of interaction safer, but also opens up for new applications, which may enhance safety. One such application is a speech-based hazard reporting system. A small fixed-base simulator study showed that drivers adapt the timing of the hazard reports to the situation at hand, such that an increase in reported workload was avoided.

Keywords: speech-based interface, natural language, compensatory behaviour, hazard reporting, human factors, VUI, strategic driving behaviour, simulated driving, IVIS.

1 Introduction

The use of speech-based interaction over traditional means of interaction in secondary tasks may increase safety in demanding situations, like when driving an automobile or flying an aircraft, where the requirements on operator attention and vigilance are high. Traditional means of interaction such as screens, buttons and touchscreens usually involve rather complex menu structures, input fields and controls. Using those means while performing a high workload spatial task such as driving decreases overall performance significantly as this type of interaction competes for our limited resources [1-3].

Speech interfaces have until recently suffered from issues similar to those of visual displays, as they often rely on a complex menu structure corresponding to that of

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 394–405, 2014.

visual systems, with the added drawback that it is normally inappropriate or even impossible to make said structure immediately visible on a screen. Instead, it has to be envisioned and/or remembered by the user. Typical examples are modern in-car navigation systems and speech recognition-based telephone services. These systems typically employ a rigid menu structure with clear expectations on the next input from the user, such that the input method is very similar to a traditional visual/manual input, except that verbal commands replace the hand movements.

Recent advances in speech technology allow the use of natural language and thereby eliminate the need for menu structures. The use of a natural language based speech interface offers a tighter coupling between the intention to act and the completion of the action as the user can jump between topics and deal with several conversation threads in parallel, as well as spontaneously introduce novel topics, without having to allocate any resources to accommodate the system structure. The conversation is essentially user-paced, which means that it can be interrupted, suspended and resumed just as the user sees fit. Modern speech technology has many potential application areas in driving, both within the domain of comfort systems and of safety systems. It may not only make already existing types of interaction safer, but also opens up for new applications, which may enhance safety.

In this article, although mainly theoretical, we will therefore also test a concept for a traffic hazard reporting system that is based on novel types of spoken language interfaces, which are characterised by being user-paced and menu-free, as opposed to more traditional types of speech interfaces. The main point of this paper is the assumption is that this new type of interface can be handled by drivers with only negligible interference with the driving task. Thus, its net effect is expected to be an increase in traffic safety, as drivers will be warned of hazards, while the action of reporting the hazard does not have any measurable negative impact.

1.1 Spoken Language Interfaces in Driving

There is ample evidence from different fields of research that adding a secondary task to a primary task usually deteriorates performance in the primary task [4-7]. Controlled studies of driving behaviour have shown time and again that adding a task like using a mobile phone or navigation system to the driving task often leads to increased reaction times [e.g. 8, 9]. Given that driving is generally agreed to have a large visual component [10], this concern is especially pronounced for tasks classified as having heavy visual/manual components, that is, where the driver needs to look away from the traffic and manipulate dials or screens. At the same time, contemporary naturalistic studies show that the act of talking on the phone was not associated with an increased crash risk when separated from dialling and reaching for the phone [11, 12].

These data therefore indicate that behavioural effects observed in simulator trials, in controlled test track tests or even in on-road studies cannot be translated one to one to behaviour in real traffic. Drivers in controlled studies are typically assigned pre-specified tasks to be carried out under certain conditions and are required to execute the secondary task in a given situation. In real traffic, however, the same drivers would have much greater room to compensate for their expected temporary attentional deficits. It is probably partly therefore that crash numbers have not increased substantially as the entertainment technology increasingly has found its way into our vehicles. Drivers do not use the technology completely uncritically, but they adapt their behaviour on the operational, the tactical and even on the strategic level, for example by slowing down, by choosing not to overtake, and by selecting situations of low complexity for secondary task execution [11, 13]. They also adapt how they execute the secondary task, for example by keeping telephone calls brief [14]. Some studies exist, although admittedly impressionistic in nature, indicating that drivers employ a whole range of linguistic devices and other communicative strategies in their spoken interaction within the vehicle or over a mobile phone line, in order to accommodate the cognitive demands of simultaneous talking and driving [15]. In a more recent study it was also investigated how cognitive load affects the degree of disfluencies during in-vehicle spoken dialogue between drivers and passengers performing a consciously demanding interview task while navigating in real traffic. The passengers acted as interviewers while also giving navigation instructions. On a side note, the authors made the collateral finding that all passengers actually spoke less disfluently when their drivers experienced high workload, which could also be seen as a tell-tale sign of co-operative adaptation from the side of the passengers [16].

While it is good news that drivers adopt strategies to improve their safety, it is of course advisable to offer methods of interaction that interfere as little as possible with the driving task. It may very well be that a certain type of secondary task can be impossible to perform with one type of interface, but very easily with another.

In order to make predictions about how different interface solutions affect the driving task, we lean on the concepts provided by the theory of threaded cognition [3]. In this theory it is assumed that different tasks within a multitasking environment are made up of different threads. Threads compete for the same cognitive, perceptual and motor resources. Different resources can operate in parallel, but each resource can only handle one request at a time. This explains why certain multitasking processes lead to degraded performance in either one or several tasks, how this can change over time with additional practice, and, furthermore, why multitasking in some cases does not have to lead to degraded performance in any of the subtasks.

According to the theory of threaded cognition, as well as other multiple resource theories [e.g. 2, 17, 18, 19], task interference resulting in degraded performance is more likely to occur and persist when different task threads contend for the same resources. Procedural resources are central and used frequently by any thread, but by avoiding the use of the same peripheral resources, parallel processing of different threads can be achieved. A further improvement is achieved by a reduction of the use of declarative resources, which occurs for example when instructions for actions have to be retrieved from memory. As shown by Salvucci and Taatgen [3], a visual dialling task produced a higher level of interference with driving than did a comparable voice-dialling task. Similarly, a study by Levy et al. [20] showed that resource interference degrades performance. Based on this notion the NHTSA published guidelines on how to test built-in visual/manual interfaces for their suitability while driving [21].

This would lead to the seemingly easy conclusion that the vocal/auditory channel should be the interface of choice for tasks that are to be processed in parallel with a visual task like driving. However, empirical data show that this is not necessarily the case. Yager [22] had her participants type and send text messages manually, and verbally with two commercial speech-based mobile services (Siri and Vlingo) while driving. She did not find any differences between the input methods with respect to a number of performance indicators, like eye gaze to the forward roadway, standard deviation of lateral position, mean speed and speed variation. A closer examination shows, however, that the allegedly verbal input methods still included a rather intensive usage of visual resources, and as a result, a comparison with manual text input is not as clear-cut as it would first seem to be. Furthermore, the two speech-based input systems are also likely to have incurred an intensive use of declarative resources, since the participants were not necessarily familiar with the voice protocol that has to be used for the two services. Thus, they often had to retrieve information memorised during instructions and initial practicing.

This points to one of the major problems which Yager's study shares with many similar reports, namely that the attempted over-all comparison between two types of interfaces (here: "direct manipulation vs. speech-based") based on the direct comparison of specific systems fails to generalise from the specific case to the canonical. Furthermore, as rightly pointed out by Green [23], "the demand characteristics of in-vehicle tasks in question are not well quantified", and even if they are split up (as they often are, following an idea popularised by McCracken and Aldrich [24]) into visual, auditory, cognitive and psychomotor demands (VACP), the question of what exact levels of task demands should be considered to be excessive still remains largely unanswered. Also according to Green [23], this is of course further aggravated by the fact that the workload of the main task of driving is not well quantified, which, in turn, is why most investigations resort to some sort of indirect comparison between the influence of different secondary tasks, as exemplified by Yager's study.

It is worth pointing out that many pre-existing commercial and research systems involving spoken input for mobile use, both in and outside of vehicles, in fact do not profit from the main advantages of spoken language interaction, but do little more than "push buttons using voice". For reasons unknown, the hierarchical menu design brought about by WIMP1-style computer GUIs already more than three decades ago, has had a remarkable but undesirable tendency to carry over to Voice User Interface (VUI) design, with many awkward and unintuitive system designs as a result. We would like to argue that making decisions and developing safety guidelines based on studies of such systems is misleading. Instead, we suggest that studies be made involving (real or simulated) spoken language systems where:

1. The (secondary) task lends itself to verbal interaction
2. The system design takes advantage of the intrinsic benefits of using human language and properly exploits the verbal and auditory channels
3. Users are allowed the benefit of training
4. Users are allowed the benefit of (tactical) planning

[1] windows, icons, menus, pointer.

The first issue might seem trivial, but is in fact often overlooked. There are many cases where operating a button via direct manipulation and with instant tactile, visual or auditory feedback is optimal, such as when turning on the headlights or honking the horn. On the other hand, dashboards would soon be completely cluttered if each and every function stemming from the infotainment escalation in recent years had got a button of its own. The second item is perhaps the most generally ignored, and consequently holds the most development potential. We will come back to that later in the choice of experimental task. The third point reminds us that it is otherwise considered acceptable to allow considerable amounts of training, which is normally required for example when learning how to operate a stick-shift transmission, or when adapting to the levers and buttons of a new car. The fourth bullet highlights that the leeway introduced by having a user-paced scenario may be enough to accommodate a range of secondary tasks without negative interference with the primary task of driving.

1.2 A Tentative Service Scenario

Today, many radio channels and TV stations provide live (non-critical) traffic congestion, hazard, and obstacle reporting. Reporting is typically done by telephone. Eventually the information will be relayed to a large number of road users, for example by public broadcasting on the radio, possibly with prioritised reception by virtue of the Traffic Announcement bulletin handling present in RDS-enabled radio receivers since the late 1980s.

As both driving and making the phone call depend on procedural and visual resources, dual task performance is likely to be degraded. There are further technical disadvantages associated with making hazard reports by phone. The driver needs to give the precise location of the hazard, which puts demands on declarative memory. The hazard report has to be processed by a human, causing an unavoidable delay from reporting to broadcasting. Finally, distribution via broadcast radio inevitably precludes the possibility to individualize the message and to convey it only to those affected by the hazard.

Given all these issues it is worth investigating whether a traffic hazard reporting system could be automated. The idea is to use verbal reports from drivers, connect them to the position at which the report was made, have a backend that evaluates all reports, and then distribute appropriate information, warnings and alerts selectively to drivers in the affected area. As a first step, in this study we investigate how a simulated voice controlled reporting system would be operated by a driver in traffic. For such a system to be safe for use in traffic it should neither have a substantial impact on reported workload nor on driving performance. Given the low fidelity of the simulator that was available for the study, we did not consider it meaningful to assess driving performance directly. Instead, we decided to investigate how drivers would time their reports with respect to the hazards and obstacles on the road, in keeping with the notion of tactical self-regulation. Based on this it may be possible to draw conclusions about the likelihood that voice based hazard reporting will affect driving performance. We therefore specifically included a comparison between the drivers making reports at their own discretion and a forced "report-as-soon-as-possible" condition.

1.3 Requirements on a Speech-Based Hazard Reporting System

Based on the arguments given in the introduction, we would like to suggest that the following requirements on system design and system performance need to be fulfilled to avoid an excessive increase in workload and to gain user acceptance:

Satisfactory Speech Recognition. In a study by Kun, et al. [25] drivers were given a speech interaction task with a system capable of a simulated high recognition rate (89% of system dialogue turns) and with a system with a simulated poor recognition rate (44%). Their results show that the system with the low recognition rate caused a significant increase in lane position variance when using the push-to-talk (PTT) button, but this effect did not transfer to the better-performing system. These results may be caused by increased workload in the auditory system imposed by the system's poor recognition in combination with the visuospatial task of reaching for the PTT button in the centre console, resulting in an overall increase in workload.

Responsive, Intuitive and Effortless. Auditive and verbal feedback needs to be immediate, just as in human-human conversation. Reporting should be as easy as putting words on thoughts. Verbal reporting is already the primary means today, although it is done by talking over the phone with a human operator, so the task is obviously feasible. Furthermore, as Green [26] has pointed out, issuing a brief verbal command (like reporting a traffic congestion – in his example a voice-controlled radio is operated) is likely to require minimal thought.

Training. The envisioned mobile speech-based system covers many functions with a similar and consistent VUI, and users of the system are typically guided through interactive tutorials for each function specifically for training purposes. Users can also re-visit these tutorials and should have ample time to practice reporting prior to use.

Tactical Choice of Reporting Time. The system should possess functionality for the interruption of on-going dialogues as workload increases and should be able to resume the dialogue where the driver left off as the workload level decreases [27, 28]. We therefore suggest a speech-based system that is completely user-paced, such that users are free to choose the time of reporting to suit their driving pattern.

2 Method

2.1 Participants and Equipment

A convenience sample of 17 participants between 20 and 27 years of age (Mean=24, SD=2.3) took part in the study. All participants were required to have a valid driving license and a minimum of 2 years of driving experience on Swedish roads. The participants were students at Linköping University. Each participant was given two cinema tickets as compensation for participating in the study.

The VTI fixed-base simulator was used for the purposes of this study. The driver environment is constructed from parts of a Ford Focus and has all the essential controls such as the transmission stick, clutch, wheel and brakes. The simulation software was executed on a distributed computer system and the graphics were rendered with a resolution of 720x1280 pixels at 60 Hz and displayed on three 40" flat-screen TVs with a 1080x1920 resolution providing an approximated 120° field of view. Sound effects were provided through a 5.1 Logitech® surround sound system. A 7" resistive touchscreen was placed on the centre console representing a full screen button activating the speech interface.

2.2 Primary and Secondary Task

The primary task was to drive approximately 20 km on a rural road with a village in the middle. The following nine hazards and obstacles were placed along the road:

- a truck trailer parked on the shoulder of the road
- a moose moving towards the road and then stopping at the road side
- a car parked at a bus stop
- a broken down car in an intersection
- a road construction site on an urban road
- a road construction site in an intersection
- a broken down car in an intersection with a cyclist crossing the road with oncoming traffic
- a truck partially parked on the shoulder of the road and a connecting road
- oncoming traffic and a cyclist in the participant's lane, cycling in the opposite direction

The secondary task used in this experiment was a speech interface for reporting roadside hazards and hindrances. The system was simulated using the Wizard of Oz method [29] where a hidden experimenter provided feedback using a synthesised voice. The reason for this was to ensure that VUI design issues would not adversely affect the drivers' performance. The system was activated by pressing the touchscreen, which resulted in auditory and visual feedback. The subject then provided a verbal hazard report that the system associated with the vehicle's physical position along the simulated road. Upon completing a report the subject received auditory feedback by a synthesised voice triggered by the experimenter.

2.3 Procedure and Design

The participants were asked to drive as they would normally do while paying attention to the traffic regulations. They had an initial training phase in the simulator for about five minutes before proceeding to the experimental conditions. There were four different experimental conditions:

- a *baseline* in which participants drove along the route without any additional tasks
- a *self-paced* experimental run in which the participants were asked to use the speech-based hazard reporting system as they saw fit
- a *video* run in which the drivers watched a film of a run in the simulator and used the speech-based reporting system as instructed in the self-paced run
- a *forced* experimental run in which the drivers were instructed to use the system as soon as they detected any traffic hazard or obstacle (externally paced)

It was decided that the order of the reporting conditions should go from least to most specific and enforcing, such that reporting behaviour in the self-paced conditions would not be influenced by the more externally paced condition. This meant that baseline driving always came first, followed by the self-paced and the video condition, the two latter of which were counterbalanced. The forced condition always came last. The ensuing risk for learning effects was considered, but viewed as less problematic than a carry-over effect of the forced reporting behaviour to the self-paced behaviour.

2.4 Data Collection and Analysis

A log was kept of when the participants started reporting each hazard/obstacle. Audio was recorded during the experimental runs using a microphone mounted on top of the dashboard. All participants were asked to fill out the NASA-RTLX form [30] after each run to obtain self-reported workload measurements.

The road around each obstacle was divided into Area 1 (before the hazard/obstacle), Area 2 (next to the hazard/obstacle) and Area 3 (behind the hazard/obstacle). Area 1 started as soon as the obstacle became visible and ended 20 m in front of the obstacle. Area 2 lasted from 20 m in front of the obstacle to 20 m behind the obstacle. Area 3 started 20 m behind the obstacle and ended 150 m behind it. The number of hazard reports per obstacle, area and condition were counted based on where the report was initiated by pressing the touch screen.

3 Results

3.1 Reporting Strategy

For the self-paced condition a total of 115 obstacles were reported, in the video condition 110 obstacles were reported, and for the forced condition 127 obstacles were reported. The obstacle that was least likely to be reported across conditions was a road construction in town, followed by a road construction in a crossing and a bicyclist cycling on the wrong side of the road.

The number of reports per area and condition is displayed in

. In the self-paced condition 57% of the reports were initiated in Area 1, that is, before the area immediately surrounding the obstacle, was reached. More than half of the 22% of reports initiated in Area 2 in the self-paced condition were associated with the moose standing next to the road and with the bicyclist crossing the intersection. The share of reports (21%) that were initiated in Area 3, after the obstacle, were most frequently connected to the truck trailer and the truck parked on the roadside.

In the video condition 63% of the reports were initiated in Area 2 and 31% were initiated in Area 1. Only 6% were initiated in Area 3.

In the forced condition the vast majority of reports (89%) was initiated in Area 1, 7% of the reports were initiated in Area 2, and 4% were initiated in Area 3.

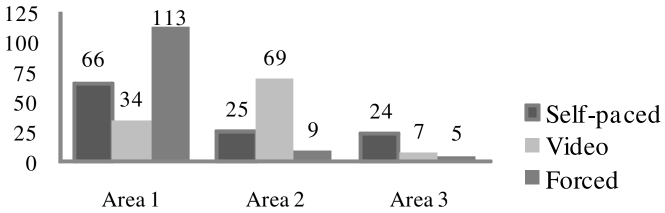


Fig. 1. Report count per road area and reporting condition for 17 drivers performing a simulated speech-based hazard-reporting task

Workload Measurements. Differences in workload ratings between conditions were analyzed using an ANOVA, the results of which are displayed in Table 1. There was a significant difference between conditions (F(3, 42)=5.565, p=.003, ω2=.284, power=.985). Post-hoc tests at the uncorrected α=.05 level showed that the reported workload for baseline was significantly higher than for the self-paced and the forced condition, workload was also higher for the video condition than for the forced condition.

Table 1. Mean workload ratings on the NASA-RTLX and p-values for the post-hoc comparisons in the ANOVA with the factor condition

		baseline	video	self-paced	forced
mean ± sd		164.0±65.0	146.3±64.3	117.7±52.3	105.0±53.4
post-hoc test results	video	p = .091			
	self-paced	p = .021	p = .092		
	forced	p = .011	p = .025	p = .401	

4 Discussion

There is a large contrast between the reporting strategies employed in the different driving conditions, indicating that drivers adapt flexibly to the situation at hand. In the self-paced and in the video condition the drivers were instructed to use the system in the same way, except that in the video condition the driving task was excluded. In the self-paced condition drivers tended to make their reports in Area 1, which could be a strategy to avoid an increase in workload when passing the obstacles. Notable exceptions are the moose on the roadside as well as the bicyclist in the crossing. Here, the workload may be higher in Area 1, as drivers need to brake and assess the somewhat unpredictable behaviour of the moose or bicyclist. In those two cases reporting is often delayed to Area 2, where there is no longer any immediate threat. This claim of the drivers' self-regulation is further supported by the fact that the general strategy changes in the video condition when no workload from driving is contributing to the overall workload. As the drivers do not have to focus on driving they tend to use the reporting system in close physical proximity to the obstacle.

In the forced condition practically all reports are initiated in Area 1, which shows that the participants followed the instructions. As this reporting behaviour does not correspond completely to the drivers' natural strategy, it is important to consider how instructions are phrased in further studies. It is recommended to give participants more leeway in executing secondary tasks, as this enables the employment of compensatory strategies and has more ecological validity.

The ANOVA on workload unexpectedly showed significant differences between the different driving conditions. It is worth noting, however, that there is a decrease in workload as the drivers work their way through the different conditions. This most likely reflects a learning effect with the drivers familiarizing themselves with the road conditions, the obstacles, the simulator and the reporting task over time. Still, the addition of the reporting task does not lead to reported workload levels above baseline, which is promising. This could possibly be explained in light of Salvucci's theory, which suggests that verbal tasks should lend themselves to being integrated with driving. In future studies learning effects have to be addressed more carefully, for example by working with highly trained participants, or by employing a between-group design. Also, performing workload assessment more frequently could provide further insight into how drivers are affected by driving and/or the reporting task.

5 Conclusions

The results show that drivers employ compensatory strategies when making speech-based hazard reports, and that the strategy seems to be dependent on the complexity of the traffic situation. The results also indicate that overall reported workload does not increase when using a speech-based hazard reporting system, providing initial support for the presented theory. However, the scope of the present study was limited, therefore it is recommended to follow up on the results and expand them, taking the recommendations given here into account.

References

1. Maciej, J., Vollrath, M.: Comparison of manual vs. speech-based interaction with in-vehicle information systems. Accident Analysis and Prevention 41, 924–930 (2009)
2. Wickens, C.D.: Situation awareness and workload in aviation. Current Directions in Psychological Science 11, 128–133 (2002)
3. Salvucci, D.D., Taatgen, N.A.: Threaded cognition: an integrated theory of concurrent multitasking. Psychological Review 115, 101–130 (2008)
4. Wu, X., Li, Z.: Secondary task method for workload measurement in alarm monitoring and identification tasks. In: Rau, P.L.P. (ed.) HCII 2013 and CCD 2013, Part I. LNCS, vol. 8023, pp. 346–354. Springer, Heidelberg (2013)
5. Grant, R.C., Carswell, C.M., Lio, C.H., Seales, W.B.: Measuring surgeons' mental workload with a time-based secondary task. Ergonomics in Design: The Quarterly of Human Factors Applications 21, 7–11 (2013)

6. Alm, H., Nilsson, L.: The effects of a mobile telephone task on driver behavior in a car following situation. Accident Analysis and Prevention 27, 707–715 (1995)
7. Beede, K.E., Kass, S.J.: Engrossed in conversation: The impact of cell phones on simulated driving performance. Accident Analysis and Prevention 38, 415–421 (2006)
8. Horrey, W.J., Wickens, C.D.: Examining the impact of cell phone conversations on driving using meta-analytic techniques. Human Factors 48, 196–205 (2006)
9. Caird, J.K., Willness, C.R., Steel, P., Scialfa, C.: A meta-analysis of the effects of cell phones on driver performance. Accident Analysis and Prevention 40, 1282–1293 (2008)
10. Sivak, M.: The information that drivers use: Is it indeed 90% visual? Perception 25, 1081–1089 (1996)
11. Fitch, G.M., Hanowski, R.J.: The risk of a safety-critical event associated with mobile device use as a function of driving task demands. In: Second Conference on Driver Distraction and Inattention, Gothenburg, Sweden (2011)
12. Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E., Dingus, T.A.: Distracted driving and risk of road crashes among novice and experienced Drivers. New England Journal of Medicine 370, 54–59 (2014)
13. Cooper, J.M., Vladisavljevic, I., Medeiros-Ward, N., Martin, P.T., Strayer, D.L.: An investigation of driver distraction near the tipping point of traffic flow stability. Human Factors 51, 261–268 (2009)
14. O'Brien, N.P., Goodwin, A.H., Foss, R.D.: Talking and texting among teenage drivers: A glass half empty or half full? Traffic Injury Prevention 11, 549–554 (2010)
15. Esbjörnsson, M., Juhlin, O., Weilenmann, A.: Drivers using mobile phones in traffic: An ethnographic study of interactional adaptation. International Journal of Human-Computer Interaction 22, 37–58 (2007)
16. Lindström, A., Villing, J., Larsson, S., Seward, A., Åberg, N., Holtelius, C.: The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. Interspeech, Brisbane, Australia (2008)
17. Wickens, C.D.: Processing resources in attention. In: Parasuraman, R., Davies, D.R. (eds.) Varieties of Attention, pp. 63–102. Academic Press, New York (1984)
18. Wickens, C.D.: Multiple resources and mental workload. Human Factors 50, 449–455 (2008)
19. Derrick, W.L.: Dimensions of operator workload. Human Factors: The Journal of the Human Factors and Ergonomics Society 30, 95–110 (1988)
20. Levy, J., Pashler, H., Boer, E.: Central interference in driving: is there any stopping the psychological refractory period? Psychological Science 17, 228–235 (2006)
21. NHTSA: Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Docket NHTSA-2010-0053 (2012)
22. Yager, C.: An evaluation of the effectiveness of voice-to-text programs at reducing incidences of distracted driving. Texas A&M Transportation Institute, The Texas A&M University System, College Station, Texas 77843-3135, Technical Report SWUTC/13/600451-00011-1 (2013)
23. Green, P.A.: Driver interface/HMI standards to minimize driver distraction/ overload. UMTRI, SAE Paper 2008-21-2002 (2008)
24. McCracken, J.H., Aldrich, T.B.: Analyses of selected LHX mission functions: Implications for operator workload and system automation goals. Anacapa Sciences Inc., Research note ASI-479-024-84B (1984)
25. Kun, A.L., Paek, T., Zeljko, M.: The effect of speech interface accuracy on driving performance. Interspeech, Antwerp, Belgium (2007)

26. Green, P.A.: Crashes induced by driver information systems and what can be done to reduce them. In: Conference of the Society of Automotive Engineers (SAE), Warrendale, PA, USA (1999)
27. Shioya, M., Nishimoto, T., Takahashi, J., Daigo, H.: A study of dialogue management principles corresponding to the driver's workload. In: Abut, H., Hansen, J.L., Takeda, K. (eds.) Advances for In-Vehicle and Mobile Systems, pp. 251–265. Springer, US (2007)
28. Villing, J., Larsson, S.: Speech, buttons or both? A comparative study of an in-car dialogue system. In: Third International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Salzburg, Austria (2011)
29. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of Oz studies — why and how. Knowledge-Based Systems 6, 258–266 (1993)
30. Byers, J.C., Bittner, A.C., Hill, S.G.: Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In: International Industrial Ergonomics and Safety Conference, Cincinnati, Ohio (1989)

Chinese Romanization and Its Application in HCI

Zhiwei Feng

Hangzhou Normal University, China
zwfengde2010@hotmail.com

Abstract. Chinese Romanization can transcribe Chinese characters to Romanized Pinyin, It is very useful for natural language processing, documentation, language learning. It became an important tool for human-computer interaction.

Keywords: Chinese Romanization, Pinyin, documentation, Chinese characters, human-computer interaction.

1 Challenges in Computational Processing of Chinese Characters

We are in the information epoch. In this epoch, computer and network play more and more important rule in human life. The language is an effective carrier of information. In information epoch, the computer with only more than 60 years challenged to the Chinese characters with more than 6000 years. The Chinese character is a kind of ideophonographic character. The ideophonographic character is a graphic character that represents an object or a concept and associated sound element. The Chinese characters are a big character set. The most of character set in the world only includes a limited number of characters. The character number included in the character set of different languages is as following (Figure 1). The number of Chinese characters is much more than above languages. Following is the Chinese character number in different Chinese dictionaries from ancient China to Modern China (Figure 2)

The Chinese character number in ZHONGHUA ZIHAI arrives to 85,000, but some Chinese characters in this dictionary only are meaningless or soundless signs, they can't be considered as the authentic Chinese characters. Generally the number of Chinese characters is more than 60,000. It is the biggest character set in the world. In 20th century, some experts try to invent the Chinese typewriter to type Chinese characters. The Chinese character typewriter is different from the Remington typewriter which based on Latin alphabet. It is extremely complicated and cumbersome. For example, the Chinese typewriter invented by Wally Johnson, which now is kept in the office of Vickie Fu Doll, Chinese and Korean Studies Librarian in the East Asian Library of the University of Kansas, USA[1].

[1] Victor Mair, Chinese typewriter, *Language Log*, June 30, 2009.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 406–416, 2014.

Language:	Number of characters:
Latin	26
Slavic	33
Armenian	38
Tamil	36
Birma	52
Thai	44
Lao	27
Tibet	35
Korean	24
Japanese	48

Fig. 1. Natural Languages and Its Number of Characters

Editor:	Dictionary/Year & Number of Chinese characters:	
Xu Shen	说文解字(SHUOWENJIEZI) / 100 A.C.	9,353
Gu Yewang	玉篇(YUPIAN) / 543	16,917
Chen Pengnian	广韵(GUANGYUN) / 1008	26,194
Ding Du	集韵(JIYUN) / 1067	53,525
Mei Yingzuo	字汇(ZIHUI) / 1615	33,179
Chen Tingjing	康熙字典(KANGXIZIDIAN) / 1716	47,043
Zhang Qiyun	中文大字典(ZHONGWEN DAZIDIAN) / 1971	49,888
Xu Zhongshu	汉语大字典(HANYU DAZIDIAN) / 1990	54,678
Leng Yulong	中华字海(ZHONGHUA ZIHAI) /1994	85,000

Fig. 2. Chinese character number in different Chinese dictionaries

The main tray - which is like a typesetter's font of lead type - has about two thousand of the most frequent Chinese characters (Figure 3). Two thousand Chinese characters are not nearly enough for literary and scholarly purposes, so there are also a number of supplementary trays from which less frequent Chinese characters may be retrieved when necessary. The pieces of character type are tiny and all of a single metallic shade in the tray, it becomes a maddening task for typist to find the right character.

Another problem is the principle upon which the characters are ordered in the tray. By radical of Chinese character? By total stroke count of Chinese character? Both of these methods would result in numerous Chinese characters under the same heading. By rough frequency of Chinese character? By telegraph code of Chinese character? Both of these methods need the good memory of typist. Unfortunately, nobody seems to have thought to use the easiest and most user-friendly method of arranging the Chinese characters according to their pronunciation. For all of the above reasons, using a Chinese typewriter was an excruciating experience. Following is a precious photograph of Wally Johnson working at his typewriter (Figure 4). These photos vividly convey the suffering that is associated with using a Chinese typewriter.

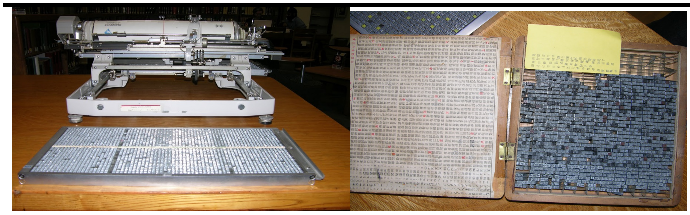


Fig. 3. Wally Johnson’s Chinese typewriter and the tray of typesetter's font of lead type

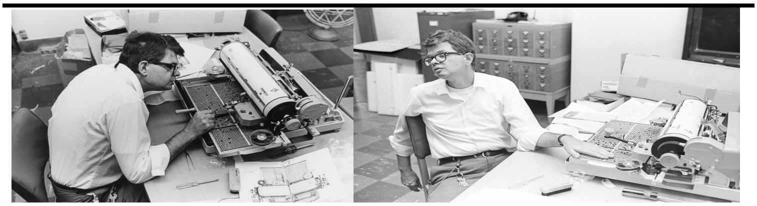


Fig. 4. Wally Johnson working at his typewriter and taking a short break in place

The computer also uses the Remington typewriter as the keyboard for human-computer interaction. Obviously, above Chinese typewriter cannot be used as the keyboard of computer for human-computer interaction. The design of computer keyboard is based on Latin alphabet system. If we use Latin alphabet to represent the pronunciation of Chinese characters, then we can get the easiest and most user-friendly method to input or output the Chinese characters according to their pronunciation. Therefore the Romanization of Chinese is very helpful for human-computer interaction [1].

2 Romanization of Chinese

The words in a language, which are written according to a given script (the converted system), sometimes have to be rendered according to a different system (the conversion system). The conversion is indispensable in that it permits the univocal transmission of a written message between two countries using different writing systems or exchanging a message, the writing of which is different from their own. There are two basic methods of conversion of a system of writing: transliteration and transcription. Transliteration is the operation which consists of representing the characters of an entirely alphabetical character or alphanumeric character system of writing by the characters of the conversion alphabet. In principle, this conversion should be made character by character: each character of the converted alphabet is rendered by one character, and only one character of the conversion alphabet, to ensure the complete and unambiguous reversibility of the conversion alphabet into the converted alphabet (re-transliteration).

Transcription is the operation which consists of representing the characters of a language, whatever the original system of writing, by the phonetic system of letters or signs of the conversion language. A transcription system is of necessity based on the orthographical conventions of a conversion language and its alphabet. The users of a transcription system must therefore have the knowledge of the conversion language to be able to pronounce the characters correctly. Transcription is not strictly reversible. The transcription may be used for the conversion of all writing systems. It is the only method that can be used for systems that are not entirely alphabetical and for all ideophonographic writing systems as Chinese.

Romanization is the conversion of non-Latin writing systems to the Latin alphabet by means of transliteration or transcription. To carry out Romanization it is possible to use either transliteration or transcription or a combination of these two methods, according to the nature of the converted system. Many years ago, in 1958-02-11, the National People’s Congress of China approved The Scheme for the Chinese Phonetic Alphabet (Hanyu Pinyin, or Pinyin)[1][2]. This scheme is based on the principle of the transcription in Romanization. So we call this scheme as Chinese Romanization.

Fig. 5. The Scheme for the Chinese Phonetic Alphabet was approved

3 Pinyin Scheme of Chinese

This scheme provides rules for alphabetic spelling of syllables in Standard Chinese Language of China (*Putonghua*). In the *Pinyin* scheme, each Chinese character generally represents one syllable. One word may consist of one or more syllables. A Chinese syllable can be divided into two parts: initial part and final part (Figure 6). The table of syllabic forms is depicted in Figure 7 and 8.

Initial part of Chinese syllable:

- Bilabial: *b p m*
-Labio-dental: *f*
-Dorso-prepalatal: *d t n l*
-Dorso-velal: *g k h*
-Apico-alveolar: *z c s*
-Apico-postalveolar: *zh ch sh r*
-Dorso-palatal: *j q x*

-Zero initial: nothing before the far left of the final.

Final part of Chinese syllable:

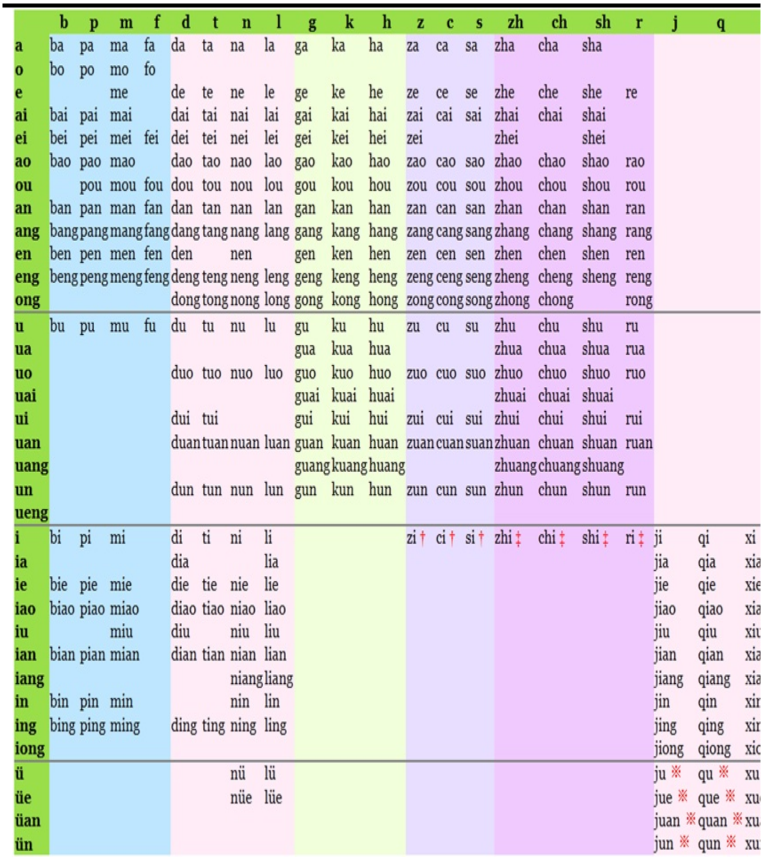
-Articulation A: Articulation with *a*, *o*, *e* as medial or main vowel (For example, *a*, *o*, *e*, *ei*, *ao*, *ou*, *an*, *ang*, *en*, *eng*, *ong, er)* and with *i* in *zi*, *ci*, *si*, *zhi*, *chi*, *shi*, *ri* as main vowel.

-Articulation B: Articulation with *u* as medial or main vowel.
For example, *u*, *ua*, *uo*, *uai*, *ui*, *uan*, *uang*, *un*, *ueng*.

-Articulation C: Articulation with *i* as medial or main vowel.
For example, *i*, *ia*, *ie*, *iao*, *iu*, *ian*, *iang*, *in*, *ing*, *iong*.

-Articulation D: Articulation with *ü* as medial or main vowel.
For example, *ü*, *üe*, *üan*, *ün*.

Fig. 6. Initial part and final part of Chinese syllable



	b	p	m	f	d	t	n	l	g	k	h	z	c	s	zh	ch	sh	r	j	q	
a	ba	pa	ma	fa	da	ta	na	la	ga	ka	ha	za	ca	sa	zha	cha	sha				
o	bo	po	mo	fo																	
e			me		de	te	ne	le	ge	ke	he	ze	ce	se	zhe	che	she	re			
ai	bai	pai	mai		dai	tai	nai	lai	gai	kai	hai	zai	cai	sai	zhai	chai	shai				
ei	bei	pei	mei	fei	dei	tei	nei	lei	gei	kei	hei	zei			zhei		shei				
ao	bao	pao	mao		dao	tao	nao	lao	gao	kao	hao	zao	cao	sao	zhao	chao	shao	rao			
ou		pou	mou	fou	dou	tou	nou	lou	gou	kou	hou	zou	cou	sou	zhou	chou	shou	rou			
an	ban	pan	man	fan	dan	tan	nan	lan	gan	kan	han	zan	can	san	zhan	chan	shan	ran			
ang	bang	pang	mang	fang	dang	tang	nang	lang	gang	kang	hang	zang	cang	sang	zhang	chang	shang	rang			
en	ben	pen	men	fen	den		nen		gen	ken	hen	zen	cen	sen	zhen	chen	shen	ren			
eng	beng	peng	meng	feng	deng	teng	neng	leng	geng	keng	heng	zeng	ceng	seng	zheng	cheng	sheng	reng			
ong					dong	tong	nong	long	gong	kong	hong	zong	cong	song	zhong	chong		rong			
u	bu	pu	mu	fu	du	tu	nu	lu	gu	ku	hu	zu	cu	su	zhu	chu	shu	ru			
ua									gua	kua	hua				zhua	chua	shua	rua			
uo					duo	tuo	nuo	luo	guo	kuo	huo	zuo	cuo	suo	zhuo	chuo	shuo	ruo			
uai									guai	kuai	huai				zhuai	chuai	shuai				
ui					dui	tui			gui	kui	hui	zui	cui	sui	zhui	chui	shui	rui			
uan					duan	tuan	nuan	luan	guan	kuan	huan	zuan	cuan	suan	zhuan	chuan	shuan	ruan			
uang									guang	kuang	huang				zhuang	chuang	shuang				
un					dun	tun	nun	lun	gun	kun	hun	zun	cun	sun	zhun	chun	shun	run			
ueng																					
i	bi	pi	mi		di	ti	ni	li				zi †	ci †	si †	zhi ‡	chi ‡	shi ‡	ri ‡	ji	qi	xi
ia					dia			lia											jia	qia	xia
ie	bie	pie	mie		die	tie	nie	lie											jie	qie	xie
iao	biao	piao	miao		diao	tiao	niao	liao											jiao	qiao	xia
iu			miu		diu		niu	liu											jiu	qiu	xiu
ian	bian	pian	mian		dian	tian	nian	lian											jian	qian	xia
iang							niang	liang											jiang	qiang	xia
in	bin	pin	min				nin	lin											jin	qin	xir
ing	bing	ping	ming		ding	ting	ning	ling											jing	qing	xir
iong																			jiong	qiong	xio
ü							nü	lü											ju ※	qu ※	xu
üe							nüe	lüe											jue ※	que ※	xu
üan																			juan ※	quan ※	xu
ün																			jun ※	qun ※	xu

Fig. 7. Chinese syllable form

⁂ Represents a zero initial (i.e. where nothing comes before the final sound in the far left column)

* Whenever *u* comes at the beginning of a syllable, it is written *w*. However, *w* must not appear without an additional vowel, so *u* as a complete syllable is not written as *w* by itself but as *wu*.

† The *i* in **zi**, **ci**, **si** is different from most other uses of *i* in that it is short, not long.

‡ The *i* in **zhi**, **chi**, **shi**, **ri** is different from most other uses of *i* in that it is short, not long.

+ Whenever *i* comes at the beginning of a syllable, it is written *y*. However, *y* must not appear without an additional vowel, so not *y*, *yn*, *yng* but *yi*, *yin*, *ying*.

※ Hanyu Pinyin simplifies the spellings of syllables with *ü* by using the *u* form instead in cases where no ambiguity could result. This is merely a spelling convention; the *u*'s here are still pronounced *ü*.

[1] **wei**: *ui* is actually an abbreviation of *uei*. This is why *Hanyu Pinyin* uses, for example, *shui*, not *shuei*, and *dui*, not *duei*.

[2] **wen**: *un* is actually an abbreviation of *uen*.

[3] **you**: *iu* is acutally an abbreviation of *iou*. Thus, since *i* is written *y* at the beginning of a syllable, the spelling becomes *you* instead of *yu* (which would be not only misleading but wrong).

Syllable *ê* and retroflexion syllable have been omitted from this table.

Syllable *er* (it is different from the retroflexion syllable) has been omitted from this table.

Fig. 8. Notes to Table in Figure 7

This table covers all syllables of Chinese Putonghua except syllable ê, syllable er and retroflexion syllables. This table includes 392 syllables, plus syllable ê, syllable er and retroflexion syllables, the basic syllables of Chinese Putonghua are 405. The structure of Chinese syllable is simple. It is easy to learn and to remember. Generally speaking, a Chinese character can be represented by a syllable. Therefore we can use the syllables in Pinyin form to represent all Chinese characters in order to realize Chinese Romanization. Because the keyboard of computer is designed on the basis of Latin-alphabet, so we can use Pinyin to represent Chinese character in the human-computer interaction.

4 ISO 7098 Information and Documentation: Romanization of Chinese

In 1979, Chinese delegate proposed to take the scheme of Chinese phonetic alphabet as the international standard in ISO TC46 meeting (Paris, Warsaw). In 1982, *ISO 7098 Documentation and Information – Chinese Romanization* was approved at ISO TC46 meeting (Nanjing) as the first edition. In 1991, ISO 7098 was technically revised. It became the second edition (ISO 7098:1991). In China, *Pinyin*, the international standard for Romanization of Chinese, gives impetus to new information technique in the information epoch. In computer application and mobile communication, it is used to input and output Chinese characters in computer, web and mobile phone. Now more than 80% Chinese used *Pinyin* to deal with Chinese information processing. *Pinyin* became a useful tool for human-computer interaction. In China, *Pinyin* also is effectively used in natural language processing and language engineering (machine translation, information extraction, information retrieval, text data mining, etc.). In the international level, *Pinyin* has been adapted by most libraries around the world. It provides access to bibliographic material of the Chinese language in documentation (including traditional documentation and computerized documentation). In the computerized documentation field, *Pinyin* plays active role in human-computer interaction. In the end of 20 century, Library of Congress (USA) used *Pinyin* to catalogue Chinese books (700,000 books) in the library. In the same time, the *Bibliothèque universitaire des langues et civilisations* in Paris asked a team of sinological librarians from all over the country, including the *Bibliothèque Nationale de France*, to ask their opinion on Chinese word segmentation of ISO 7098, in order to establish a common guideline on Chinese word segmentation in *Pinyin*. The National Library of Australia also adapted *Pinyin* for Chinese Romanization in documentation. Now more and more people in the world learn Chinese as a foreign language by the means of *Pinyin*. *Pinyin* became an important tool for teaching and learning Chinese. In Computer-Assisted Chinese Language Learning, *Pinyin* is used for input and output of Chinese characters in the human-computer interaction.

These facts show, *Pinyin* is a useful tool in human-computer interaction not only in China, but also in the world.

5 Index of Ambiguity for Chinese Syllables

However, the number of basic Chinese syllables is only 405. These 405 Chinese syllables can represent the pronunciation of all Chinese characters (more than 8,000 characters)2. In this case, one Chinese syllable has to represent averagely more than 19 Chinese characters (8,000/405 = 19.75). For example, The Pinyin syllable /bei/ can represent following 66 Chinese characters (Figure 9) and the Pinyin syllable /jing/ can represent following 88 Chinese characters (Figure 10). This means that Pinyin syllable has ambiguity in representation of Chinese characters.

We can use the ambiguity index to describe the degree of ambiguity of *Pinyin* syllable. The ambiguity index of a *Pinyin* syllable (I) equals the number of Chinese characters represented with this *Pinyin* syllable (N) minus 1. The formula is as following:

$$I = N - 1 \tag{1}$$

[2] *General Standardization List of Chinese Characters* (Beijing: Language Publishing House, 2011.) includes 8105 commonly-used Chinese characters.

北	邶	茾	輩	鈚	貝	狈	呗	琪
	钡	垻	唄	鋇	梖	蛽	郥	备
	惫	鞴	俻	偹	備	備	僃	犕
	糒	卑	碑	椑	諀	庳	箄	鞞
	鹎	背	褙	偝	揹	鄁	褙	倍
	蓓	碚	棓	焙	輩	悲	棑	琲
	被	陂	骳	杯	盃	孛	臂	鐴
	㸬	桮	誖	韝	悖	鐾	□	龍
	昁							

Fig. 9. The *Pinyin* syllable /bei/ can represent following 66 Chinese characters

京	惊	猄	澋	燝	綡	鶁	景
	鲸	璟	憬	婛	暻	倞	幜
	经	径	劲	莖	泾	胫	迳
	痉	俓	桱	殌	涇	痙	鵛
	颈	弪	巠	刭	静	精	婧
	菁	晴	靓	儬	腈	箐	睛
	鶄	敬	儆	擏	曔	璥	驚
	警	憼	井	刱	汫	妌	穽
	丼	肼	荆	镜	竟	境	璄
	獍	璄	傹	净	瀞	婙	竫
	婙	晶	楅	粳	伶	旌	劢
	茎	坙	旍	粢	竧	殑	旌
	兢	麠					

Fig. 10. The *Pinyin* syllable /jing/ can represent following 88 Chinese characters

This formula means that if one *Pinyin* syllable can represent N Chinese characters, its ambiguity index (I) equals N – 1. Therefore we may use the ambiguity index of *Pinyin* to describe the ambiguity degree of *Pinyin* syllable in representation of Chinese characters. If one *Pinyin* syllable can represent one Chinese character, its ambiguity index is zero. If one *Pinyin* syllable can represent two Chinese characters, its ambiguity index is 2 – 1 = 1. If one *Pinyin* syllable can represent three Chinese characters, its ambiguity index is 3 – 1 = 2. ...etc. In our example, the *Pinyin* syllable /bei/ can represent 66 Chinese characters, its ambiguity index is 66 – 1 = 65; the *Pinyin* syllable /jing/ can represent 88 Chinese characters, its ambiguity index is 88 – 1 = 87. However if we combine these two monosyllables /bei/ and /jing/ to form a bi-syllabic word /beijing/, the ambiguity index will reduce, because /beijing/ can only represent three Chinese bi-syllabic words:

北京, 背景, 背静

The ambiguity index of /beijing/ reduced to 3 –1 = 2. And if we capitalize the first letter of /beijing/ as /Beijing/, the ambiguity index will be reduced to 1 – 1 = 0. It means that /Beijing/ is a *Pinyin* word without ambiguity, its sense number is only 1. The sense of /Beijing/ exactly is the name of the capital of China:

北京

Therefore if we link different *Pinyin* monosyllables to form a polysyllabic Chinese word, the ambiguity index of *Pinyin* syllable will be reduced. It is the advantages of linking different monosyllables to form one polysyllabic Chinese word. However, at present days, in Chinese linguistics, there is not clear definition of common Chinese word, it is difficult to decide the boundary (dividing line) of a common Chinese word, and of course it will bring the difficulty to link the monosyllables to form a polysyllabic common Chinese word. But the boundary of Chinese proper noun is relatively clear. It is not so difficult to link different monosyllables to form a Chinese polysyllabic proper noun (the naming entity as personal names, geographic names, language names, ethnic names, tribe names, religion names … etc), because the boundary of Chinese polysyllabic proper noun is easy to decide according to the standards or regulations of China. By this reason, at the 38th plenary meeting of ISO/TC 46 (6 May 2011, Sydney), the Chinese delegate proposes to further update ISO 7098:1991 to reflect current Chinese Romanization practice and new development not only in China, but also in the world. At the 39th plenary meeting of ISO/TC 46 (11 May 2012, Berlin), ISO TC 46 resolves to accept the China's proposal at Working Draft (WD) stage. In 5 November 2013, the CD ballot is approved. At the 41th plenary meeting of ISO/TC 46 (5 May 2014, Washington D. C.), the Chinese delegate shall submit the Draft of International Standard (DIS) revised according to the comments at the CD ballot stage. In ISO 7098 updating version, Chinese delegate proposed and shall propose the following suggestions for the transcription rules of personal names, geographic names, language names, ethnic names, tribe names and religion names in Chinese language. We believe that this kind of transcription for the naming entity will be the first step for Chinese transcription based on the Chinese word (including polysyllabic common word and polysyllabic proper noun, etc).

6 Suggestions for Updating ISO 7098

We shall propose following suggestions (Suggestions 6.1- 6.11) for updating ISO 7098: (6.1)Chinese personal names are to be written separately with the surname first, followed by the given name written as one word, with the initial letters of both capitalized. The traditional compound surnames are to be written together without a hyphen. The double two-character surnames are to be written together with a hyphen and the initial letters of both capitalized. For example, Li Hua (李华), Wang Jianguo (王建国), Zhuge Kongming (诸葛孔明), Zhang-Wang Shufang (张王淑芳). Pen names and other aliases are to be treated in the same manner: For example, Lu Xun (鲁迅), Wang Pangzi (王胖子). (6.2)The surname, given name, seniority order after the adjuncts “Xiao”, “Lao” are to be written separately and with the initial letter both capitalized. For example, Xiao Liu (小刘， younger Liu), Lao Qian (老钱，older Qian). (6. 3)Certain proper names and titles have already fused and are written as one word with the initial letter capitalized. For example, Kongzi (孔子, Master Confucius), Xishi (西施， acme of beauty, 5th cent. B.C.). (6.4)Chinese place names should separate the geographical proper name from the geographical feature name and capitalize the first letter of both. For example, Beijing Shi (北京市, Beijing Municipality), Hebei Sheng (河北省, Hebei Province). (6.5) If a geographical proper name or geographical feature name has a monosyllabic adjunct, write them together as one word. For example, Jingshan Houjie (景山后街, Jingshan Back Street), Chaoyangmennei Nanxiaojie (朝阳门内南小街, South Street inside Chaoyangmen Gate). (6.6)The names of smaller villages and towns and other place names in which it is not necessary to distinguish between the proper place name and the geographical feature name are to be written together as one unit. For example, Wangcun (王村， Wang Village) , Zhoukoudian (周口店， an historical site).(6.7)In accordance with the principle of adhering to the original, non-Chinese personal names and place names are to be written in their original Roman (Latin) spelling. While personal names and place names from non-Romanized scripts are to be spelled according to the rules for Romanization for that language. For reference, Chinese characters or their Hanyu Pinyin equivalent may be noted after the original name. Under certain conditions, the Hanyu Pinyin may precede or replace the original spelling. For example, Marx (马克思, Makesi), Pairs (巴黎, Bali). (6.8)Transcribed names which have already become Chinese words are to be spelled according to their Chinese pronunciation. For example, Feizhou (非洲, Africa) , Nanmei (南美, South America), Deguo (德国, Germany), Dongnanya (东南亚, Southeast Asia). (6.9)In some cases, all the letters in personal name and geographical name may be capitalized. For example, BEIJING (北京， Beijing), LI HUA (李华，Li Hua). (6.10)In the abbreviation of personal names, the surnames are to be written with initial capitalized letter or with all capitalized letters; the given names are to be written with first capitalized letter in every syllables and are to be added a dot after the capitalized letter. For example, Li H. or LI H. for Li Hua (李华) , Wang J.G. or WANG J.G. for Wang Jianguo (王建国). (6.11)The abbreviation of geographical names written together as one word, is to be written with first capitalized letter in every syllable; all capitalized letters in the syllable are to be linked together. For example, BJ for Beijing (北京), HZ for Hangzhou (杭州).

The detailed spelling rules of personal names and geographical names should be alphabetized according to the regulations *Spelling Rules for Chinese Personal Names* and *Spelling Rules for Chinese Geographical Place Names (the part of Chinese Geographical Names).* The detailed spelling rules of common words are more complex than the rules of these proper nouns (naming entity). The rules of *pinyin* orthography for Chinese common words are included in the National Standard of China *Basic Rules for Hanyu Pinyin Orthography (GB/T 16159-2011).* This National Standard will further give impetus to the Chinese Romanization.

The Chinese Romanization will play more and more important roles in human-computer interaction.

References

1. Scheme of Chinese phonetic alphabet, Selections of Norms and Standards for Language and Script of China, p. 441. Standards Press of China, Beijing (1997)
2. Directives for the promotion of Putonghua, promulgated by the State Council of China, Selections of Norms and Standards for Language and Script of China, pp. 439–440. Standards Press of China, Beijing (1997)

Driving with a Speech Interaction System: Effect of Personality on Performance and Attitude of Driver

Ing-Marie Jonsson and Nils Dahlbäck

Department of Computer and Information Science, Linköping University, SE-581 83
Linköping, Sweden
ingmarie@ansima.com, nils.dahlback@liu.se

Abstract. Personality has a huge effect on how we communicate and interact with others. This study is one in a series of three that investigates how a speech based in-car system matched with dominant and submissive drivers affects performance and attitude drivers. The study was conducted with 30 participants at Linköping University in Sweden. Data show that using a voice that combines feature from submissive and dominant speech patterns work well for both dominant and submissive drivers. The voice showed the same performance gain as when matching car voice personality with personality of driver, without the negative attitude ratings associated with the submissive car voice found in previous studies. Drivers assessment of the car system show that even though both dominant and submissive drivers find the system helpful, dominant drivers find the system more annoying and more likely to turn the system off. Design implications of in-vehicle systems are discussed.

Keywords: In-car System, Driving Simulator, Driving Performance, Speech system, Attitude, Personality, Dominant and Submissive.

1 Introduction

Humans can easily detect characteristics in a voice and will use that skill when communicating with both humans and speech-based computer systems [1]. The linguistic and para-linguistic properties of a voice can influence people's attention and affect performance, judgment, and risk-taking [2, 3]. Previous studies show that voices used by in-car systems can influence driving performance and driver attitude [4, 5, 6]. Characteristics of the voice affects listeners perception of liking and credibility of what is said, regardless of if the speaker is human or computer-based system [3]. "Speaking is the most social and human thing we do", stated Professor Clifford Nass, professor and director of the Communication between Humans and Interactive Media Lab at Stanford University. "The minute you start speaking or listening to speech, the part of your brain that associates 'humanness' kicks in."[7]

In the context of in-car information systems, Nass et al. [8] show a clear positive effect of matching the emotional characteristics of the in-car voice to the emotional state of the driver. People prefer people to interact with people that are like themselves; it makes it easy to establish common ground and to communicate. Lazarsfeld

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 417–428, 2014.

and Merton [9] showed that most successful human communication will occur between a source and a receiver who are alike, i.e., homophilous, and have a common frame of reference.

In general terms, theories of similarity-attraction and consistency-attraction [10] would suggest that personality has a huge effect on how we communicate and interact with others. Previous studies show that matching personality when communicating with a computer systems matters [11] and Dahlbäck, Swamy et al. [12] show that even matching accents matters. A system is always rated higher, and the user's perception of the systems performance better in matched cases. For in-car systems and driving performance, Jonsson and Dahlbäck [13], show a clear positive effect on driving performance when matching personality of the in-car voice with personality or driver. There is however a complex interaction between personality, perceived similarity, attitude and performance. Even though performance numbers are better for matched conditions, attitude towards the in-car systems does not necessarily improve with matched conditions.

To further investigate the effects of matching personality of in-car system with personality of driver. The authors designed an in-car system exhibiting properties that can be considered personality neutral, i.e. rating in the neutral zone between dominant and submissive.

The study reported here was designed to investigate if the voice of an in-car system, rated to be neither dominant nor submissive, would be perceived similar enough to trigger positive effects of similarity-attraction on driving performance without exhibiting negative effects on attitude.

2 Study Design and Apparatus

To investigate the effect of a personality-neutral voice on dominant/extrovert and submissive/introvert drivers a study with 30 participants was designed. The study was conducted at Linköping University in Sweden and is a follow-up of a study conducted at Oxford Brookes University in the UK [13].

2.1 Study Design and Participants

The design was a 1 (personality of car voice) x 2 (Personality of driver: dominant, submissive) between subject and gender balanced study.

There were 30 participants in the study (18 assessed as extrovert/dominant and 12 as introvert/submissive) Participants were screened based on the NEO-FFI [14]. It is an abbreviated version of the NEO Personality Inventory Revised (NEO-PI-R) (Costa & McCrae, 1992). It is intended for individuals aged 17 and older and requires a sixth grade reading level. The test items take the form of first person statements which participants are asked to rate on a five point Likert scale ranging from "Strongly Disagree" to "Neutral" to "Strongly Agree." The inventory typically takes 10-15 minutes to complete (Costa & McCrae, 1992).

All participants were students at Linköping University and they were awarded 140SEK for their participation.

2.2 Apparatus

Driving Simulator

The study was done using a driving simulator. This means that results provide an indication rather than a determination of behavior in real cars and real traffic.

There are many factors that motivate the use of driving simulators, the most pertinent being the ability to fully control the experimental setting and driving environment. The average driver will have very few accidents in their lifetime despite the dangers involved in driving. Due to the rarity of incidents, it would be extremely time consuming to set-up an experiment with the characteristics of real driving within the defined parameters of the study, and wait for a significant number of events to occur. Hence, the best way to examine new in-car systems is to challenge people using a driving simulator. Even though the degree of immersion varies with the fidelity of the simulator, the immersive effect is there even for very low fidelity simulators [15].

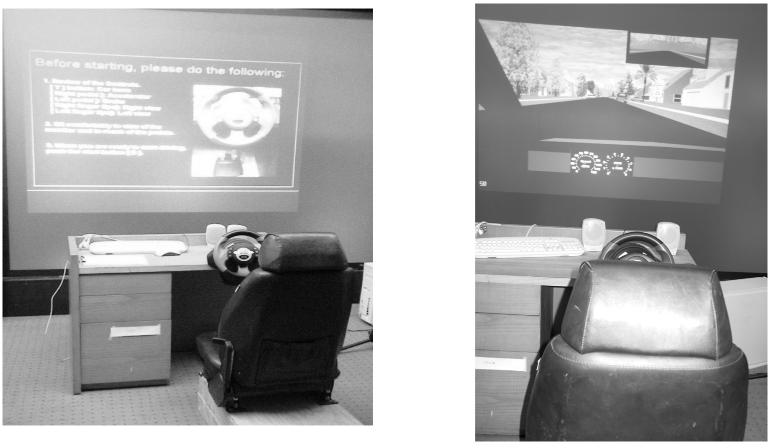
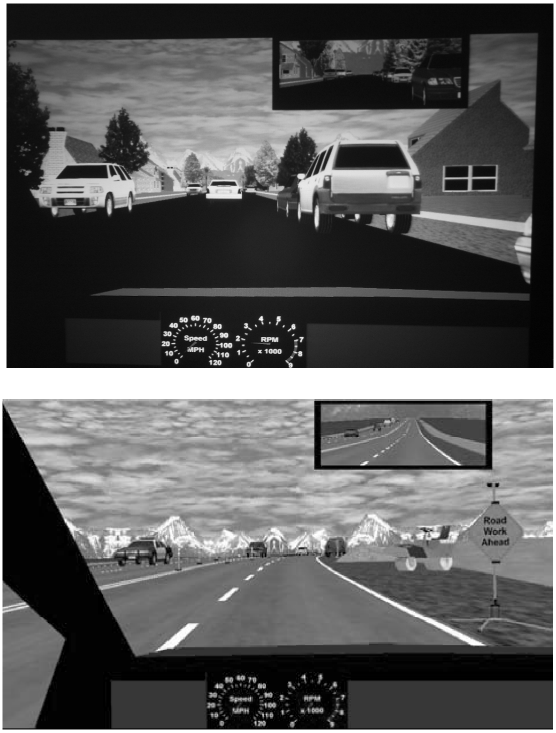


Fig. 1. STISIM Drive - Driving simulator. Setup Screen and random road scene depicted.

A commercial driving simulator, STISIM Drive model 100 with a 45-degree driver field-of-view, from Systems Technology Inc. was used in the studies. Participants sat in a real car seat and "drove" using a Microsoft Sidewinder steering wheel and two pedals, accelerator and brake. The authors selected a driving experience based on an automatic gearbox, and the simulated driving course was viewed on three large screen monitors in front of participants. The screens were setup as one screen right front, and two screens angled towards the driver on the left and right side respectively.



The view from the driver seat. There are two gauges visualized at the bottom of the screen, a tachometer and a speedometer. Please note the rearview mirror located at top of screen.

Traffic can either be programmed to follow traffic regulations or drive without adherence to traffic regulations. This includes behavior at stop signs, traffic lights and driving speed.

Fig. 2. STISIM Drive – Properties of driving setup and traffic

Driving scenarios in STISIM Drive consist of a road with objects placed along that road. Note that a driving scenario in STISIM Drive is static. Drivers are driving the exact same road regardless of if they turn left, right or continue straight ahead at any intersection along the way. This ensures a consistent and repeatable driving environment from start to finish for all participants.

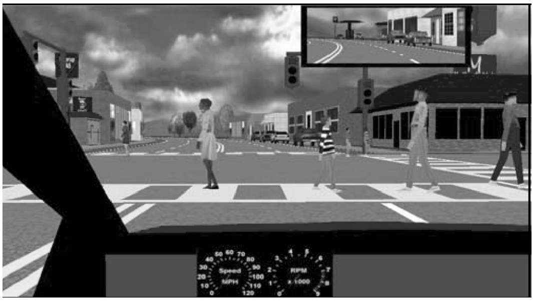


Fig. 3. STISIM Drive – Driving scenario with a small village, an intersection and pedestrians

The driving scenario that was used was the same as in previous studies on personality of voice in cars [13]. It is a varied and realistic road scenario of 52 000 feet (15.85 kilometers), especially designed to take the drivers through rural areas, villages and intersections. In addition to driving the exact same scenario, all properties of the simulator, car, vehicle dynamics, weather conditions and traffic were set to be the identical for all participants.

In-Car System. The authors used the same navigation system as designed for previous studies on personality of voice in cars [13]. It takes the driver to five locations by interacting with drivers at certain locations along the way.

The navigation system consists of 40 utterances. 30 of the utterances are directions or suggestions, and 10 utterances are facts about the immediate surroundings. Directions and suggestions were designed to guide the drivers to the pre-programmed destinations. The facts were added to investigate how much attention drivers were paying to the system and the voice. All 40 utterances were translated to Swedish.

The Swedish voice that was used by the navigation system was selected to be personality neutral, neither rated as dominant, nor rated as submissive. The linguistic features used by the voice were a mix between those used by a dominant and a submissive voice. Choice of words was selected to match the dominant style. Using words such as “will”, “must” and “definitely, in contrast to submissive style words such as “might”, “could” and “perhaps”. Overall the navigation system used assertive language “You should definitely turn right” in contrast to the submissive language style of “Perhaps you should turn right”.

The voice was then recorded with lower overall frequency, flat pitch range and slower speed than a typical dominant voice [11]. The male voice used for the systems was reviewed and rated on the same NEO FFI inventory [14] used to screen participants.

3 Procedure and Measures

3.1 Procedure

All participants were informed that the experiment would take one hour and started the experimental session by signing a consent form. This was followed by a five-minute test run of the simulator, where participants could familiarize themselves with the simulator and the controls. This enabled participants to experience feedback from the steering wheel, the effects of the accelerator and brake pedals, a crash, and for us to screen for participants with simulator sickness [16]. None of the signed up participants felt nauseous or discomfort during the training course. All 30 participants proceeded to fill in the first questionnaire consisting of general information such as gender and age and real-life driving experience.

In this study, all participants but one drove the driving simulator from start to finish. One participant retired from the diving session due to simulator sickness. The remaining 29 completed the driving scenario with the exact same navigation system using the same voice, scripted to take the driver to five destinations. During the drive all participants were subjected to the factual information inserted at 10 locations along the road.

After the driving session, participants filled in a set of post driving questionnaires. One of the questionnaires asked participants to assess the voice of the navigation system in terms of how similar it was to them. A second questionnaire asked the driver to assess their driving experience and how the navigation system was perceived to affect their driving performance. The final questionnaire asked participants to recall information volunteered by the navigation system during the drive.

3.2 Measures and Dependent Variables

This study used the same measures for personality, similarity, driving performance and navigation system as used in previous studies on personality of voices in cars, [13]. The authors used these measures in all three personality studies in this suite of studies to ensure consistency and enable comparisons between the different studies.

Personality

Participants were screened based on the NEO FFI inventory [14]. The inventory consists of 60 first person statements which participants were asked to rate on a five point Likert scale ranging from “Strongly Disagree” to “Neutral” to “Strongly Agree.” The NEO inventory measures differences among normal individuals, and will assess individuals on the five factors or dimensions of the five-factor model (FFM) of personality.

Similarity

Similarity-attraction is an important aspect of how voices influence attitude and perception of spoken messages. Similarity-attraction predicts that people will be more attracted to people matching themselves than to those who mismatch. It is a robust finding in both human-human and human-computer interaction [9, 11]. The theory predicts that users will be more comfortable with computer-based personas that exhibit properties that are similar to their own. Attraction leads to a desire for interaction and increased attention in human-computer interaction [17, 18].

A standard questionnaire on homophily [19] was used to assess similarity. The index for similarity used in the study was constructed as a combination of attitudinal similarity and behavioral similarity. Participants were asked to rate the statements of the inventory based on the question "On the scales below, please indicate your feelings about the person speaking?" Contrasting statements were paired on opposite sides of a 10-point scale such that, 'similar to me' and 'different from me' would appear at different ends.

Driving Performance. This is a collection of measures that consists of accidents and adherence to traffic regulations. The driving simulator automatically collected the data for these measures. Accidents is comprised off-road accidents, collisions, and pedestrian incidents. Adherence to Traffic Regulations is comprised of speeding, running stop signs, and running red lights.

Because it is much more difficult to drive in a simulator than to drive a real car in real traffic, the number of incidents are much higher than in real traffic, which makes this a useful measure of driving performance.

Navigation System. This is a collection of measures related to the voice used by the navigation system and how drivers perceive and react to it. The measure *Instructions followed* simply counts how many of the driving instructions drivers followed. There were a total of 30 instructions given by the system to navigate the driver from start to finish. *Time to destination* measures drivers' time to complete the driving scenario to the last destination. *Facts remembered* measures how many of the 10 driving scenario facts that drivers remembered after the driving session ended.

Driver Self-Assessment and Perception of Navigations System. Participants self-assessed their Normal driving style based on 8 terms using a 10-point Likert scale. In addition to this, participants also rated the perceived Influence by navigation system on their driving performance using a 10-point Likert scale for 9 terms.

Participants were specifically asked assess the driving session and navigation system. The driving session rated in terms of Fun and Liking, the navigation system in terms of being Annoying and Helpful. Finally, participants were asked to disclose their Willingness to use, i.e. to install and use in their own cars.

4 Results

The effects of using a "neutral" car voice in a navigation system with personality of drivers were measured by a one (Personality of Navigation System voice) by two (Personality of Driver) between-participants ANOVA.

4.1 Prior Driving Experience

To ensure that there was no bias based on drivers' prior driving experience, data from the two most recent years of driving was collected. The data that included number of miles driven per year, number of accidents, and number of tickets, was averaged for each group of drivers. No significant differences were found across conditions.

4.2 Similarity – Homophily

Data from the similarity assessment show that both groups of drivers felt similar to the car voice. There was no significant difference between the two groups of drivers, dominant drivers felt similar to the person behind the car voice Mean=5.9 SD=1.1, and submissive drivers felt equally similar to the person behind the car voice Mean=5.9, SD=1.0, $F(1, 28) = 0.006$, $p < 1.0$.

4.3 Driving Performance

Bad Driving. There was no significant difference between the two groups of drivers on the bad driving indices, accidents and adherence to traffic regulations. There was no significant difference between *Accidents* for dominant and submissive drivers. Dominant drivers show Mean=3.5, SD=2.0, submissive drivers show Mean=3.1, SD=1.3, F(1, 28)= 1.8, p < 0.7. Similarly, there was no significant difference between adherence to traffic regulations between dominant drivers (Mean=7.8, SD=5.5) and submissive drivers (Mean=11.1, SD=5.9), F(1, 28)= 2.1, p < 0.2.

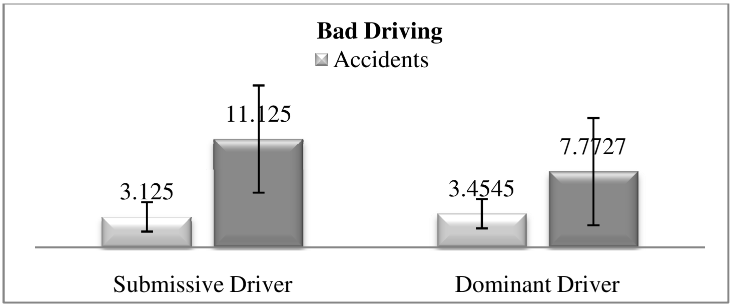


Fig. 4. Bad driving- accidents and adherence to traffic regulations

4.4 Navigation System

Instructions Followed. Data show that there was no significant difference between the two groups of drivers when following instructions. Dominant drivers follow the same amount of instructions (Mean=25.1, SD=1.8) as submissive drivers (Mean=24.4 SD=2.0), F(1, 28) = 0.84, p < 0.4.

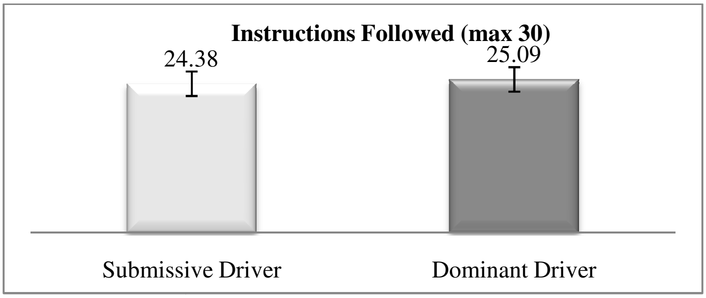


Fig. 5. Instructions followed

Facts Remembered. Both submissive drivers and the dominant drivers paid attention to and listened to the voice equally.

Submissive drivers and dominant drivers remembered approximately 80% of the facts uttered during a 25-minute drive. Mean for submissive drivers was 7.9 (SD = 1.2) and mean for dominant drivers was 8.0 (SD = 1.2), F(1, 28) = 0.04 p < 0.9.

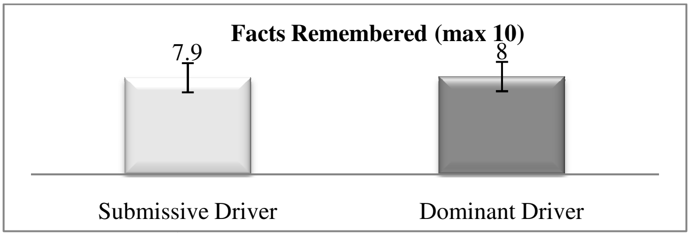


Fig. 6. Facts Remembered

Time to Destination. The driving simulator automatically collected completion time, i.e. the time it took for drivers to reach their fifth and final destination. There was no significant difference between how long it took submissive drivers (M=24min. 20sec., SD=3min. 20sec.) and dominant drivers (Mean= 25min. 58sec., SD=3min. 30sec.) to reach the last destination, F(1, 28) = 1.3, $p < 0.3$.

4.5 Driver Self-Assessment

Normal Driving Style. Results from participants' self-assessment of their normal driving style show no significant difference between submissive and dominant drivers. Submissive drivers with Mean=4.8 (SD=0.7) and dominant drivers with Mean=5.3 (SD=1.3), F(1, 28)=1.1, $p < 0.3$.

Influence by Navigation System. Data from participants rating the influence of the navigation system on their driving performance show no significant difference between the two groups of drivers. Assessing the positive influence, both submissive (Mean= 5.8, SD=1.2) and dominant (Mean=6.5, SD=1.2) drivers perceived that the system made them slightly more safe and careful drivers, F(1, 28) = 2.2, $p < 0.15$, than their normal driving style.

Fun and Liking. Data from participants rating their experience with the driving session show that both submissive drivers (Mean=8.2, SD=1.5) and dominant drivers (Mean=7.4, SD=1.5) had fun and liked the driving session.

Navigation System. There was a significant difference in how annoying the two groups of drivers found the navigation system. Dominant drivers found the navigation system to be more annoying (Mean=5.7, SD=2.4), than submissive drivers (Mean 2.8, SD=1.4), F(1, 28)= 10.6, $p < 0.005$. Please note that both groups of drivers found the navigation system to be helpful (Mean=7.0, SD=2.3 and Mean=7.4, SD=1.7), F(1,28)=0.32, $p < 0.6$.

Willingness to Use. When specifically asked if they wanted to install and use the system in their own cars, there were also significant differences between the two groups. Submissive drivers (Mean=8.2, SD=0.9) were more willing than dominant driver (Mean=7.0, SD=1.2) to install in the car, F(1, 28) = 6.7, p < 0.01.

Dominant drivers (Mean=4.8, SD=2.0) were also more likely to turn the system off than submissive drivers (Mean=2.4, SD=0.8), F(1, 28) = 10.1, p < 0.005.

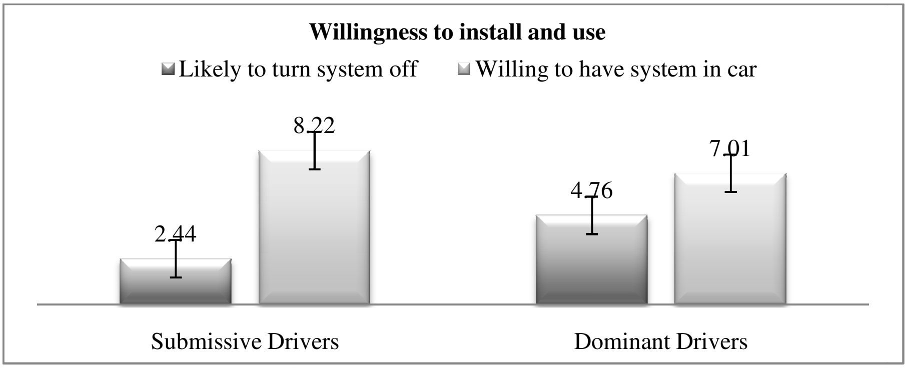


Fig. 7. Willingness to install and use system in own car

5 Conclusions and Discussion

Results from previous studies show that similarity-attraction predicts drivers' performance and attitude, for emotional drivers. The theory however, only partially predicts performance and attitude when matching personality of drivers with personality of car voice [13]. Result from this study show that selecting a car voice neutral on the dimension of dominant/submissive retains the performance benefits seen in matched conditions and lessening the negative attitudinal effects. Both submissive and dominant drivers feel similar to and like the car voice. The differences emerge with the unwillingness and resistance that dominant drivers exhibit in accepting instructions and advice from the system.

Data from a previous study [13] show that both dominant and submissive drivers felt less at ease after driving with the submissive voice, than after driving with the dominant voice. Data from the current study show that submissive drivers had a more positive experience with the personality-neutral car voice than with the submissive car voice in the matched case scenario [13]. Making the car voice personality neutral, reduced the negative impact on perception and willingness to listen, but did not entice dominant drivers to engage and interact. Data furthermore showed that dominant drivers were annoyed by the systems, and that they also were significantly more likely to turn the system off than submissive drivers.

The data from this, and previous studies, show complex interactions between personality, perceived similarity, attitude and performance. It emphasizes that it is important, to find the balance between matching-efforts and efficacy. Having a system that can accurately match drivers' personalities, is a remarkable technological feat, if drivers are not positively influenced by it on all dimensions, it is however a wasteful expense. Previous studies [13] showed that a system could be perceived as annoying and undesirable, regardless of its actual performance.

As one study in a suite of personality based studies investigating effects of matching car voice with drivers, this study refines attitudinal results. Even though the data clearly show improvements over the matched cases investigated in a previous study [13], there are still more dimensions to be investigated. Dominant drivers perceive the in-car system tested in this study as a mixed blessing, seen as both helpful and annoying.

The bottom line is that even the technologically-best system may not satisfy or help all drivers: While in-vehicle information systems represent exciting technological advances, their deployment should be guided by significant caution.

References

1. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636 (1996)
2. Jonsson, I.-M., Nass, C., Endo, J., Reaves, B., Harris, H., Le Ta, J., Chan, N., Knapp, S.: Don't blame me I am only the Driver: Impact of Blame Attribution on Attitudes and Attention to Driving Task. In: SIGCHI, pp. 1219–1222. ACM Press (2004)
3. Nass, C., Brave, S.: Wired for speech" how voice activates and advances the human-computer relationship. MIT Press, Cambridge (2005)
4. Zajicek, M., Jonsson, I.-M.: A Complex Relationship, Older People and In-Car Message System Evaluation. Journal of Gerontology 6, 66–78 (2007)
5. Jonsson, I.-M.: Conversational Interfaces and Driving: Impact on Behaviour and Attitude. In: IASTED Human-Computer Interaction, pp. 224–229 (2008)
6. Jonsson, I.-M., Dahlbäck, N.: The effects of different voices for speech-based in-vehicle interfaces: Impact of young and old voices on driving performance and attitude. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, pp. 2795–2798 (2009)
7. La, L.: CNET Reviews, iSheep, Fandroids, and why we care so damn much about oru smartphones. Online Magazine (November 2013), http://reviews.cnet.com/8301-6452_7-57612654/isheep-fandroids-and-why-we-care-so-damn-much-about-our-smartphones/
8. Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L.: Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In: SIGCHI, pp. 1973–1976. ACM Press (2005)
9. Lazarsfeld, P., Merton, R.: Mass Communication, Popular Taste, and Organized Social Action. In: The Communication of Ideas, pp. 95–188 (1948)
10. Byrne, D.: The Attraction Paradigm. Academic Press, New York (1971)
11. Nass, C., Lee, K.M.: Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In: SIGCHI, pp. 329–336. ACM Press (2000)
12. Dahlbäck, N., Swamy, S., Nass, C., Arvidsson, F., Skågeby, J.: Spoken Interaction with Computers in a Native or Non-native Language - Same or Different? In: Proceedings of INTERACT, pp. 294–301 (2001)

13. Jonsson, I.-M., Dahlbäck, N.: In-Car Information Systems: Matching and Mismatching Personality of Driver with Personality of Car Voice. In: Kurosu, M. (ed.) HCII/HCI 2013, Part II. LNCS, vol. 8005, pp. 586–595. Springer, Heidelberg (2013)
14. Costa Jr., P.T., McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources, Odessa (1992)
15. de Winter, J., van Leuween, P., Happee, P.: Advantages and Disadvantages of Driving Simulators: A Discussion. In: Proceedings of Measuring Behavior, pp. 47–50 (2012)
16. Brooks, J., Goodenough, R., Crisler, M., Klein, N., Alley, R.: Simulator sickness during driving simulation studies. Accident Analysis and Prevention 42, 788–796 (2010)
17. Dahlbäck, N., Wang, Q., Nass, C., Alwin, J.: Similarity is More Important than Expertise: Accent Effects in Speech Interfaces. In: SIGCHI, pp. 1553–1556. ACM Press (2007)
18. Nass, C., Lee, K.: Does Computer synthesized speech manifest personality? Experimental tests of recognition, similarity attraction and consistency attraction. Journal of Experimental Psychology: Applied 7, 171–181 (2001)
19. Rubin, R., Palmgreen, P., Sypher, H.: Communication Research Measures: A Sourcebook. Guilford Press, New York (1994)

Effects of Language Variety on Personality Perception in Embodied Conversational Agents

Brigitte Krenn[1], Birgit Endrass[2], Felix Kistler[2], and Elisabeth André[2]

[1] The Austrian Research Institute for Artificial Intelligence,
Freyung 6/6, A-1010 Vienna, Austria
brigitte.krenn@ofai.at
[2] University of Augsburg,
Universitätsstr. 2, 86159 Augsburg, Germany
{endrass,kistler,andre}@hcm-lab.de

Abstract. In this paper, we investigate the effects of language variety in combination with bodily behaviour on the perceived personality of a virtual agent. In particular, we explore changes on the extroversion-introversion dimension of personality. An online perception study was conducted featuring a virtual character with different levels of expressive body behaviour and different synthetic voices representing German and Austrian language varieties. Clear evidence was found that synthesized language variety, and gestural expressivity influence the human perception of an agent's extroversion. Whereby Viennese and Austrian standard language are perceived as more extrovert than it is the case for the German standard.

Keywords: virtual agents, personality, extroversion-introversion, language variety and non-verbal behaviour.

1 Introduction

In the present contribution, we address cultural implications of multimodal expressive behaviours in artificial agents. In particular, we investigate effects of language variety in combination with linguistic and gestural expressivity on the assessment of a virtual agent's personality on the dimension extroversion-introversion. As regards language variety, we concentrate on German and Austrian standard and Viennese dialect. While the combination of linguistic and bodily expression of extroversion in virtual agents has already been assessed in previous work, e.g. (Neff et al. 2010, Isbister and Nass 2000), studying the effects of language variety on an agent's perceived personality is novel.

In the past, synthetic voices available for text-to-speech systems typically have represented standard varieties, e.g. voices for the German, the British English, the American English standard, etc. More recently, localized standard varieties have been made available. Examples are Cereproc's synthetic voices for Scottish, Northern and Southern British, Irish, Catalan, and Austrian German (https://www.cereproc.com,

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 429–439, 2014.

last accessed 7.2.2014). Moreover, the creation of synthetic voices representing varieties of smaller regions is pursued in research contexts. See, for example, work on Austrian varieties (Neubarth et al. 2008, Pucher et al. 2010a, Pucher et al. 2010b). Under these preconditions, localizing virtual agents technically becomes feasible. At the same time, this opens up questions regarding the effects of language variety on the human perception and socio-emotional evaluation of such an agent. Language attitude studies are a well-established means to assess human evaluation of language varieties. Results from a major, recent language attitude study on German varieties (natural voices), for instance, indicate that speakers of Bavarian are perceived as more extrovert (high in spirits) than speakers of the German standard (Gärting et al. 2010). In a study on synthetic standard Austrian German and Viennese dialectal varieties, (Krenn et al. 2012) demonstrated that: (i) language attitudes towards natural voices transfer to synthesized voices, and (ii) the dialectal Viennese variety is characterized by attributes that also characterize extrovert behaviour. For instance: the dialectal voice compared to the standard Austrian voice is associated with sense of humour, emotionality, self-confidence, open-mindedness and an easy-going nature. These findings together with the fact that Bavarian and Austrian both are varieties of the same German dialect, East Upper German, theoretically back the hypotheses that language variety influences the perceived personality of an agent regarding the dimension extroversion-introversion.

In the next chapter we summarise the theoretical background of our work. Afterwards, we explain how we implemented the introvert and extrovert versions of our embodied conversational agent. In Section 4, we describe our study design and execution, followed by its analysis and interpretation. In the last section, we summarise our findings.

2 Theoretical Background

2.1 Expression of Extroversion-Introversion in Language Variety

Results from Gärtig et al. (2010) indicate that speakers of German standard variety are evaluated as friendly, educated and calm. Speakers of Bavarian are evaluated as even more friendly as speakers of the German standard and in contrast to speakers of the German standard variety Bavarians are perceived as full of spirit. For a summary of the findings see Table 1. 1017 participants rated the "typical German", 501 participants rated the “typical Bavarian”.

Krenn et al. (2012) show that language attitudes towards synthesized voices representing local varieties are comparable to language attitudes towards natural voices. 91 subjects of (Austrian) German mother tongue were presented with a semantic differential comprising 19 adjective pairs representing positive and negative extremes on a value dimension, where each pair had to be rated on a 5-point Likert scale. Comparing two synthetic varieties (an Austrian standard male voice and a dialectal Viennese male voice) the following differences could be identified applying a Wilcoxon tests for pairwise comparison and Bonferroni correction for multiple comparisons. Significant differences were found for all dimensions of the semantic differential except for likeability, friendliness and arrogance. Summing up, the Austrian standard is evaluated as more trustworthy, competent, polite, intelligent, educated and serious as the dialectal Viennese. Whereas the dialectal Viennese is evaluated as more emotional, self-confident, natural, relaxed, with more sense of humour and less strict than the Austrian standard, but also as more aggressive, less gentle and less refined. See for an overview. More information can be found in (Krenn et al. 2012).

Table 1. Evaluation of Standard German speakers and speakers of Bavarian on the dimensions friendliness, educatedness and spiritedness. Summary from Gärtig et al. 2010, cf. pages 103, 106, 109, 113, 116, 119.

Dimension	**Typical German** N=1017	**Typical Bavarian** N=501
Friendliness		
(very) friendly	38,7 %	52.5 %
partly	51.4 %	31.7 %
(very) unfriendly	8.3 %	8 %
Educatedness		
(very) educated	45 %	40.8 %
partly	48,9 %	43.4 %
(very) uneducated	4.1	2.6 %
Spiritedness		
(very) spirited	14.5 %	50.6 %
partly	51 %	27.2 %
(very) quiet	30 %	14.1 %

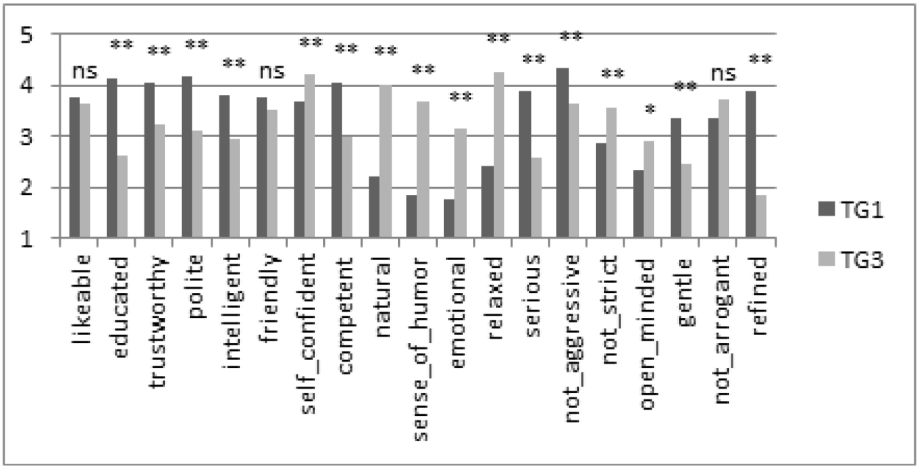


Fig. 1. Pairwise comparison of Austrian standard (TG1) and Viennese dialect (TG3) employing Wilcoxon test and Bonferroni correction; ** α=0.01, * α=0.05, ns not significant.

2.2 Expression of Extroversion-Introversion in Gestural Correlates

For designing the agents' body behaviours, the following gestural correlates for extroversion/introversion were exploited, see. Indicators for gesture rate, amplitude, gesture direction and body part were taken from (Neff et al. 2010). Similar findings are reported in studies by (Knapp and Hall 2009 and Lippa 1998) stating that extroversion correlates with a higher spatial extent when gesturing. The gestural speed also tends to be higher for extroverts (Lippa 1998 and Brebner 1985) making their gestures look more powerful. As regards self-adaptor gestures, we follow a study on virtual agents from (Neff et al. 2011) where self-adaptors were identified as signalling low emotional stability, as well as earlier findings from (Campbell and Rushton 1978) indicating negative association of self-adaptors and outward-directed gestures, i.e. gestures that signal extroversion.

Table 2. Gestural correlates for extroversion-introversion

	Introversion	Extroversion
Gesture amplitude	Narrow gestures	Wide gestures
Gesture speed	Low	High
Gesture direction	Inward self-contact gestures	Outward gestures
Gesture rate	Low	High
Body part	Elbows/arms close to the body	Elbows away from body
Self-adaptor gestures	Yes	No

3 Building Extrovert-Introvert Agents

For the technical realization, we use the *Virtual Beergarden* running in the AAA application (Damian et al. 2011). In this scenario, an arbitrary number of agents can be loaded that are able to speak and to exhibit animations. For our aim, we employ the *Charlie* character, a male middle-aged virtual agent with a western appearance. The character initially plays an idle animation that includes eye blinking. Other animations are added according to the current body behaviour script.

Verbal behaviour is realized by a text-to-speech component, in which different voices can be used for the characters. The generated speech consists of audio with synchronised lip movements of the characters. To avoid an influence of the semantics of speech on the perceived extroversion of the character, we choose content and wording as linguistically neutral as possible with respect to extroversion/introversion. For our setting, the following text is uttered:

Willkommen im Biergarten. Wir bieten Sitzplätze für hundert Personen und warme oder kalte Speisen. Von allen Plätzen haben Sie einen guten Ausblick auf die Landschaft. Für Firmenfeiern oder private Feste bieten wir Spezialkonditionen an. Und wenn sie kurzfristig mit mehr Personen kommen wollen, macht das auch nichts.

(En gloss: Welcome to the Beergarten. We offer room for 100 persons and cold and hot dishes. From all places, you get a good view on the landscape. For official or private parties we offer special conditions. And in case you wish to bring additional people on short notice, this does not matter.)

3.1 Realization of the Language Varieties

The virtual character is equipped with three different language varieties: standard German, standard Austrian and Viennese dialect. Following (Gärtig et al. 2010 and Krenn et al. 2012), we hypothesize these language variants as representative for different grades of extroversion-introversion, with Viennese dialect being the most extrovert and standard German being the most introvert.

The agents' speech was generated with two different text to speech engines which are:

- The CereVoice SDK (https://www.cereproc.com/en/products/sdk) for generating the standard German and standard Austrian utterances. For German, the male voice Alex was used, and for Austrian the male voice Leopold (https://www.cereproc.com/de/storede).
- The Festival Multisyn TTS engine (http://www.cstr.ed.ac.uk/projects/festival/, Clark et al. 2007) for generating the utterances in Viennese dialect. See http://vsds.ofai.at.vsds_synthesize.cgi, HPO: Festival Unitsel 16 kHz for the respective voice.

3.2 Realization of the Body Behaviour

In the current version of the *Virtual Beergarden* over 70 animations are available for each character. Non-verbal behaviours are divided into gestures, body postures and facial expressions. Predefined gestures and body postures, per se, vary in their expressivity, e.g. an adaptor gesture (such as scratching the nose) has a lower spatial extent compared to a waving movement as used for greeting. In addition to this gesture-inherent expressivity, animations can be further customized to show different levels of expressivity, e.g. an animation can be played with a higher frame rate to increase the speed of the gesture. In previous work, Damian et al. (2011) conducted evaluation studies to test, amongst others, the effects of variations in non-verbal behaviour on the perceived personality (introvert vs. extrovert) of virtual characters in a conversational setting. The same animation technology is employed as in the *Beergarden*. We take this as further evidence for the validity of the animations in the present study to adequately transport extrovert, introvert and neutral behaviour.

Three different non-verbal variations were created to match the text spoken by the agent: introvert, neutral and extrovert. The introvert version is characterised by a lower animation rate compared to the neutral and extrovert version. It includes adaptor gestures and has low spatial extent in gestures and body postures. The neutral version contains more animations than the introvert version, but fewer gestures than the extrovert version. Animations are at a middle level of spatial extent and speed. The extrovert version shows the highest gesture rate, spatial extent, and speed. No adaptor gestures are used in the neutral and extrovert settings. Fig. 2 shows screenshots of the agent exhibiting animations in the extroversion and introversion setting.

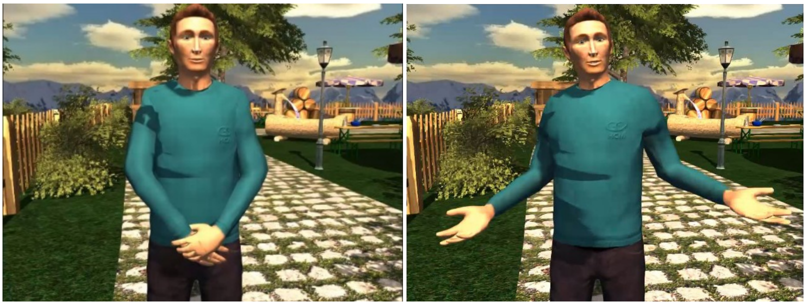


Fig. 2. The Charlie character exhibiting animations in the extrovert (left) and introvert (right) versions

4 The Study

4.1 Experimental Design

We follow Neff et al. (2010) in the experimental design of the study and in the design of extrovert-introvert body behaviour, and extend it by employing different synthetic voices which stand for different language varieties, namely standard German, standard Austrian German and Viennese dialect. The main goal of the present study is to assess how far the perceived extroversion-introversion of a language variety influences the human perception of an agent's extrovert or introvert personality.

Parameter Combinations. We follow a 3x3 design where the synthetic voices/language varieties and bodily behaviours are varied while the agent's text, graphical appearance and setting of the scene are kept constant. This results in 9 settings represented by 9 videos as summarized in Table 3. For the sake of simplicity, from now on we will refer to the number of each video instead of the respective parameters combinations.

Table 3. Combination of parameters as used in the study

Variation in speech/ Non-verbal behaviour	Standard German	Standard Austrian	Viennese Dialect
Introvert	Video 1	Video 4	Video 7
Neutral	Video 2	Video 5	Video 8
Extrovert	Video 3	Video 6	Video 9

Questionnaire. The assertions presented in are adapted from the Big-Five questionnaire published in (Satow 2012), in order to assess the agents' extroversion-introversion dimension as perceived by humans. Whereas the original assertions are made from the human agent's perspective, we have adapted the assertions to be uttered from the observer's perspective in the form of ***I believe that the agent*** **in the video ASSERTION.** Each assertion is rated on a 7-point Likert scale ranging from 'not at all' (*Trifft überhaupt nicht zu*) to 'fully agree' (*Trifft voll und ganz zu*). See Table 4 for the respective German wording.

In addition, the participants were asked for their age, gender, mother tongue, nationality, where they have mainly lived for the last 5 to 10 years, and whether they had experience with virtual agents and speech synthesis before.

Hypotheses. The videos shown to the participants contain variations in non-verbal behaviour and language variety. The hypotheses are theoretically grounded in earlier findings: (i) East Upper German is perceived by German speakers as extrovert as opposed to Standard German which is perceived as introvert (Gärtig et al. 2010); (ii) Viennese dialect is evaluated as more extrovert as Austrian standard (Krenn et al. 2012); and (iii) gestural expressivity increases the perception of extroversion (Neff et al. 2010).

The hypotheses tested in the present study are:

- **H1:** Language variety increases the perceived extroversion of an agent.
- **H2:** Increased extroversion in non-verbal behaviour leads to an increased perceived extroversion of the agent.
- **H3:** A combination of parameters that are perceived as more extrovert leads to even higher perceived extroversion of the agent.
- **H4:** Viennese dialect is perceived as more extrovert than Austrian and German standard.
- **H5:** German standard is perceived as least extrovert in comparison to Austrian standard and Viennese dialect.

Table 4. Assertions to assess the perceived extroversion-introversion of an agent.

Ich glaube, dass der Agent im Video	gerne mit anderen Menschen zusammen ist. (‘likes to be in company’)
	ein Einzelgänger ist. (‚is a loner‘)
	in vielen Vereinen aktiv ist. (‚is active in many clubs‚)
	ein gesprächiger und kommunikativer Mensch ist. (‘is a talkative and communicative person’)
	im Grunde oft lieber für sich allein ist. (‚prefers to be on his own‘)
	sehr kontaktfreudig ist. (‚is very sociable‚)
	schnell gute Stimmung verbreiten kann. (‚can quickly spread a good mood‘)
	gerne auf Partys geht. (‚likes to go to parties‘)
	unternehmungslustig ist. (is adventurous)
	gerne im Mittelpunkt steht. (‚likes to be in the center of attention‘)

Table 5. Likert scale rating for the extroversion-introversion assertions as presented to the participants

Trifft überhaupt nicht zu	Trifft größtenteils nicht zu	Trifft eher nicht zu	Weder zutreffend noch unzutreffend	Trifft eher zu	Trifft größtenteils zu	Trifft voll und ganz zu
Applies not at all	Applies not for the most parts	Applies rather not	Neither applies or not applies	Rather applies	Applies for the most parts	Fully applies
1	2	3	4	5	6	7

4.2 Participants

The videos and questionnaires were embedded in a website. Thus, participants were not distracted by a lab setting, but able to watch the videos where and when they liked. In total, 45 people (22 female, 23 male) participated in our study. 22 of the participants have German nationality and 23 have Austrian nationality. Participants were in an age range from 23 to 54 years (Ø 32.2). On a 4-point scale ranging from no experience (1) to very much experience (4), participants reported their previous experience with virtual characters (Ø 1.69, Median 1) and speech synthesis (Ø 1.62, Median 1). All participants are of German mother tongue, except for one whose mother tongue is English and another one who is bilingual (German/Bosnian). Both participants are of Austrian nationality and have lived in Austria for the last 5-10 years, and therefore were included in the analysis.

5 Analysis and Interpretation

A two-way repeated measures ANOVA (using SPSS) with the factors language variety and body behaviour was computed on the participants' evaluation of the perceived extroversion of the agent in videos 1 to 9. As a post-hoc test, Bonferroni corrected t-tests were applied.

The ANOVA showed highly significant effects ($\alpha < 0.01$) of language variety (F(2, 88) =30.81) as well as of body behaviour (F(1.42, 62.34)=13.57 with Greenhouse-Geisser corrections) on the perceived extroversion of the agent. The results confirm H1 and H2. However, there are no significances for joint effects of language variety and bodily behaviour. Accordingly, H3 has to be rejected.

As regards language variety, the pairwise comparisons showed differences ($\alpha <$ 0.01) between standard German on the one hand, and standard Austrian and Viennese dialect on the other hand, whereby both standard Austrian and Viennese dialect were evaluated as more extrovert than standard German. Thus H5 is confirmed. H4 must be rejected, as there are no differences between standard Austrian and Viennese.

Although the ANOVA suggested no significances for the interaction between language variety and body behaviour, a closer look at the within-subjects contrasts revealed a significantly higher difference (α < 0.05) in the perceived extroversion of the agent between the Austrian standard combined with the extrovert versus the introvert body, than it is the case for the German standard when combined with the extrovert versus the introvert body (F(1, 44)=7.06). For illustration see Fig. 3, the distances in the "rated degree of extroversion" between standard German with introvert and extrovert body, and between standard Austrian with introvert and extrovert body.

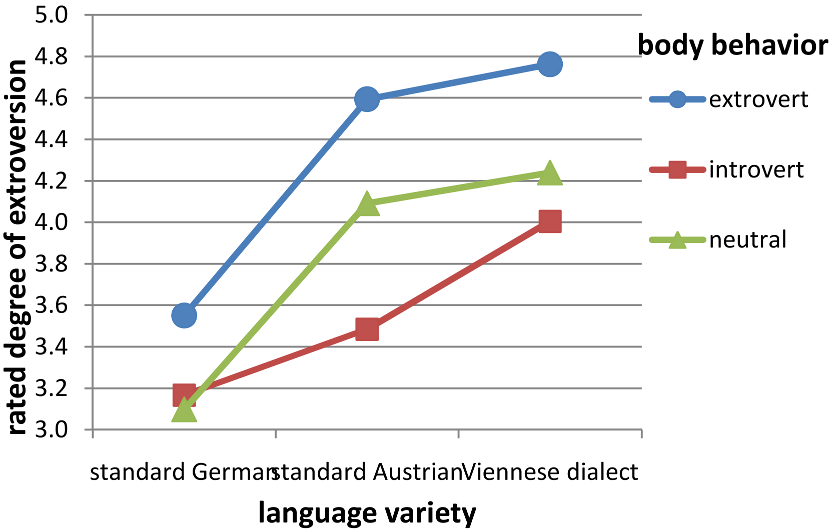


Fig. 3. Rated degree of the agent's extroversion based on the factors language variety and body behaviour

6 Conclusion

An online-study was conducted with speakers of German and Austrian standard language variety who were to assess a virtual agent's degree of extroversion based on the agent's synthesized language variety and its body language. The agent was equipped with synthetic speech representing German and Austrian standard language variety and Viennese dialect, and with gestural behaviour suggesting an extrovert or introvert personality. Clear evidence was found that: (i) Synthesized language variety increases the perceived extroversion of an agent, with Viennese dialect and Austrian standard being perceived as more extrovert than German standard. This confirms findings from language attitude studies on the effects of natural language variety on the perception of a speaker's personality. Thus the present study provides further evidence for the fact that effects of natural speech transfer to synthesized speech. (ii) Increased extroversion in non-verbal behaviour leads to an increased degree in perceived extroversion of the agent. Whereby aspects such as gesture rate, amplitude and speed, or the presence or absence of self-adaptor gestures influence whether a virtual agent is perceived as extrovert or introvert. Here again, the results from the present study corroborate findings from earlier works on the effects of non-verbal behaviour on perceived personality. In addition, the data also provide some evidence for joint effects of language variety and non-verbal behaviour, namely the Austrian standard language variety combined with extrovert body behaviour leads to a strong increase of perceived extroversion compared to the effect of Austrian standard combined with introvert body language, whereas the respective effect is significantly smaller when the German standard language variety is combined with extrovert or introvert non-verbal behaviour.

Overall, the results clearly demonstrate the effects language variety, on the one hand, and body language, on the other hand, have on the human perception of an agent's extrovert or introvert personality. Thus the presented work offers respective guidance for the design of artificial agents. Moreover, the current work further supports previous findings (cf. Krenn et al. 2012) that evidence from effects of natural language varieties transfers to synthetic speech. This is relevant for agent designers as their design process may directly profit from existing results of language attitude studies on natural speech. Which may be a factor, given the growing availability of synthetic voices representing language varieties, see for example Cereproc's commercially available synthetic voices for Scottish, Northern and Southern British, Irish, Catalan, or Austrian German.

In future work, we aim to take a closer look at potential differences in the assessment of the agent's personality by the Austrian as opposed to the German user group, and include the gender dimension analysing and comparing the answers from male and female participants separately. In the present study, the assertions referring to the agent's extroversion-introversion in the questionnaire are closely related to human contexts, as they are taken from a typical personality questionnaire which is designed for human self-assessment. See for instance the wording in, in particular the explicit use of the word *Mensch*. This may bias the human ratings towards "neither-nor" answers, especially in combination with the 7-point rating scale applied in the questionnaire. Thus, an analysis of the participants' rating behaviour per assertion is called for in the first place. Moreover, the assertions might need better adaption to an artificial agent context where, for instance, *Mensch* is exchanged by *Charakter* (En.: character). In addition, the rating scale may be changed to 6 or even to 4 in order to (i) prevent the participants from giving "neither-nor" answers as it is possible with the currently used 7-point scale, and to (ii) force answers towards the extroversion or introversion end of the dimension.

References

1. Brebner, J.: Personality theory and movement. Individual differences in movement, pp. 27–41. MTP Press Limited (1985)
2. Clark, R., Richmond, K., King, S.: Multisyn voices from ARCTIC data for the Blizzard challenge. In: Proceedings of Interspeech, pp. 101–104 (2007)
3. Campbell, A., Rushton, J.: Bodily communication and personality. The British Journal of Social and Clinical Psychology 17(1), 31–36 (1978)

4. Damian, I., Endrass, B., Huber, P., Bee, N., André, E.: Individualized Agent Interactions. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 15–26. Springer, Heidelberg (2011)
5. Gärtig, A.-K., Plewnia, A., Rothe, A.: Wie Menschen in Deutschland über Sprache denken. Ergebnisse einer bundesweiten Repräsentativerhebung zu aktuellen Spracheinstellungen. amades - Arbeitspapiere und Materialien zur deutschen Sprache. Band 40. Institut für Deutsche Sprache, Mannheim (2010)
6. Gosling, S.D., Rentfrow, P.J., Swann Jr., W.B.: A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality 37, 504–528 (2003)
7. Knapp, M.L., Hall, J.A.: Nonverbal communication in human interaction. Holt, Rinehart (2009)
8. Krenn, B., Schreitter, S., Neubarth, F., Sieber, G.: Social Evaluation of Artificial Agents by Language Varieties. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 377–389. Springer, Heidelberg (2012)
9. Lippa, R.: The nonverbal display and judgement of extraversion, masculinity, femininity, and gender diagnosticity: A lens model analysis. Journal of Research in Personality (32), 80–107 (1998)
10. Neff, M., Toothman, N., Bowmani, R., Fox Tree, J.E., Walker, M.A.: Don't scratch! self-adaptors reflect emotional stability. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 398–411. Springer, Heidelberg (2011)
11. Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the Effect of Gesture and Language on Personality Perception in Conversational Agents. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS (LNAI), vol. 6356, pp. 222–235. Springer, Heidelberg (2010)
12. Neubarth, F., Pucher, M., Kranzler, C.: Modeling Austrian dialect varieties for TTS. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia, pp. 1877–1880 (2008)
13. Pucher, M., Neubarth, F., Strom, V., Moosmüller, S., Hofer, G., Kranzler, C., Schuchmann, G., Schabus, D.: Resources for speech synthesis of Viennese varieties. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, pp. 105–108 (2010)
14. Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V.: Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. Speech Communication 52(2), 164–179 (2010)
15. Satow, L.: Big-Five-Persönlichkeitstest (B5T): Test- und Skalendokumentation (2012), `http://www.drsatow.de`

Long Text Reading in a Car

Ladislav Kunc[1], Martin Labsky[1], Tomas Macek[1], Jan Vystrcil[1], Jan Kleindienst[1],
Tereza Kasparova[1], David Luksch[1], and Zeljko Medenica[2]

[1] IBM Prague Research and Development Lab, Prague, Czech Republic
{ladislav_kunc1,martin.labsky,tomas_macek,jan_vystrcil,
jankle,tereza.kasparova,david.luksch}@cz.ibm.com
[2] Nuance Communications, Inc., Burlington, United States
Zeljko.medenica@gmail.com

Abstract. We present here the results of a study focused on text reading in a car. The purpose of this work is to explore how machine synthesized reading is perceived by users. Are the users willing to tolerate deficiencies of machine synthesized speech and trade it off for more current content? What is the impact of listening to it on driver's distraction? How do the answers to the questions above differ for various types of text content? Those are the questions we try to answer in the presented study. We conducted the study with 12 participants, each facing three types of tasks. The tasks differed in the length and structure of the presented text. Reading out a fable represented an unstructured pleasure reading text. The news represented more structured short texts. Browsing a car manual was an example of working with structured text where the user looks for particular information without much focusing on surrounding content. The results indicate relatively good user acceptance for the presented tasks. Distraction of the driver was related to the amount of interaction with the system. Users opted for controlling the system by buttons on the steering wheel and made little use of the system's display.

Keywords: Architectures for interaction, CUI, SUI ad GUI, HCI methods and theories, Interaction design, Speech and natural language interfaces, Long text reading, car, UI, LCT.

1 Introduction

Drivers are well accustomed to listening to radio, music or audio books. The quality of machine synthesized speech is however still inferior to performance of a professional speaker reading out a text tailored for audio presentation. However, it is much slower, less flexible and more expensive to create such content.

The purpose of the study presented in this text is to learn to what extent the user is willing to cope with the deficiencies of text to speech synthesis (TTS).

Text processing is one of the activities humans do frequently. It ranges from passive reading to text creation, error correction and team collaboration. Users tend to shift most of their activities conducted previously on desktop to mobile environment. They even want to perform certain tasks in a car while driving. User interfaces for

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 440–449, 2014.

mobile devices however have to respect a smaller form factor, less efficient input methods and distraction caused by using the system in a car. We addressed the tasks of text creation and correction in our previous work [2]. In this paper we focus on an apparently less difficult but important task of text reading.

2 Related Work

Significant attention was devoted in the past to assessing the impact of various in-car activities [1]. The Lane Change Test (LCT) [9] and subjective tests using questionnaires such as NASA TLX [5] and DALI [6], [7] are examples of popular methods used to assess the impact of various secondary in-car tasks on the primary task of driving.

Although electronic 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 [4]. The negative impact on driving performance due to having conversation with someone while driving was assessed by various studies [15], [16].

Several approaches to designing speech-based UIs for in-car usage including menu-based and search-based UIs were described [8], [10].

General quality of various TTS systems can be effectively measured only on the basis of reliable and valid listening tests, e.g. using mean opinion scale [18]. TTS quality was also assessed in terms of its suitability for various tasks such as computer assisted learning of foreign languages [17]. In this study we try to show that the quality of today's state-of-the-art TTS systems is sufficient for reading out texts in a car.

3 Research Goals and Experiment Design

The purpose of this study is to analyze the usability and distraction aspects of text reading in a car in general. The research questions that we search answers for are of three categories: usability, distraction and performance.

- **Usability:** Is the TTS quality sufficient for this kind of task? What part of the implemented functionality is actually used by the user? What are the preferred control mechanisms (buttons vs. swipe gestures, audio vs. visual feedback)? What are the preferred usage patterns (auto-playback vs. manual browsing through the text)? Is there correlation between the results and personal information about the subjects?
- **Distraction:** What levels of distraction can we observe for each of the tasks? How is distraction perceived subjectively? How often and for how long do the users look at the screen?
- **Performance:** Does the user remember what has been read?

We decided to carry out tests using three scenarios: 'Fable', 'News' and 'Car Manual'. They differ in the complexity of information, in the structure of the presented text and in the ways the user is allowed to interact with the text being read.

- **The fable scenario** represents a task of reading a plain unstructured text such as a short book chapter or an article. The user is only able to navigate within the text and may navigate by sentences and paragraphs.
- **The news scenario** involves reading multiple shorter texts (news articles). It demands more interactivity. The user can navigate between the articles or within the text of an article.
- **The car manual scenario** represents a complex task of looking for specific information in a car owner's user manual. It requires formulation of a query by the user, navigation in multiple search results and finally navigation in the retrieved user manual section to find the relevant piece of information. The user manual text was presented without modifications as extracted from a standard PDF car owner's manual.

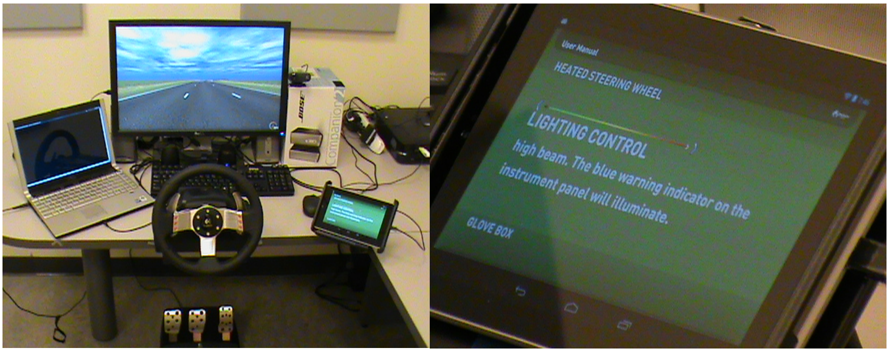
Testing procedure consisted of the following steps. Initially, the whole procedure was explained to the participants. All training and evaluated drives were conducted at a constant speed of 60km/h on a standard straight 3-lane road in a Lane Change Test Simulator [9]. All drives were approximately 3.5 km long and took 3.5 minutes. First, our subjects trained the primary task of driving during a single drive and filled in a pre-test questionnaire. Prior to driving with secondary tasks, participants conducted one undistracted ride which was used to estimate an ideal LCT track adapted to each participant's style of driving. Another undistracted ride was conducted at the end of the testing session and was used as a reference to compare against distracted rides.

Training for each reading task was done shortly before evaluating it. The order of tasks was counterbalanced to compensate for a possible learning effect. Three distracted rides were conducted and each was followed by filling in the DALI [6] and SUS [13] questionnaires. In addition, for the car manual task, participants first searched by voice for a pre-specified topic, such as "turning fog lamps on", and only then they navigated through the retrieved set of articles to locate the relevant piece of information. For this task, participants also filled in an additional SASSI [14] form at the end of the drive.

Tests were conducted in a laboratory environment. The drivers were using a low-fidelity driving simulator to mimic driving on a highway. The primary task was performed using the standard LCT [9] used according to ISO 26022:2010 [12]. Fig.1 depicts the physical location of the devices during the experiment.

As a test bed, we used a prototype of an in-car infotainment system with a dedicated component for text reading (right part of Fig.1). We used the Nuance Vocalizer TTS system with a Premium US English voice named Ava.

The tested system presented text primarily through the audio channel via TTS playback. It allowed both for passive listening and for active navigation in the text using steering wheel buttons and touch screen swipe gestures. Participants could make use of up to 6 steering wheel buttons in a layout depicted in Fig.2, which allowed for advanced navigation in the presented text, including navigation between articles and within an article at the level of individual sentences and paragraphs.



User Manual

HEATED STEERING WHEEL

LIGHTING CONTROL

high beam. The blue warning indicator on the instrument panel will illuminate.

GLOVE BOX

Fig. 1. Testing setup (left) and sample text shown on the system's display (right)

Control using swipe gestures was limited to navigation between articles only, using vertical swipe gestures. The double tap gesture activated speech recognition with automatic end-of-speech detection. Horizontal swipe gestures were reserved for swiping between different applications and thus were not used in this test.

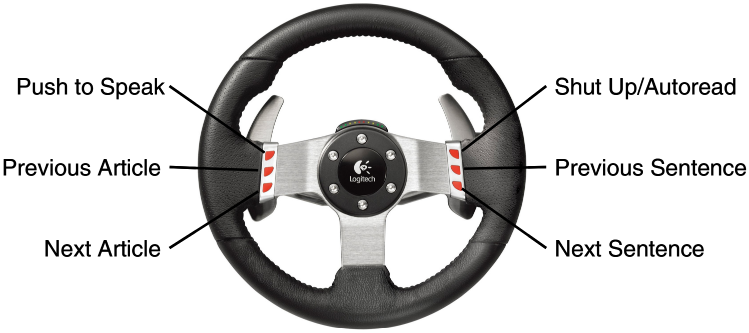


Fig. 2. Steering wheel buttons layout

The visual presentation of the text on a display was intended as complementary information only. The display showed three lines of text in large fonts. The word currently being read was underlined. A progress indicator above the text showed the current reading position within a text block (e.g. news article).
Speech recognition (Nuance Vocon Hybrid) was used to search for relevant Car Manual articles.

4 Testing Results

The study was piloted with one subject and then conducted with 14 subjects in Burlington, USA. All participants were US English native speakers. Two subjects were excluded due to an error in recording of the data. Half of the test subjects were females; all of them were driver's license holders, age varied from 20 to 55 with 7 participants under 29 years. 11 subjects drove and used radio daily; all had at least high school education.

We collected both usability feedback and objective distraction statistics, and also evaluated performance of test subjects on the reading task using simple reading comprehension tests.

Distraction. We measured driving distraction both objectively [9] and subjectively [6]. Fig.3 depicts objective measurements using LCT driving logs. We report SDLP (Standard Deviation of Lateral Position) and MDev (Mean Deviation) calculated both for the whole evaluated drive and for lane keeping segments only (excluding lane change segments).

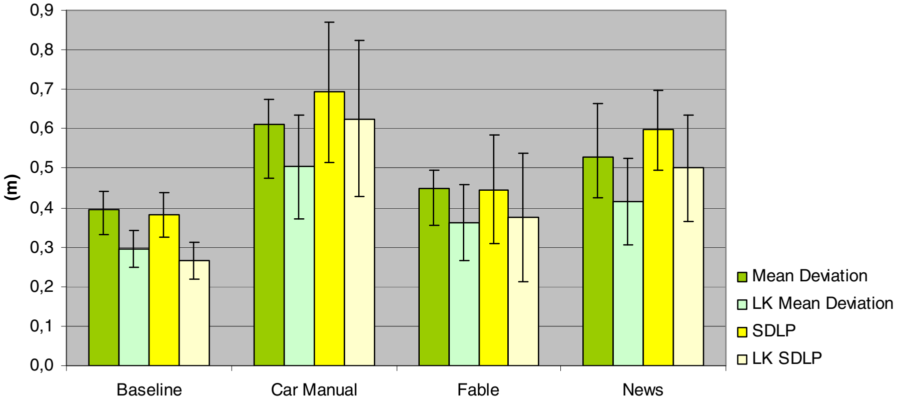


Fig. 3. Distraction measured objectively using mean deviation and SDLP; LK denotes lane-keeping versions of the statistics. 95% confidence intervals are shown.

There are statistically significant differences in the distraction for the news and car manual tests when compared to the undistracted ride. The fable was found to be the least distracting task with impact on driving that did not reach statistical significance with α=0.05.

Fig.4 depicts distraction measured subjectively using the DALI [6] test. The distraction ranking of tasks for all of the observed domains is the same as for the objective statistics. Fig.5 shows numbers of glances that each user made at the application screen. The counts vary. Some participants did not look at the screen at all, while others used the screen more frequently. Overall, the observed distraction results confirm the hypothesis that the more interactive tasks (car manual and news) cause more distraction.

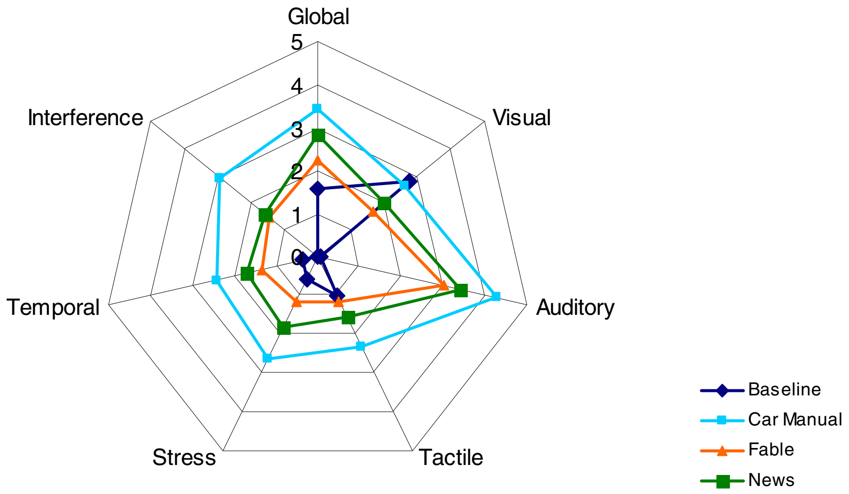


Fig. 4. Subjective distraction using DALI

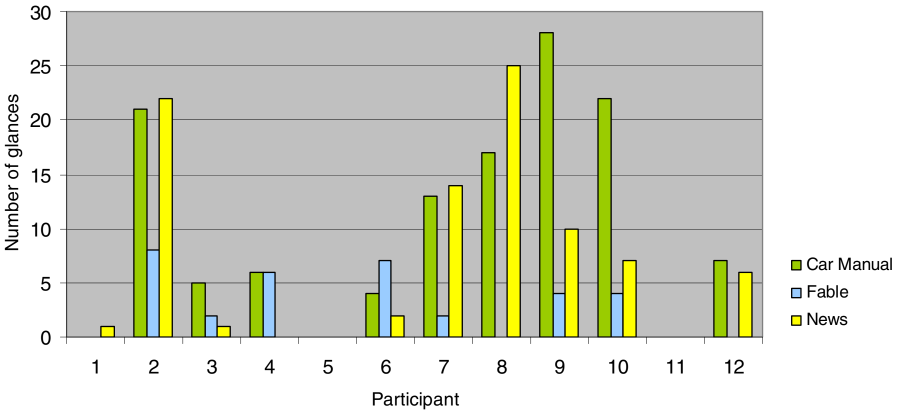


Fig. 5. Number of glances at the screen

The application screen was subjectively perceived as distracting. Most of the participants preferred to use the system without a screen or would move the screen to a position closer to the windshield. This finding is similar to the results presented in [11]. Most of the glances at the application screen occurred during the car manual task as the participants often checked the results of their voice search commands. The number of glances tended to be higher in the case when the retrieved content included irrelevant search results.

Performance. We evaluated efficiency of the system by asking participants several questions regarding the presented content at the end of each task, to verify that the content was understood and remembered.

The **car manual** test consisted of three tasks, each assigned immediately after the previous one was completed. Each task was rated successful or unsuccessful based on whether the user was able to find the requested information. Success rate was calculated as the number of successful tasks normalized by the total number of tasks (3).

The **fable** test was followed by asking three questions to the participant. The success rate was calculated as the number of right answers normalized by the number of questions (3).

The **news** test was evaluated as follows. For each article, three important facts were chosen that were expected to be remembered by the participants. The subjects were asked to repeat what they remembered and the experimenter could ask complementary questions. The success rate was the number of facts correctly remembered normalized by the overall number of facts (9).

The results in Fig.6 indicate that the tasks were reasonably complex. It may however still be problematic to compare the difficulty of the tasks using the achieved average success scores as they depend on the complexity of questions that were constructed subjectively. Overall, the mean values of success rate for the fable task were the highest (mean 89, deviation 0.16) followed by the news task (mean 75, deviation 0.18) and the car manual task (mean 67, deviation 0.24).

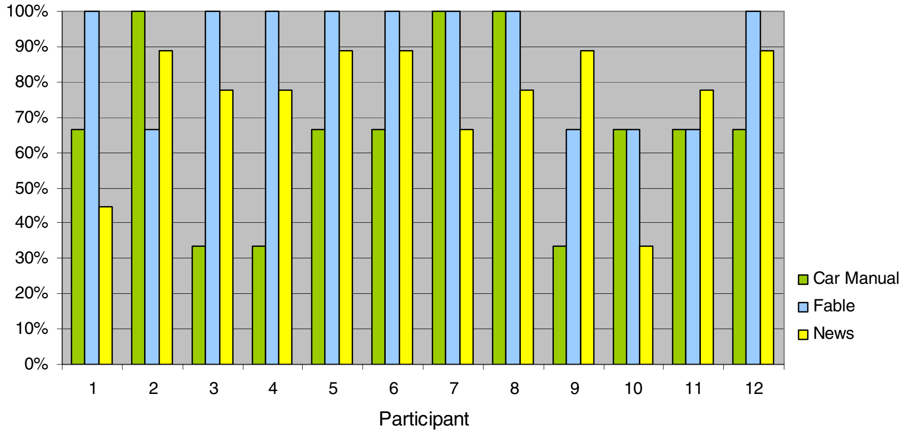


Fig. 6. Success rates for all tasks measured for all participants

Usability. The important part of the study was to collect usability feedback from participants; both about the text reading task in general and also concerning the utilized prototype. We collected feedback by analyzing video recordings, by interviewing the participants and by asking them to fill in several questions that were specific to each task.

The Car Manual task was also evaluated by collecting the SASI factors [14] as it was the only task that included search functionality that could be evaluated for accuracy. The scaled SASI factors are shown in Fig.7, indicating that part of the users considered it difficult to find a specific piece of information in a list of retrieved user manual sections.

In general, the subjects found the system useful. It was clear how to use it and easy to learn. Younger and more educated users liked the system more, and they performed better. The participants who were used to process information audibly (preferred radio to TV or newspaper) also performed better than others. We observed the reading process and analyzed how the users handled the related tasks of browsing and searching for specific content. We wanted to understand the degree to which the users exploit some of the advanced features of the prototype such as multiple browsing granularities or the way of displaying text on the screen.

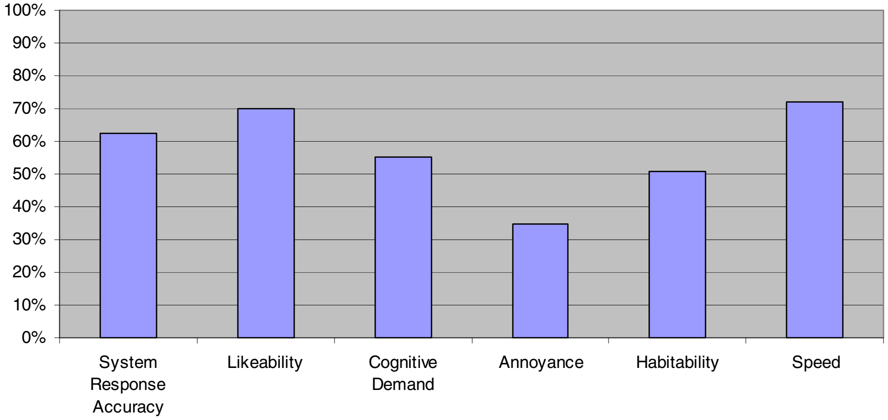


Fig. 7. Scaled SASI factors for the car manual task

Although some participants complained about the quality of the synthesized voice, they declared that they would like to use the system for reading of a wide spectrum of texts (news, emails, books and instant messages). They preferred not to use it for browsing car manual content in its original form. The reasons for that included distraction caused by navigating technical text, the limited suitability of the original user manual for presentation by voice and limited need to perform the task while driving. The users opted for controlling the UI by buttons on the steering wheel instead of using swipe gestures on a touch screen.

5 Conclusion

We presented here the results of a long text reading study performed on three types of texts. Although the number of tested subjects is relatively small to make a major quantitative evaluation, the study provides useful qualitative observations and distraction estimates. A longitudinal study should be carried out to observe adoption of long text reading components in a daily driving scenario. We paid special attention to the impact of the content type presented and to the acceptance of TTS for each of the tasks. The study showed that most of the participants would use the system for reading various kinds of texts in spite of some complaints about the quality of TTS. The results suggest that car manual content in its original form may be appropriate for browsing and searching while parked, but other forms of presentation such as question answering should be considered for use while driving. The answers should be specially tailored for in-car presentation by voice. The application GUI was perceived as distracting. However some participants still used the GUI during the car manual tasks, mainly to verify the correctness of the retrieved search results.

Acknowledgment. This study was done jointly by Nuance and IBM as part of their joint research and development agreement.

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. 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)
4. 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)
5. 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. North Holland Press, Amsterdam (1988)
6. Pauzié, A.: A method to assess the driver mental workload: The driving activity load index (DALI). IET Intelligent Transport Systems 2(4), 315–322 (2008)
7. Pauzie, A.: Evaluation of Driver Mental Workload Facing New In-vehicle Information and Communication technology. 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)
9. Stefan, M.: The lane-change-task as a tool for driver distraction evaluation. In: Proceedings of the Annual Spring Conference of the GFA/ISOES, vol. 2003 (2003)
10. Labsky, M., Kunc, L., Macek, T., Kleindienst, J., Vystrcil, J.: Recipes for building voice search UIs for automotive. Submitted to EACL 2014 - Dialogue in Motion Workshop, Sweden (2014)

11. Vystrcil, J., Macek, T., Luksch, D., Labsky, M., Kunc, L., Kleindienst, J., Kasparova, T.: Mostly Passive Information Delivery - A Prototype. Submitted to EACL 2014 - Dialogue in Motion Workshop, Sweden (2014)
12. Road vehicles-Ergonomic aspects of transport information and control systems-Simulated lane change test to assess invehicle secondary task demand, International Standard ISO/DIS 26022:2010
13. Brooke, J.: SUS-A quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor and Francis, London (1996)
14. Hone, K.S., Graham, R.: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). Natural Language Engineering 6(3-4), 287–303 (2000)
15. Kubose, T.T., Bock, K., Dell, G.S., Garney, S.M., Kramer, A.F., Mayhugh, J.: The effects of speech production and speech comprehension on simulated driving performance. Applied Cognitive Psychology 20(1), 43–63 (2006)
16. Drews, F.A., Pasupathi, M., Strayer, D.L.: Passenger and cell phone conversations in simulated driving. Journal of Experimental Psychology: Applied 14(4), 392 (2008)
17. Handley, Z.: Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Communication 51(10), 906–919 (2009)
18. Viswanathan, M., Viswanathan, M.: Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. Computer Speech & Language Iss. 19(1), 55–83 (2005)

Let's Get Personal

Assessing the Impact of Personal Information in Human-Agent Conversations

Nikita Mattar and Ipke Wachsmuth

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

Abstract. Agents that are able to build relationships with the people they are interacting with are envisioned to be more successful in long-term interactions. Small talk about impersonal topics has been found an adequate tool in human-agent interactions for manipulation of such relationships. We suspect that an agent and the interaction with it will be evaluated even more positively when the agent talks about personal information it remembers about its interlocutor from previous encounters. In this paper a model of person memory that provides virtual agents with information needed in social conversations is presented. An interaction study demonstrates the impact of personal information in human-agent conversations and validates the performance of our model.

Keywords: conversational agents, intelligent virtual agents, human-agent interaction, person memory, social conversations, interaction study.

1 Introduction

In conversational and intelligent virtual agent research an important goal is to create agents that are able to build relationships with the people they are interacting with. This goal is motivated by the fact that virtual agents develop from tools to human-like partners [20].

Small talk has been found an adequate tool in human-agent interaction to, e.g., increase trust, which is an important prerequisite for close relationships. Considering theories on politeness strategies and *face work*, initial approaches focused on impersonal topics, like the weather, when engaging the agent in small talk with its interlocutor [3]. According to these theories, talk going beyond safe impersonal topics would seem inappropriate in initial encounters and therefore could threaten the development of a closer relationship.

For a relationship to develop from superficial acquaintance towards a level of closer friendship, personal matters are important. Bringing personal topics to the table is a sign of high involvement and signals willingness to deepen a relationship. One common approach to introduce more personal information in repeated human-agent conversations is to let the agent use a strategy of self-disclosure [12]. However, this information is centered around the agent.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 450–461, 2014.

To enable conversational agents to exhibit more appropriate behavior in social encounters, we [14] proposed to equip such agents with a person memory and demonstrated how this memory can be populated with personal information during initial encounters [16]. In more recent work we showed how the personal information can be exploited by our agent during conversation [17], [18].

Based on findings on *relational work* [21], we expect that an agent and the interaction with it will be evaluated more positively by human interlocutors, with regards to, e.g., likability and communication satisfaction, when the agent talks about personal information it remembers about its interlocutor from previous encounters.

In this paper we present our model of person memory and the results of a first interaction study we conducted to test our hypotheses. In Section 2, an overview of related work is given. We describe the key ingredients used in our model of person memory in Section 3. In Section 4, the interaction study and its results are presented.

2 Related Work

Various approaches have been proposed where different levels of behavior of virtual agents are adapted to achieve increased believability. It has been examined how to adapt the agent's display of emotions [2], and gesturing [10]. Also, influences of personality [13], and the interlocutors' cultures [7], on conversational behavior have been investigated. These approaches have in common that they focus on processes that affect the interpersonal relationship between the agent and a person the agent interacts with. As conversational agents start to appear in everyday interaction scenarios, the question arises how to provide agents with information fundamental for being able to handle repeated social encounters.

Most approaches dealing with virtual agents that are to operate in long-term scenarios rely on human-like memory systems, i.e. autobiographic and episodic memory [9], [6]: Autobiographic memories can be used to increase the performance of storytelling or narrative agents [9]. Episodic memories can be used in domains where learning and reasoning about actions is a crucial task [19], [6]. However, both these kinds of memories have in common that they are egocentric systems with the experiences of the agent in focus.

We question if an egocentric memory component is sufficient to handle the requirements that come up in social encounters. So, what would be the requirements of a person memory for a conversational agent?

3 Ingredients of a Person Memory for a Conversational Agent

We identified the following ingredients as crucial for a person memory for conversational agents (see. Fig. 1):

1. Representations of **persons** and **social categories**

2. A representation of the **interaction context** that integrates knowledge of the **social situation** and representations of the individuals
3. **Social memory tasks** and **social strategies** that function as operating rules and instructions on how to deal with the provided information
4. A **Person Memory Processing Unit** that provides an interface between the person memory and the agents cognitive architecture

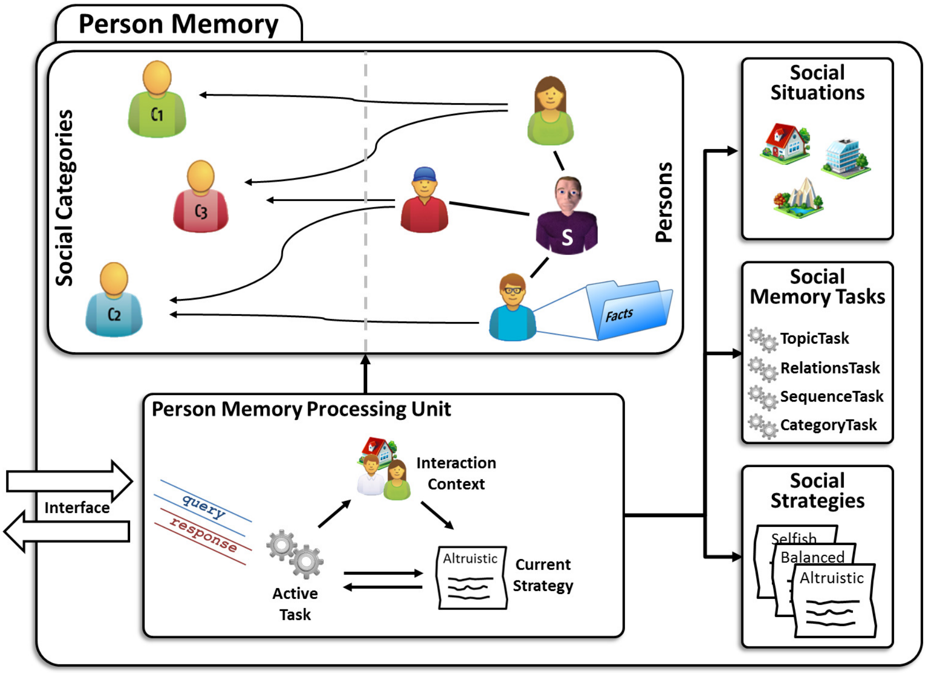
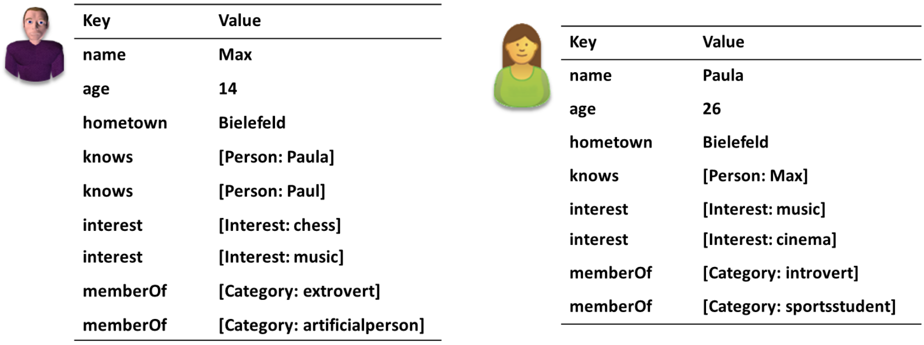


Fig. 1. Model of the person memory. Besides individual and generic representations (social categories) of persons, the model contains information of social situations, social memory tasks, and social strategies. The Person Memory Processing Unit delegates incoming queries to memory tasks appropriate in a given interaction context. The white *S* denotes the representation of the agent's self.

3.1 Persons and Social Categories

The heart of our model of person memory consists of representations of the persons the agent interacts with. We consider the following information as fundamental to be remembered about a person (cf. [16]): biographical facts, preferences and interests, personality traits, events, and relationship information.

The *individual* representation of a person consists of instances of such types of information. During encounters such information is stored directly in the individual representation. Figure 2 depicts example representations of the embodied conversational agent Max and a person known by the agent. In addition to the information stored in the individual representation, further generic representations may be linked which derive from *social categories*.



Key	Value
name	Max
age	14
hometown	Bielefeld
knows	[Person: Paula]
knows	[Person: Paul]
interest	[Interest: chess]
interest	[Interest: music]
memberOf	[Category: extrovert]
memberOf	[Category: artificialperson]

Key	Value
name	Paula
age	26
hometown	Bielefeld
knows	[Person: Max]
interest	[Interest: music]
interest	[Interest: cinema]
memberOf	[Category: introvert]
memberOf	[Category: sportsstudent]

Fig. 2. Two examples of individual representations in the person memory of the conversational agent Max: a representation of Max himself and a representation of a person Paula known by Max

Within the person memory social categories are used to represent stereotypical information of groups of people, for instance shared interests and preferences. Furthermore some information, like personality traits and relationship, can be inferred from such generic representations (see [18]). During conversation with a person of a certain category, the agent can use the stereotypical information as hints of what the interlocutor might like to talk about.

3.2 The Interaction Context and Social Situations

Not only the personality of individuals, and the relationship between them, affect the interactant's behavior, but also the social situation the interaction takes place in [22]. So it is not sufficient to consider an interaction between two individuals without its context.

In our model, the interaction context consists of the social situation and the combined representations of the agent and the interlocutor (see Fig. 1). By joining the representation of the *I* (agent's self) and the representation of the *You* (interlocutor), a representation of the *We* is constructed, consisting of generic and individual information that is relevant to the current social situation.

A representation of a social situation contains a description of the situation (e.g., name, type, location) and information that can have influence on the agent's behavior towards its interlocutor. The social categories described in Sect. 3.1 may contain triggers that are sensitive to certain situations. This allows to include information from specific categories when triggered by a social situation (cf. [18]).

To exploit the information provided by the interaction context, the agent is equipped with *Social Memory Tasks* and *Social Strategies*.

3.3 Social Memory Tasks and Social Strategies

As described in [17], three different kinds of tasks are associated with the information of a memory: storage, access, and manipulation. Two groups of social memory tasks (*core* and *extended* tasks) fulfill these actions in the person memory: Tasks of the first group handle basic actions, like storage of new information, or retrieval of existing information. Tasks of the second group carry out more context based information retrieval and manipulation on the data provided by, e.g., the interaction context. Examples for tasks of the second group are:

- Calculating probabilities for the use of
  - dialogue sequences (*"Question/Answer"* vs. more complex sequences like *"Question"/ "Counter"/ "Probe"/ "Reply"*, cf. [15])
  - topics from different topic categories (*"communication"*, *"immediate"*, *"external"*, cf. [5])
- Selecting a topic category according to the calculated probabilities
- Selecting a topic (from this category) for conversation

While core tasks are predefined operations, extended tasks can be exchanged dynamically at run time. This allows to define tasks that include different information when, for instance, selecting a topic the agent should bring up during conversation: While one task takes all available information into account that is located in the interaction context about the interlocutor and the agent, a second task may only consider the representation of the agent.

To activate appropriate tasks for a given situation the person memory contains a set of instructions in the form of *social strategies*. Each social strategy contains at least one *trigger* that is sensitive to certain social situations. Furthermore social strategies contain a mapping of social memory tasks to *keywords*. The keywords are predefined and used to identify tasks in the person memory.

3.4 Person Memory Processing Unit

The *Person Memory Processing Unit* (PMPU) provides an interface between the person memory and the cognitive architecture of an agent. In that, it handles the communication between different components, like a dialog manager, or further memory components.

This way, the proposed model of Person Memory enables an agent to cope with social encounters that may occur in different application scenarios. The information provided about the persons an agent interacts with, and the strategies that influence how the information is exploited, allow for an adaptation of the agent to different settings [18].

4 Assessing the Impact of Personal Information in Human-Agent Conversations

One of the assumptions that motivated the development of the person memory presented in this paper, is that personal information can be exploited to

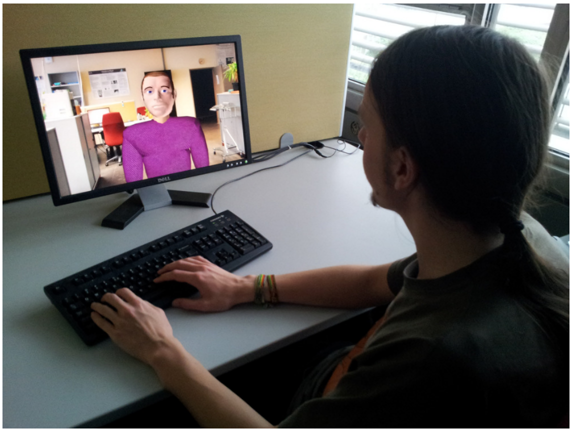


Fig. 3. Setup used for our interaction study: A human interlocutor and the virtual agent Max conducting a conversation

enhance the interactions between a virtual agent and its human interlocutors. For instance, based on work by Deborah Tannen, Svennevig [21] states that in human-human conversations *"the preference for personal topics is a case of involvement in the interlocutor."* (p. 52). To assess the impact of personal information in human-agent conversations, and thereby evaluating the performance of our model of person memory, we conducted an interaction study with 22 particpants (see below).

4.1 Setup and Hypotheses

An embodied conversational agent, Max, is used in our work as research platform for human-computer interaction (see Fig. 3). Max's usefulness as a conversational partner is already demonstrated in a museum setting where he explains exhibits and engages visitors in small talk conversations since 10 years [11].

The study was split into two sessions: an initial interaction (*getting to know*) and a second encounter (*meeting again*). In both sessions, particpants were asked to get seated in front of a 24" computer monitor that was used to display the agent (Fig. 3). The particpants had to input their utterances using a keyboard, while Max's questions and answers were generated by a voice synthesizer.

During the first encounter participants were engaged in a short small talk with our agent. They were instructed that the first interaction is to get comfortable with the agent and the way of interacting with it, and that Max would end the conversation. Max used this first encounter to gather personal information, like interests and hobbies (e.g., *"Do you like to read?"*), about its interlocutor. The participants were asked to come back for a second conversation with Max.

In advance of their second encounter with Max, participants were randomly assigned to two groups (while maintaining gender balance between both groups), leading to a *between-subject* design of our study. In conversations with members of the first (control) group (**Group A**, $N = 11$), Max did not exploit the information stored in his person memory, by using customized tasks and strategies (see section 3.3). Instead he sticked to impersonal topics about the immediate and external situation, like the weather and recent events in the surrounding (e.g., *"A lot of construction going on here in the university building, right?"*).

During conversations with members of the second group (**Group B**, $N = 11$), Max recalled personal information it gathered in the first encounter and used this information as topics for the ongoing conversation (e.g., *"Are you reading something special right know?"*). Again participants were told to wait until Max ended the conversation.

Our hypotheses were as follows: If the agent Max exploits the personal information of his person memory during conversation, then

H1 the social presence of Max will be rated higher.
H2 Max's interlocutors will be more satisfied with the conversation.
H3 the impression that Max knows and remembers oneself is stronger.
H4 the participants' trust in the agent is stronger.
H5 the overall impression of Max will be more positive (in terms of sympathy, friendliness etc.).

4.2 Questionnaire

To assess how the agent is perceived in terms of social presence (*SP*), 3 items from a social presence questionnaire from [1] and 6 items of the *Networked Minds Social Presence Inventory* [4] were selected. Factor analysis with varimax rotation revealed two underlying factors that explain 58.09% of the variance. Cronbach's α was .82 for the first factor (5 items) and $\alpha = .8$ for the second factor (4 items).

In addition, 11 items from the *Interpersonal Communication Satisfaction Inventory* [8] were used to test how the interlocutors liked the overall conversation (*CS*). Again a factor analysis with varimax rotation resulted in two factors explaining 54.83% of the variance. Cronbach's α was .91 for the first factor (8 items) and $\alpha = .78$ for the second factor (3 items).

Three items (only included in Session 2) were used to test hypothesis **H3** (Cronbach's $\alpha = .81$). To test hypothesis **H4** an additional item (*"If nobody else was in the room, I would have no problem telling personal secrets to Max."*) from the questionnaire of [1] was used. As manipulation checks (*MC*), two items were added that focused on whether the topics of the conversation were considered personal or impersonal. All items were rated on a 7-point Likert-scale (with 1=*strongly disagree*, 7=*strongly agree*).

A semantic differential with 15 bi-polar adjective pairs (7-point scale) was used to assess the overall impression of Max (**H5**).

4.3 Participants

Participants were recruited within Bielefeld University through postings and a mailing list of people who previously attended interaction studies in the VR lab of our group. Initially a total of $N = 26$ people (7 females and 19 males) participated in our study.

However, two people had to be excluded from the final evaluation as they did not fully complete the questionnaires. Two further participants were excluded, since the conversational agent system got stuck during the interaction. Thus, in the final evaluation $N = 22$ particpants (11 particpants per group) were considered, with a mean age of 30.27 ($SD = 12.78$), ranging from 20 to 64. 17 of the participants were students.

The average time between first and second session was 6.23 ($SD = 1.82$) days, with an average duration of conversations of 6.17 ($SD = 1.58$) minutes in Session 1 and 4.2 ($SD = 0.80$) minutes in Session 2. No significant difference was found in terms of time between sessions and duration of conversations.

4.4 Results

The non-parametric *Mann-Whitney U* test was used to compare the results of the questionnaires of both groups.

Session 1. Since the first session did not differ between groups, it was expected that there should be no statistical significant differences in the results. This held true for the *SP* and *CS* factors. However, a significant difference was found for one questionnaire item (*"The conversation flowed smoothely."*) in the first session ($Mdn_A = 5$, $Mdn_B = 6$, $U = 22.00$, $z = 2.61$, $p < 0.05$, $r = -0,56$), and for one dimension of the semantic differential (*superficial - profound*, $Mdn_A = 3$, $Mdn_B = 5$, $U = 28.50$, $z = 2.14$, $p < 0.05$, $r = -0.46$). These differences did not emerge in the results of the second session. For the remaining items no significant differences were found for the first session, as expected.

Session 2. Hypothesis **H1** predicted that Max's social presence will be rated higher when he uses more personal information during conversations. However, there were no significant differences in the factors that constitute the *SP* measure. Thus, hypothesis *H1* was not supported. Still, a significant difference was found in one item directly targeting the presence of Max (table 1, item 1).

Hypothesis **H2** predicted that the participants will be more satisfied with the conversation when Max uses personal topics. Both *CS* factors showed a significant difference (CS_{F1}: $Mdn_A = 38.00$, $Mdn_B = 44.00$, $U = 31.00$, $z = -1.94$, $p < 0.05$, $r = -0.41$; CS_{F2}: $Mdn_A = 13.00$, $Mdn_B = 17.00$, $U = 15.00$, $z = -3.01$, $p < 0.05$, $r = -0.64$). Thus, hypothesis *H2* was confirmed. Four out of 11 items were found to show a significant difference (see table 1, items $2-5$).

All items that were used to test hypothesis **H3** were found to show significant differences between both groups as depicted in table 1 (items $6 - 8$). Thus, hypothesis *H3* was confirmed. Furthermore, the results of items 9 and 10 show that the topics addressed by Max were judged as more impersonal resp. personal in the corresponding conditions.

Considering the results of the item used to test hypothesis **H4**, the hypothesis was not confirmed: Participants of both groups rejected the idea of telling personal secrets to the agent ($Mdn_A = 2$, $Mdn_B = 3$), with no significant difference between groups.

Table 1. Results of the statistical analysis for selected items of the social presence (SP), communication satisfaction (CS), person memory performance (PM), and manipulation check (MC) parts of the questionnaire. Among the medians Mdn for both groups, the U value of the Mann-Whitney test, z-score, level of significance p ($* < 0.05$, $** < 0.001$), and effect size r are given.

		Item	Mdn_A	Mdn_B	**U**	**z**	**p**	**r**
SP	1.	I perceive that I am in the presence of another person in the room with me.	3	5	26.50	$-2,27$	$*$	$-0,48$
CS	2.	We talked about something I was not interested in.	3	2	27.00	-2.28	$*$	-0.49
	3.	I was very dissatisfied with the conversation.	3	1	33.00	-1.86	$*$	-0.40
	4.	Max genuinely wanted to get to know me.	3	5	10.00	-3.38	$**$	-0.72
	5.	I would like to have another conversation like this one with Max.	5	6	31.00	-1.99	$*$	-0.42
PM	6.	I had the feeling that Max knows me.	4	6	13.00	-3.18	$*$	-0.68
	7.	Max remembered me very well.	4	7	4.00	-3.81	$**$	-0.81
	8.	Max did not remember me at all.	1	1	33.00	-2.46	$*$	-0.52
MC	9.	The questions that Max posed were very personal.	2	5	23.00	-2.53	$*$	-0.54
	10.	The questions that Max posed were very impersonal.	5	3	24.00	-2.44	$*$	-0.52

The results of the semantic differential are given in fig. 4. Compared to the results of all participants of the first session, four dimensions (*superficial–profound*, *silly–serious*, *reliable–unreliable*, *offhanded–chatty*) were rated in a more positive direction and only two more negatively (*impolite–polite*, *introverted–extroverted*) by participants of *Group B*. Whereas none of the dimensions were rated better and 12 more negatively by participants of *Group A*. Between groups, five dimensions (*unsocial–social*, *reliable–unreliable*, *personal–impersonal*, *likeable–unlikeable*, *strange–intimate*) showed a significant difference ($p < 0.05$) in Session 2.

4.5 Discussion

Given the overall results, we conclude that the use of personal topics indeed has quite a *positive impact* on human-agent conversations. Participants were *more*

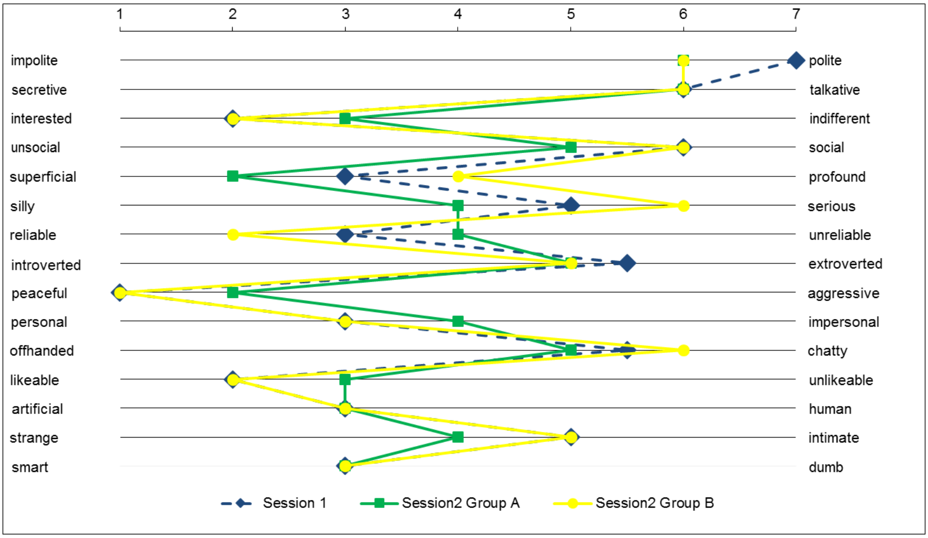


Fig. 4. Results of the semantic differentials used to assess particpants' attitudes towards the agent

satisfied with the conversation and had a *more positive attitude* towards our agent after the second conversation.

The social presence of Max was not found to be affected by the use of more personal topics in conversation. Furthermore, both groups scored rather low on the two factors of social presence (normalized median factor scores for *Group A* are $SP_{F1} = 4.4$, $SP_{F2} = 3$, and $SP_{F1} = 4.8$, $SP_{F2} = 4$ for *Group B*). An explanation could be the setup of our study: Reactions of our agent were generated based on keyboard input only, no camera or microphone were attached to provide further input to the agent. While this was necessary to prevent uncontrollable behavior of the agent, the agent did not appear to genuinely take notice of the participants.

Regarding our model of person memory, the interaction study demonstrated that it enabled the agent to *successfully remember* useful information as hypothesis *H3* was confirmed. Remarkably, most participants of the control group felt that Max somehow remembered them as well (see table 1, item 8, $Max_A = 4$), although he did not explicitly talk about their first conversation or information stored in his person memory. This could be due to the fact that Max talked about things going on in the university and most of the participants were students.

5 Conclusion

In this paper, a model of person memory for artificial agents was presented. The main ingredients we identified – representations of persons and social categories, knowledge about social situations, social memory tasks, social strategies, and a central processing unit – build a foundation for a *social memory component* in the architecture of a conversational agent.

Our initial interaction study demonstrates that our model can be successfully exploited by our agent to *obtain, store*, and *recall* information about his interaction partners. Furthermore, the mechanisms to guide the agent's conversational behavior were successfully used to *adapt* the agent according to the two conditions (i.e., use of impersonal resp. personal topics) of our study.

The results of our initial interaction study show that social conversations between an agent and its human interlocutors *benefit* from the use of personal information in subsequent encounters. This we regard as evidence that we are on the right track by stressing the importance of a specialized memory component to represent the people an agent interacts with, as done within our person memory.

We expect a more thorough interaction study, where additional aspects of an agent's behavior are adapted according to the individual representations of the agent's interaction partners (see [16]), could further underpin these findings.

References

1. Bailenson, J.N., Beall, A.C., Blascovich, J., Raimundo, M., Weisbuch, M.: Intelligent Agents Who Wear Your Face: Users Reactions to the Virtual Self. In: de Antonio, A., Aylett, R.S., Ballin, D. (eds.) IVA 2001. LNCS (LNAI), vol. 2190, pp. 86–99. Springer, Heidelberg (2001)
2. Becker, C., Wachsmuth, I.: Modeling primary and secondary emotions for a believable communication agent. In: Reichardt, D., Levi, P., Meyer, J.J.C. (eds.) Proceedings of the 1st Workshop on Emotion and Computing, pp. 31–34. Bremen (2006)
3. Bickmore, T.W.: Relational Agents: Effecting Change through Human-Computer Relationships. Ph.D. thesis, Massachusetts Institute of Technology (2003)
4. Biocca, F., Harms, C., Gregg, J.: The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In: 4th Annual International Workshop on Presence, Philadelphia, PA (2001)
5. Breuing, A., Wachsmuth, I.: Let's Talk Topically with Artificial Agents! Providing Agents with Humanlike Topic Awareness in Everyday Dialog Situations. In: ICAART 2012 - Proceedings of the 4th International Conference on Agents and Artificial Intelligence, pp. 62–71. SciTePress (2012)
6. Brom, C., Lukavský, J., Kadlec, R.: Episodic Memory for Human-like Agents and Human-like Agents for Episodic Memory. International Journal of Machine Consciousness 2(2), 227–244 (2010)
7. Endrass, B., Rehm, M., André, E.: Planning Small Talk behavior with cultural influences for multiagent systems. Computer Speech & Language 25(2), 158–174 (2011)
8. Hecht, M.L.: The conceptualization and measurement of interpersonal communication satisfaction. Human Communication Research 4(3), 253–264 (1978)
9. 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)

10. Kang, S., Gratch, J., Sidner, C., Artstein, R., Huang, L., Morency, L.P.: Towards building a Virtual Counselor: Modeling Nonverbal Behavior during Intimate Self-Disclosure. In: Eleventh International Conference on Autonomous Agents and Multiagent Systems, Valencia, Spain (June 2012)
11. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A Conversational Agent as Museum Guide - Design and Evaluation of a Real-World Application. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 329–343. Springer, Heidelberg (2005)
12. Leite, I., Martinho, C., Paiva, A.: Social Robots for Long-Term Interaction: A Survey. International Journal of Social Robotics 5(2), 291–308 (2013)
13. Mairesse, F., Walker, M.A.: Towards Personality-Based User Adaptation: Psychologically Informed Stylistic Language Generation. User Modeling and User-Adapted Interaction 20, 227–278 (2010)
14. Mattar, N., Wachsmuth, I.: A Person Memory for an Artificial Interaction Partner. In: Proceedings of the KogWis 2010, pp. 69–70 (2010)
15. Mattar, N., Wachsmuth, I.: Small Talk Is More than Chit-Chat – Exploiting Structures of Casual Conversations for a Virtual Agent. In: Glimm, B., Krüger, A. (eds.) KI 2012. LNCS, vol. 7526, pp. 119–130. Springer, Heidelberg (2012)
16. Mattar, N., Wachsmuth, I.: Who Are You? On the Acquisition of Information about People for an Agent that Remembers. In: ICAART 2012 - Proceedings of the 4th International Conference on Agents and Artificial Intelligence, pp. 98–105. SciTePress (2012)
17. Mattar, N., Wachsmuth, I.: Adapting a virtual agents conversational behavior by social strategies. In: Timm, I.J., Thimm, M. (eds.) KI 2013. LNCS, vol. 8077, pp. 288–291. Springer, Heidelberg (2013)
18. Mattar, N., Wachsmuth, I.: Strangers and friends: Adapting the conversational style of an artificial agent. In: Kurosu, M. (ed.) HCII/HCI 2013, Part V. LNCS, vol. 8008, pp. 102–111. Springer, Heidelberg (2013)
19. Nuxoll, A., Laird, J.: Extending Cognitive Architecture with Episodic Memory. In: Proceedings of the National Conference on Artificial Intelligence, vol. 22, p. 1560. AAAI Press. MIT Press, Menlo Park, Cambridge (1999, 2007)
20. Rosis, F.D., Pelachaud, C., Poggi, I.: Transcultural Believability in Embodied Agents: A Matter of Consistent Adaptation. In: Agent Culture: Designing Human-Agent Interaction in a Multicultural World, pp. 1–22. Laurence Erlbaum Associates (2004)
21. Svennevig, J.: Getting Acquainted in Conversation: A Study of Initial Interactions. Pragmatics & beyond, John Benjamins Publishing Company (1999)
22. Zayas, V., Shoda, Y., Ayduk, O.N.: Personality in Context: An Interpersonal Systems Perspective. Journal of Personality 70(6), 851–900 (2002)

Multimodal Behaviours in Comparable Danish and Polish Human-Human Triadic Spontaneous Interactions

Costanza Navarretta and Magdalena Lis

University of Copenhagen, Njalsgade 140
Build. 25, 4th floor, 2300 Copenhagen S
Denmark
{costanza,magdalena}@hum.ku.dk http://cst.ku.dk

Abstract. This is a pilot study of multimodal behaviours in manually annotated comparable video recordings of Danish and Polish triadic naturally occurring conversations. The data are comparable with respect to the conversational settings, the familiarity degree, age and gender of the participants. Furthermore, they have been annotated according to the same annotation scheme following common coding strategies. The analysis of the annotations indicates that although the conversations in the two languages differ in content, Danes and Poles use the same type of head movements and with the same frequency. In both datasets the most common facial expressions are laughter and smile, however, facial expressions are much more frequent in the Polish data than in the Danish data. Furthermore, the facial expressions in the Polish data are often used as feedback signals to the interlocutors while the Danes use facial expression to comment their own spoken contribution. Finally, the Danes use more frequently hand gestures than the Poles and their hand gestures have a deictic function while the hand gestures of the Poles are iconic. The differences in the behaviours in the two corpora can partly depend on the language, but is also due to the type of relationship between the participants and the content of the conversations.

Keywords: Multimodal Corpora, Multilinguality, Human-human Interaction.

1 Introduction

The paper presents a comparative study of multimodal behaviours in video-recorded and manually annotated triadic human-human spontaneous conversations in Danish and Polish. The study focuses on feedback signalled by facial expressions and head movements, and on iconic and deictic hand gestures. The use of facial expressions and especially head movements as signals that a conversational participant is giving (backchannelling) and/or eliciting feedback has been extensively studied both monolingually inter alia [1–3] and multilingually e.g. [4–8]. The same is the case for iconic hand gestures, that is gestures which

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 462–471, 2014.

represent aspects of the entities mentioned in the discourse, e.g. [9–12] Most comparative studies of feedback have studied head movements and, to a lesser extent, facial expressions in dyadic first encounters, inter alia [4, 6–8], while studies of iconic gestures have in many cases investigated behaviour in narratives where the participants re-told the same story from a cartoon, see inter alia [10, 12–14].

Differing from these studies, we base our work on triadic naturally occurring conversations between people who are well-acquainted. These data are more difficult to compare since the variation in discussed issues is much wider than in first encounters or in narratives of common stories. However, it is also important to investigate cultural variation between people who are well-acquainted since familiarity influences conversational behaviours, e.g. [15].

Annotated multimodal data of people who have high familiarity degree should be used to model re-current human-robot interactions, reflecting the fact that communication is affected by the growing familiarity degree of the participants [16].

The present work builds upon a preceding study [17] in which we analysed feedback head movements in the same data. This study indicated that Danish and Polish participants used many types of head movements to signal feedback. The type of non-verbal behaviour in the two groups was the same. Furthermore, there were also similarity in the type of feedback spoken expressions. However, Polish participants in these data express more often feedback multimodally, that is through both modalities, and they use significantly more repeated multimodal feedback expressions than the Danish participants. Moreover, the data showed that there are significantly more repeated head movements and speech tokens in the Polish conversations than in the Danish ones.

In the present work, we extend the preceding study which analysed exclusively feedback head gestures, by including all types of head movements, as well as facial expressions and hand gestures. We expect that feedback facial expressions and iconic gestures in the data will strongly depend on the content of the conversations, but we also expect that there will be many mirroring facial expressions and iconic gestures due to the fact that the conversational participants know each other very well.

The paper is organised as follows. We start presenting related studies in section 2 and describing the corpora and their annotations in section 3. Then, in section 4 we account for the multimodal behaviours in the two corpora and we discuss similarities and dissimilarities in the data 5. Finally in section 6, we conclude and present future work.

2 Related Studies

There is agreement in research that especially head nods, head shakes and smiles are common feedback signals which occur both unimodally and multimodally, see inter alia [1, 7, 18]. Furthermore, researchers have found differences in the way people express feedback multimodally in different cultures. For example, [4] compares the occurrences of feedback head nods in dialogues between Japanese and Americans and finds that the Japanese nod with higher average frequency of the Americans. Rehm et al. [6, 19] have investigated the occurrences of feedback behaviour of German and Japanese speakers in first encounters, and conclude that German speakers gesture more frequently than Japanese speakers. Furthermore, the amplitude and speed of the Germans' gestures are higher than those of the Japanese participants. [7] have analysed cultural differences in first encounters between Swedish and Chinese speakers, and found that the Chinese use more laughter, gaze around, gaze sideways and covering their mouth with their hands. However, head movements and tilts only occur in the Swedish conversations and both Chinese and Swedish participants give more gestural feedback when they speak English in intercultural interaction with each other than when they communicate in their mother language.

[8] compare feedback head nods and shake in first encounters in three Nordic languages, Danish, Finnish and Swedish and feedback speech expressions in Danish and Swedish. The results of the comparison show that the most frequent feedback head movement is nod in all languages. However, the Danish participants use down-nods more frequently than the Finnish and Swedish subjects, while the Swedish participants use up-nods more frequently than Finnish and Danish. Finns use more often single nods than repeated nods, differing from Swedes and Danes. The differences in the frequency of different types of nods in the three datasets are interesting because Nordic countries are not only geographically near, but are also very similar culturally. Finally, the comparison of feedback spoken expressions in the Danish and Swedish first encounters indicates that Swedes and Danes use common feedback words such as yes and no expressions with similar frequency.

Navarretta and Paggio [15] compare feedback head movements and facial expressions in Danish conversations where the participants did not know each other and in conversations where the participants were well acquainted. The results indicate that familiarity degree seems to affect the frequency of feedback signals so that subjects who are familiar with one another move their heads more than people who do not know each other: this effect seems parallel to the increased flow of speech in relation to familiarity that has been observed in the literature [16].

In this study we compare feedback head movements and facial expressions in naturally occurring corpora in Danish and Polish.

3 The Data

The Danish data consists of two conversations that are part of the MOVIN database collected by Conversational Analysis researchers at Southern Danish University [20]. The participants are all Danish native speakers and talk freely while they drink and eat. The three participants were sitting in an half circle around a table in a private home and were filmed by one video-camera. The conversations were then orthographically transcribed and multimodally annotated as part of the Danish CLARIN project [21, 22].

The Polish conversation was collected, transcribed and annotated under the European CLARA project. The three Polish participants are all native speakers and were also recorded by two cameras at a private home, while sitting around a table, eating, drinking and talking freely. All the participants were aware that they were video-taped, but the recording equipment was well incorporated in the space in order to obtain as natural data as possible.

The setting, the degree of familiarity of the participants, their number and age are similar in the two corpora. It must,however, be noted that the content of the conversations differs, since they were naturally occurring. Finally, our data are also comparable because they were transcribed according to the same principles and were multimodally annotated according to the same annotation scheme [23]. This scheme describes the shape and function of body behaviours with independent attribute and value pairs, and allows the annotators to link body behaviours to speech if they judge that speech and body behaviours are semantically or pragmatically related. The annotations, which are relevant to this study, concern speech, head movements, facial expressions, and hand gestures.

Speech was orthographically transcribed at the word level. The transcriptions were performed by native speakers in Praat [24]. Speech tokens were time aligned. The annotations comprise both lexical and non-lexical words such as *hm* and *mm*. Multimodal behaviours were annotated in the ANVIL tool [25].

The annotations of the shape of the body behaviours that are used in these study are in Table 3. The relevant features for head movements are information about the type of movement and repetitiveness of the moment. For facial expression we have only included general features describing the face while for hand gestures we have considered features describing handedness, the shape of fingers and repetitiveness of the movement. In this paper, we also use the annotations indicating whether the behaviours have a feedback function. In particular we distinguish between feedback giving (backchannelling), feedback eliciting and selffeedback (see [3, 23] for more details on the annotation of feedback in MUMIN). We also use the semiotic characteristics of hand gestures. The annotation

Table 1. Shape Features of Body Behaviours

Attribute	Value
HeadMovement	Nod, Jerk, HeadForward, HeadBackward,Tilt, SideTurn, Shake, Waggle, HeadOther
HeadRepetition	Single, Repeated
General face	Smile, Laugh, Scowl, FaceOther
Handedness	SingleHand, BothHands
Hand-Repetition	Single, Repeated
Fingers	IndexExtended, ThumbExtended, AllFingersExtended,FingersOther

of semiotic types builds upon the types proposed by [26]. The types relevant to this study are the following: *indexical gestures* which have a real and direct connection with the objects they denote and comprise *deictic* (pointing) and *non-deictic* gestures, e.g. beats and displays; *iconic gestures*, also known as *emblems* or *illustrators*, which denote their objects by similarity, and *symbolic gestures* which are established by means of an arbitrary conventional relation.

Inter-coder agreement experiments on the Danish annotations measured in terms of Cohen's kappa [27] resulted in agreement features between 0.7 and 0.9 depending on the annotations [21]. The annotations used in this study comprise over 3000 words and 1000 body behaviours in each language.

4 Comparison of the Data

This section contains an analysis of the speech tokens and body behaviours annotated in the two corpora. In table 4 the frequencies per second of the various behaviours in the two datasets are given. The frequency of speech tokens in the Danish data is 3.86 per second, while it is 4.14 per second in the Polish data. The frequency of body behaviours is 1.3 in the Danish data and 1.28 in the Polish data.

Table 2. Behaviour type per second

Behaviour	Danish	Polish
Speech	3.86	4.14
All Head	0.9	0.93
Feedback Head	0.56	0.59
Face	0.06	0.14
Hand	0.34	0.21
All	1.3	1.28

In each conversation, there is one participant who talks more than the other two participants, but the Danish data are more balanced with respect to speech token production than the Polish where one participant speaks much more than the two others.

The most frequent body behaviour in both datasets are head movements and their frequency (0.9 head movements per second in the Danish data vs. 0.93 head movements per second in the Polish data) is similar. The frequency of feedback head movements in the two datasets is also similar (0.56 per second in the Danish data and 0.59 in the Polish data) and the types of feedback head movements produced in the two datasets are the same. However, the Poles use significantly more repeated head movements and feedback spoken expressions than the Danes as reported in [17].

The majority of facial expressions in both languages are coded as smiles and laughs. Smile and laughs are equally frequent in the Danish data and most of them are coded as selffeedback. In the Polish data, the most common facial expression is laugh and two thirds of the laugh and smile occurrences signal feedback give. Furthermore, facial expressions are much more frequent in the Polish corpus than in the Danish one: the frequency of facial expressions in Polish is 0.14 per second while it is 0.06 in the Danish data.

The analysis of the data also shows that there are more co-occurrences of smile and laugh (overlapping occurrences of the behaviours between at least two participants) than in the Danish data.

Different is the situation for hand gestures. The Danish participants gesture more often than the Polish participants (0.34 hand gestures per second in the Danish data vs. 0.21 hand gestures per second in the Polish data). The most common hand gesture type (over two third of the occurrences) in the Polish data is iconic, while in the Danish data it is deictic (90 % of the gesture occurrences). The iconic gestures in the Polish data are often repeated more times, and in some cases the participants repeat each other gestures.

5 Discussion

The analysis of the behaviours annotated in the triadic Danish and Polish conversations indicate that the Poles utter more speech tokens per second than the Danes. The average length of the speech tokens in the two languages should be investigated in future. The data also indicates that in all conversations there is a subject who talks more than the two others, but the Danish data is more balanced than the Polish data under this respect. The frequency of body behaviours per second in the two languages is exactly the same, thus the ratio gesture per speech token is 0.34 in Danish and 0.31 in Polish. Whether these figures reflect a characteristics of the two languages should be investigated further.

Although the conversations have different content, the type and frequency of head movements in the two languages is similar. The Poles, however, use significantly more repeated feedback head movements and spoken expressions in these data [17]. Future studies should investigate whether the frequency of repeating feedback signals is a Polish characteristics or it is specific for these data. Navarretta and Paggio [15] compared the triadic conversations used in this study and the dyadic DK-CLARIN conversations with Danish dyadic first encounters and find out that the triadic DK-CLARIN data contained most feedback head movements and least feedback facial expressions. Thus the frequency of feedback head movements ins Danish seems to be influenced by the number of subjects involved in the conversations and their degree of familiarity. It should be investigated in future whether this is also the case in Polish, but we found interesting that the frequency and type of head movements in general and feedback head movements are so similar in the two datasets.

Facial expressions are more common in the Polish data than in the Danish data, and the Polish participants laugh much more than the Danish subjects. This is certainly due to the content of the conversations, but it might also depend on the relationship between the participants. In the two Danish conversations the participants are family members and belong to different generations: two of the participants have approximately the same age while the third one is much older. The Polish participants, on the other hand, have the same age being old schoolmates.

The facial expressions in the Danish data are often produced by the speaker as comments to her own spoken contribution, while the facial expressions in the Polish conversation signal feedback giving and/or eliciting and they are often overlapping with the same behaviour of the addressees. This is not so surprising since laughter, which is the most frequent facial behaviour, is often a collective activity. However, the synchronic facial expressions also include smiles, thus they might indicate that the participants are aligning to each other, see inter alia [28, 29].

The Danes use more hand gestures than the Poles, furthermore deictic hand gestures are most common in the Danish data while iconic gestures are more common in the Polish data. Furthermore, the Polish participants reproduce their own and the other participants' iconic gestures, which seems to confirm that they are in synchrony. In future, we should investigate whether there is also synchrony in speech and head movements.

6 Conclusion

In this paper, we have presented a pilot comparative study of multimodal body behaviours in two small corpora of annotated video-recordings of Danish and Polish triadic naturally occurring conversations. The data in the two languages are comparable because the conversational settings are similar as well as the age and degree of familiarity of the participants. They are also comparable because the data have been manually annotated according to the same annotation scheme and using the same annotation tools. Differing from most other multilingual comparable corpora, our data record friends or family members talking at their private homes. While these data are natural and reflect natural conversations, they are problematic with respect to the content which varies from conversation to conversation. Thus, they are more difficult to compare than map-task dialogues, narratives of pre-defined stimuli and first encounters interactions which have often been used in comparative multimodal studies.

Since the participants in the conversations in the current study are well acquainted, these data can be valuable for modelling re-current human-robot interactions behaviours of robot or software agent. These models should reflect the fact that communication is affected by the growing familiarity degree of the participants [16].

The analysis of data indicates that the frequency of speech tokens is higher in the Danish than in the Polish corpus, while the frequency of body behaviours per second is similar in the two datasets. However there are both differences and similarities in the occurrences of the various types of behaviours. The frequency and types of head movements as well as their use is similar even though the content of the conversations is different. This confirms a preceding study [17] which also indicated that the Poles use more repeated multimodal feedback expressions than the Danes.

The Danes use less facial expressions than the Poles in these data and they laugh less. Furthermore they mostly smile and laugh commenting their own spoken contribution, while the Poles laugh and smile giving and eliciting feedback. In these cases, their facial expression is the same as that of their interlocutors.

The Danes move much more the hands than the Poles. The most common hand gestures in the Danish data are deictic, while the most common gestures in the Polish corpus are iconic. Also in the case of hand gestures the Poles repeat more often their own gestures and those of their interlocutors. These kinds of overlapping behaviours are considered to indicate synchrony between the conversants, see i.a. [28, 29]. The differences in the type and frequency of facial behaviours and hand gestures in the two datasets can depend on the specific content of the conversations and on the type of relation between the participants in the Danish and Polish data. However, they can also be idiosyncratic to these data or partially depend on the types of language. Thus these aspects should be investigated in larger datasets and in more types of conversation.

References

1. McClave, E.: Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32, 855–878 (2000)
2. Cerrato, L.: Investigating Communicative Feedback Phenomena across Languages and Modalities. PhD thesis, Stockholm, KTH, Speech and Music Communication (2007)
3. Paggio, P., Navarretta, C.: Head movements, facial expressions and feedback in Danish first encounters interactions: A culture-specific analysis. In: Stephanidis, C. (ed.) Universal Access in HCI, Part II, HCII 2011. LNCS, vol. 6766, pp. 583–590. Springer, Heidelberg (2011)
4. Maynard, S.: Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. Journal of Pragmatics 11, 589–606 (1987)
5. Jokinen, K., Navarretta, C., Paggio, P.: Distinguishing the communicative functions of gestures. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 38–49. Springer, Heidelberg (2008)
6. Rehm, M., Nakano, Y., André, E., Nishida, T.: Culture-Specific First Meeting Encounters between Virtual Agents. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 223–236. Springer, Heidelberg (2008)
7. Lu, J., Allwood, J., Ahlsén, E.: A study on cultural variations of smile based on empirical recordings of Chinese and Swedish first encounters. In: Heylen, D., Kipp, M., Paggio, P. (eds.) Proceedings of the Workshop on Multimodal Corpora at ICMI-MLMI 2011, Alicante, Spain (2011)
8. Navarretta, C., Ahlsn, E., Allwood, J., Jokinen, K., Paggio, P.: Feedback in nordic first-encounters: a comparative study. In: Proceedings of LREC 2012, Istanbul, Turkey, pp. 2494–2499 (May 2012)
9. Jacobs, N., Garnham, A.: The role of conversational hand gestures in a narrative task. Journal of Memory and Language 56(2), 291–303 (2007)
10. Kita, S., Özyurek, A.: How does spoken language shape iconic gestures? In: Duncan, S., Cassel, J., Levy, E. (eds.) Gesture and the Dynamic Dimension of Language, pp. 67–74. Benjamins, Amsterdam (2007)

11. Gullberg, M.: Thinking, speaking and gesturing about motion in more than one language. In: Pavlenko, A. (ed.) Thinking and Speaking in two Languages, pp. 143–169 (2011)
12. Lis, M., Parrill, F.: Referent type and its verbal and gestural representation: A test on English multimodal corpus and WordNet 3.1. In: Extended Abstract in Proceedings of the 1st European Symposium on Multimodal Communication, MMSym 2013 (2013)
13. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press, Chicago (2000)
14. Goldin-Meadow, S., Chee So, W., Ozyurek, A., Mylander, C.: The natural order of events: How speakers of different languages represent events nonverbally. Proceedings of the National Academy of Sciences of the USA 105(27), 9163–9168 (2008)
15. Navarretta, C., Paggio, P.: Verbal and non-verbal feedback in different types of interactions. In: Proceedings of LREC 2012, Istanbul Turkey, pp. 2338–2342 (2012)
16. Campbell, N.: Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. In: Proceedings of the COST 2007 Workshop, pp. 107–120 (2007)
17. Navarretta, C., Lis, M.: Multimodal feedback expressions in Danish and Polish spontaneous conversations. In: Allwood, J., et al. (eds.) NEALT Proceedings. Northern European Association for Language and Technology, Proceedings of the Fourth Nordic Symposium of Multimodal Communication, Gothenburg, Sweden, November 2012. Linköping Electronic Conference Proceedings, pp. 55–62 (2013)
18. VYngve, V.: On getting a word in edgewise. Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, pp. 567–578 (1970)
19. Rehm, M., et al.: Creating Standardized Video Recordings of Multimodal Interactions across Cultures. In: Kipp, M., Martin, J.-C., Paggio, P., Heylen, D., et al. (eds.) Multimodal Corpora. LNCS (LNAI), vol. 5509, pp. 138–159. Springer, Heidelberg (2009)
20. MacWhinney, B., Wagner, J.: Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. Gespraechsforschung 11, 154–173 (2010)
21. Navarretta, C.: Annotating Behaviours in Informal Interactions. In: Esposito, A. (ed.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 309–315. Springer, Heidelberg (2011)
22. Navarretta, C.: Anaphora and gestures in multimodal communication. In: Hendrickx, B., Devi, L., Mitkov (eds.) Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Edicoes Colibri, Faro, Portugal, pp. 171–181 (2011)
23. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The mumin coding scheme for the annotation of feedback, turn management and sequencing. Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation 41(3-4), 273–287 (2007)
24. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer (2013), retrieved from `http://www.praat.org/`
25. Kipp, M.: Gesture generation by imitation - from human behavior to computer character animation. Ph.D. thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com (2004)

26. Peirce, C.S.: Collected Papers of Charles Sanders Peirce, 8 vols. Harvard University Press, Cambridge (1931, 1958)
27. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
28. Campbell, N.: An audio-visual approach to measuring discourse synchrony in multimodal conversation data. In: Proceedings of Interspeech 2009, pp. 12–14 (2009)
29. Jokinen, K., Pärkson, S.: Synchrony and copying in conversational interactions. In: Paggio, P., Ahlsén, E., Allwood, J., Jokinen, K., Navarretta, C. (eds.) Proceedings of the 3rd Nordic Symposium on Multimodal Communication. NEALT Proceedings Series, vol. 15, pp. 18–24 (2011)

Building Rapport between Human and ECA: A Pilot Study

David Novick and Iván Gris

Department of Computer Science, The University of Texas at El Paso,
500 West University Avenue, El Paso, TX 79968-0518 USA
novick@utep.edu, ivangris4@gmail.com

Abstract. This study is part of a longer-term project to provide embodied conversational agents (ECAs) with behaviors that enable them to build and maintain rapport with their human partners. We focus on paralinguistic behaviors, and especially nonverbal behaviors, and their role in communicating rapport. Using an ECA that guides its players through a speech-controlled game, we attempt to measure the familiarity built between humans and ECAs across several interactions based on paralinguistic behaviors. In particular, we studied the effect of differences in the amplitude of nonverbal behaviors by an ECA interacting with a human across two conversational sessions. Our results suggest that increasing amplitude of nonverbal paralinguistic behaviors may lead to an increased perception of physical connectedness between humans and ECAs.

Keywords: Embodied conversational agent, familiarity, rapport, paralinguistic, nonverbal communication.

1 Introduction

An embodied conversational agent is a computer program that produces an intelligent agent that lives in a virtual environment and communicates through an elaborate user interface. Graphically, an embodied agent can take almost any form, often human-like, and aims to unite gesture, facial expression and speech to enable face-to-face communication with users, providing a powerful means of human-computer interaction (Cassell, 2000). We are interested in exploring how ECAs can build rapport with humans.

Face-to-face conversation is an ongoing collaborative process in which conversants coordinate their verbal and paralinguistic actions (Cassell et al., 2007). However, human-agent communication cannot yet achieve the naturalistic and spontaneous communication that humans do unconsciously; familiarity-enabled ECAs are a step towards a more naturalistic human-agent conversation. By increasing an ECA's extraversion as the relationship progresses across time, participants should experience a stronger sense of physical connection. Thus the question we address in this paper is how ECAs can build rapport with humans through behaviors linked to familiarity—the sense of knowing someone built in more than one conversation.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 472–480, 2014.

In this paper we describe several definitions and measures of rapport and familiarity. We then classify and merge these into a more comprehensive model. We test this model by having users play a variation of a speech enabled text-based game. During the game, the paralinguistic behaviors of the ECA can change over time to simulate an increase in familiarity between ECA and human. Finally, we analyze the results of the experiment, review the study's limitations, and discuss the future work.

2 Rapport Models

Previous research has identified multiple constituent factors for rapport, including positivity, attention and coordination (Tickle-Degnan & Rosenthal, 1987), and sense of connection, sense of understanding, and what and how things are said (Gratch et al., 2007). In this section, we analyze multiple approaches to rapport with a view toward creating a unified, comprehensive model that can then be implemented in an ECA as a way to test the model.

Prior research on interaction between humans and embodied conversational agents (ECAs) has studied differences in ECAs' nonverbal behaviors as expressions of extraversion or attention. Some studies were based on analysis of a single recording of a human-ECA interaction rather than through a between-subjects comparison of responses to differences in ECA behaviors (e.g., Neff et al., 2010; Huang, Morency & Gratch, 2011). Other studies, though they did use a between-subjects design, looked at rapport-building behaviors through single-session experiments. That is, they compared subjects' responses across conditions based on a single encounter with the ECA (e.g., Cafaro et al., 2012). Several studies have examined how ECAs and humans build rapport over time. In these studies, however, the agents used a multiple-choice text interface and sprite-based characters, which limited nonverbal interactions and dialog flow, particularly during turn taking (e.g., Bickmore & Cassell, 2001; Bickmore & Picard, 2004). Ideally, the way ECAs interact with humans would change as a function of prior interactions between individuals, analogous to how humans interact differently with friends than with strangers.

2.1 Natural Rapport Model

Tickle-Degnen and Rosenthal (1987) described rapport in terms of three dimensions:

- *Attentiveness:* The conversants focus is directed toward the other. They experience a sense of mutual interest in what the other is saying or doing.
- *Positivity:* The conversants feel mutual friendliness and caring.
- *Coordination:* Balance and harmony, where the conversants are "in sync." In addition to its positive valence, coordination conveys an impression of equilibrium, regularity and predictability between the conversants.

This model assumes that positivity becomes less necessary over time, while coordination increases in frequency and importance. This is one of the simpler yet robust models of rapport, and it was not developed with ECAs in mind. Due to its high level of abstraction, the model does not specifically address many of the major nonverbal rapport-building behaviors within each dimension, and these paralinguistic behaviors are particularly important when implementing ECAs.

2.2 Relational Models

Other researchers provided different approaches to modeling rapport, although these other approaches are not, by themselves, as comprehensive as that of Tickle-Degnen and Rosenthal. We look at four relational models proposed by different researchers, whose findings have not yet been unified. Each of these relations by itself only explains a part of rapport as an overall relationship. Moreover, these interaction traits relay heavily on context and verbal disclosure, which can be difficulty to implement in ECAs. These four relational models are

- *Affinity:* The process through which people try to induce others to have positive feelings towards them; this has also been described as a sense of connection (Bell & Daly, 1984).
- *Reciprocity:* A preference of similarity, often expressed as the Golden Rule: One should treat others as one would like others to treat oneself (Cole & Teboul, 2004).
- *Intimacy:* An interpersonal process, where a person expresses personally revealing feelings or information to another. The process continues when the listener responds supportively and empathically. For an interaction to become intimate, the discloser must feel understood, validated, and cared for (Reis & Shaver, 1988).
- *Continuity:* A progressive pattern of interactions, where each conversation ends with the possibility of continuing the interaction at a later time (Fisher & Drecksel (1983).

Each of these relations can be viewed through a lens of nonverbal behaviors based on extraversion across time, creating animations and expressions on behalf of the agents that are easily observable and controlled (e.g., increased) across time to create the familiarity effect.

2.3 Virtual Rapport Model

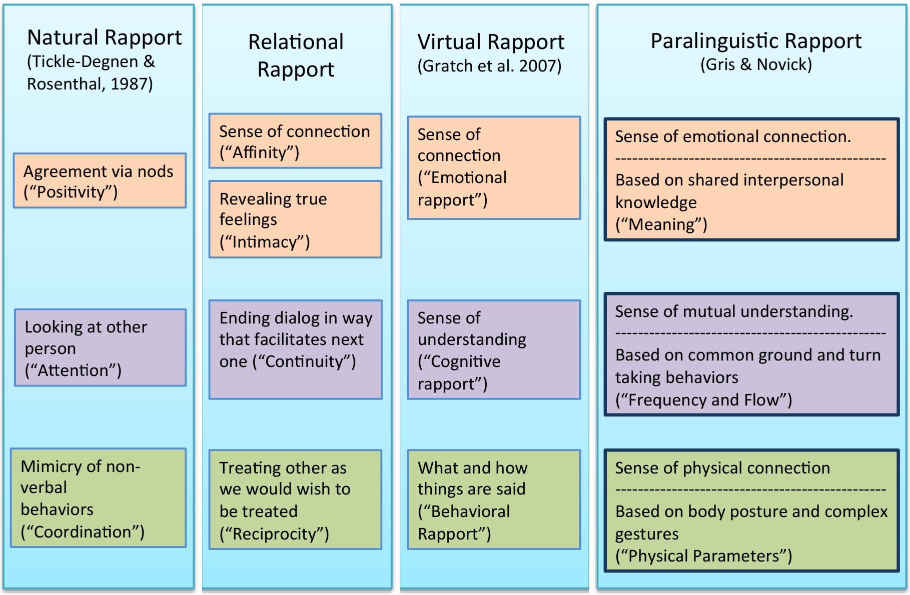
Gratch et al. (2007) proposed a model of rapport specifically for ECAs. Indeed, this model defines virtual rapport as rapport generated for human-ECA interactions. In this model, rapport comprises three dimensions:

- *Emotional Rapport:* The sense of connection with the user
- *Cognitive Rapport:* The sense of mutual understanding
- *Behavioral Rapport*: Verbal properties, such as speech duration, pitch, etc

This model, while clearly more useful for implementing ECAs, does not provide details on some of nonverbal behaviors that trigger these dimensions of rapport, especially full-body gesture and interaction.

2.4 Paralinguistic Rapport Model

With a view toward providing a model of rapport that (a) accounts for the factors identified in the natural, relational, and virtual models and (b) provides a basis for supporting full-body ECA paralinguistics, we examined the common elements of the three approaches. We suggest that these elements can be described as encompassing three dimensions: a sense of emotional connection, a sense of mutual understanding, and a sense of physical connection. Figure 1 presents our "paralinguistic" model of rapport, showing the correspondence of the model's dimensions to the dimensions or factors of the antecedent models. Two of the dimensions in the paralinguistic model—emotional connection and mutual understanding—arise from a combination of verbal and nonverbal behaviors. However, the third dimension—physical connection—arises solely from paralinguistic behaviors. Our model is broader than that of Tickle-Degnen and Rosenthal with respect to physical behaviors, in that the physical collaboration and cooperation of familiar conversants can go beyond mimicry to include many other kinds of paralinguistic behaviors, such as ways of expressing continuers and ways of indicating attitudes.



Natural Rapport (Tickle-Degnen & Rosenthal, 1987)	Relational Rapport	Virtual Rapport (Gratch et al. 2007)	Paralinguistic Rapport (Gris & Novick)
Agreement via nods ("Positivity")	Sense of connection ("Affinity") Revealing true feelings ("Intimacy")	Sense of connection ("Emotional rapport")	Sense of emotional connection. -- Based on shared interpersonal knowledge ("Meaning")
Looking at other person ("Attention")	Ending dialog in way that facilitates next one ("Continuity")	Sense of understanding ("Cognitive rapport")	Sense of mutual understanding. -- Based on common ground and turn taking behaviors ("Frequency and Flow")
Mimicry of non-verbal behaviors ("Coordination")	Treating other as we would wish to be treated ("Reciprocity")	What and how things are said ("Behavioral Rapport")	Sense of physical connection -- Based on body posture and complex gestures ("Physical Parameters")

Fig. 1. Paralinguistic rapport model and its relation to the natural rapport mode, the relational rapport model (affinity: Bell & Daly, 1984; reciprocity Cole & Teboul, 2004; intimacy: Reis & Shaver, 1988; and continuity: Fisher & Drecksel, 1983), and the virtual rapport model

Given our paralinguistic rapport model, our longer-term goal involves assessing the model's validity and usefulness for implementing ECAs that can build rapport with humans. As a first step toward this goal, we focus on the rapport-building effects of physical paralinguistic behaviors. The harmony and engagement of rapport can be seen as relatively weak during initial interactions and developing strength over time;

we refer to this development across time as familiarity. Conversants signal increased familiarity by, among other things, increasing the amplitude of nonverbal communicative behaviors such as hand gestures and head nods (Neff et al., 2010; Cafaro et al., 2012; Clausen-Bruun, Ek, & Haake, 2013). In other words, the ECA's gestures and their degree of extraversion, as expressed through greater amplitude, can build the physical-connection dimension of rapport over time.

Thus our specific question in this study is whether subjects interacting with an ECA over two sessions, where the ECA uses higher-amplitude gestures in the second session, would feel an increase in rapport in the second session.

3 Methodology

To test whether subjects would perceive more rapport with an ECA in the increased-familiarity condition, this study piloted an investigation of how to signal increased familiarity over repeated interactions as a component of rapport. In particular, we studied the effect of differences in the amplitude of nonverbal behaviors by an ECA interacting with a human across two conversational sessions. In the first session, the ECA used nonverbal behaviors with a lower-amplitude baseline. Our independent variable was whether, in the second session, the ECA used same baseline amplitude nonverbals, indicating no increase in familiarity, or used higher-amplitude nonverbals to convey an increase in familiarity.

Our experimental protocol had 20 subjects interact for a 20-minute session with an ECA and then interact for a second 20-minute session with the ECA at least one day later. The sessions involved a conversation with a life-sized, front-projected ECA in UTEP's Immersion Laboratory, where the ECA served as the narrator for an adventure game developed specifically for this study.

The game, "Escape from the Castle of the Vampire King," was inspired by early text-based adventure games such as Zork (Anderson & Galley, 1985) and Colossal Cave. Subjects can move from room to room, pick up, drop, and use objects, and kill vampires. We chose this application because such games are known to be engaging, and we wanted our human subjects to want to interact with the ECA. A text-based game was also helpful from a practical standpoint, in that it limited the amount of speech that needed to be recognized; the subject's possible utterances were both simple and highly constrained by the game's context. Based on our own experiences with text-based games, we instructed subjects to draw a map as they explored.

A trial run of the experiment strongly suggested that asking subjects to draw a map as they explored the castle meant that the subjects usually had their gaze directed at the map rather than at the ECA. This was problematic for our experiment because if subjects kept their gaze focused on their map they would not be watching the ECA and thus would not see the paralinguistic behaviors we were changing as the independent variable. As a result, we modified the game so that the map was drawn automatically on the wall behind the ECA as the subject explored the castle. We observed that subjects now directed their gaze toward the ECA to a much greater extent. Figure 2 shows a person interacting with the ECA. Figure 3 shows the game mid-way through a typical session. The game was extensive enough to support easily the two 20-minute sessions.

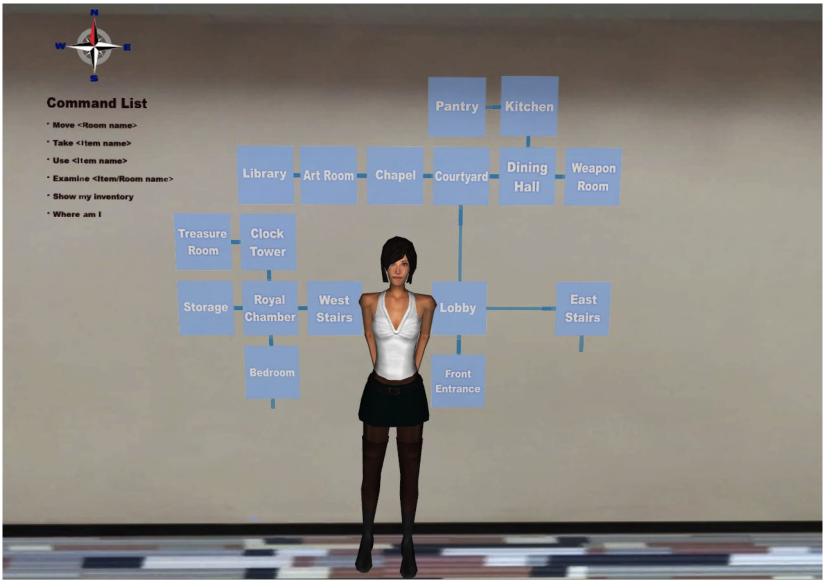


Fig. 2. "Escape from the Castle of the Vampire King" game, with the game at about the mid-point of play

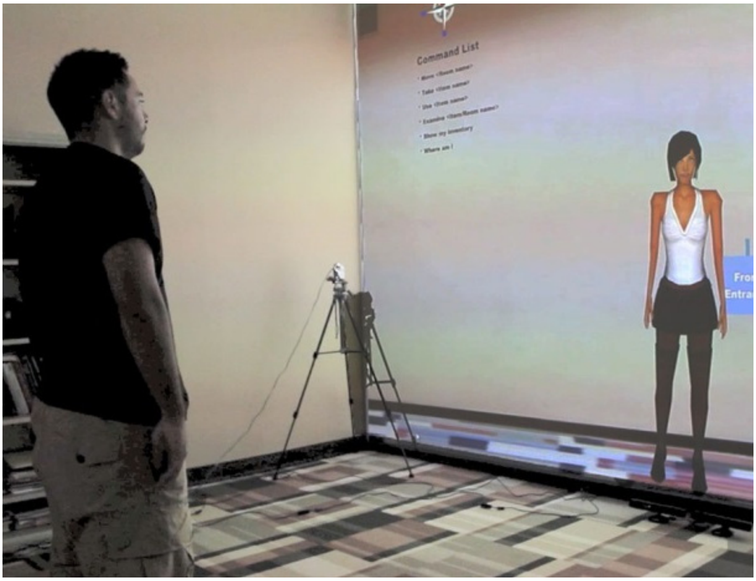
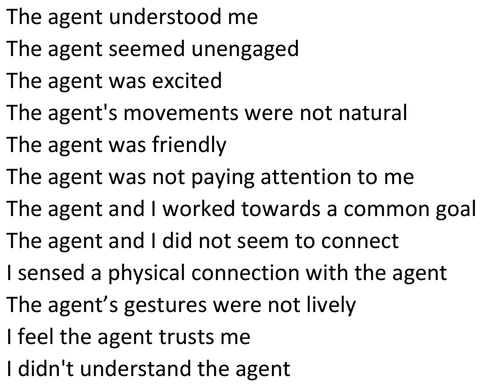


Fig. 3. Interaction between a human and the ECA during testing of the game

In the subjects' first session, the ECA used the baseline amplitude for nonverbals. In the second session, the subjects were randomly assigned to one of two conditions:

(1) the ECA continued to use the baseline nonverbal or (2) the ECA used nonverbal with increased amplitude. The subjects completed a rapport instrument after each session. We adapted and extended the survey of Acosta and Ward (2011) into an instrument of twelve Likert-scale questions, balanced for positive and negative responses that covered the three rapport factors in our model. Table 1 lists the questions in the rapport instrument.

Table 1. Rapport instrument. Subjects indicated agreement or disagreement on a five-point scale.



The agent understood me
The agent seemed unengaged
The agent was excited
The agent's movements were not natural
The agent was friendly
The agent was not paying attention to me
The agent and I worked towards a common goal
The agent and I did not seem to connect
I sensed a physical connection with the agent
The agent’s gestures were not lively
I feel the agent trusts me
I didn't understand the agent

4 Results

Our analysis compared the second-session responses across the two conditions. A one-tailed t-test indicated that there was no significant main effect (p=0.37). Similarly, there was no significant effect for each of the three rapport factors composing the instrument: emotional connection (p=0.40), sense of mutual understanding (p=0.29), physical connection (p=0.17). However, a post-hoc power analysis suggests that conducting the study with 60 subjects would likely produce a significant result for the physical-connection factor.

These results suggest that increasing the amplitude of nonverbal paralinguistic behaviors may not by itself be sufficient to induce a perception of increased rapport throughout all three major dimensions. Nevertheless, the results also suggest that increasing the amplitude of nonverbal paralinguistic behaviors may lead to increased perception of physical connectedness between humans and ECAs. We suspect that the low emotional connection observed in our study was likely due to the lack of emotion in the speech synthesizer; varied emotion would have increased engagement for critical parts of the game, such as the player fighting a vampire or dying. In our follow-on work, we are building a new game with greater interactivity that uses recorded rather than synthesized speech.

An open question at the end of the survey asked subjects to comment on their overall perception of the agent. This question was not taken into account for the analysis of the results, but it provided some interesting information. Some subjects went as far as believing that the agent was a vampire trying to kill them, and they expected a major plot twist near the end of the game. This is arguably a sense of mutual understanding on the part of the player, even if it was unintended and a complete misconception. Another reason for this perception may be a mismatch between the ECA's nonverbal and verbal behaviors, due in large part to the speech synthesizer. That is, when the player encountered a vampire, the ECA would physically show exaggerated expressions alerting the player to the presence of the vampire in the room while verbally explaining the situation in a calm, synthesized voice.

5 Conclusion

Based on our observation of the 20 second-session human-ECA interactions discussed in Section 4, we now return to the question that motivated our study: Do subjects perceive the agent in the increased-familiarity condition in the second session as having higher rapport? Although subjects did notice a behavioral change, and they scored the agent higher on the physical connection dimension, there was not enough information to affirm that this by itself leads to an overall higher sense of rapport.

Our future work will address many of the limitations in this study. In particular, we plan to have a pool of 60 rather than 20 subjects, a recorded human voice rather than a synthesized voice, a more immersive game experience to maintain the participants' attention for a longer period of time, and visual aids to accompany the verbal descriptions of the game state that the agent describes. In the longer term, we plan to address the additional two dimensions of rapport by expressing them in terms of perceivable nonverbal behaviors, to truly create a greater sense of rapport that grows as a function of additional interaction.

Acknowledgments. The authors thank Guillaume Adoneth and David Manuel for their contributions to the design of this study, Jonathan Daggerhart for permission to adapt his original text-based adventure game into "Escape from the Castle of the Vampire King," and Diego Rivera, Alex Rayon, Adriana Camacho, Baltazar Santaella, Juan Vicario, Joel Quintana and Anuar Jauregui for their help in developing the game.

References

1. Acosta, J.C., Ward, N.G.: Achieving rapport with turn-by-turn, user-responsive emotional coloring. Speech Communication 53(9), 1137–1148 (2011)
2. Anderson, T., Galley, S.: The history of Zork. The New Zork Times 4(1-3) (1985)
3. Bell, R.A., Daly, J.A.: The affinity-seeking function of communication. Communications Monographs 51(2), 91–115 (1984)
4. Bickmore, T., Cassell, J.: Relational agents: A model and implementation of building user trust. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 396–403. ACM (March 2001)

5. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction (TOCHI) 12(2), 293–327 (2005)
6. Cafaro, A., Vilhjálmsson, H.H., Bickmore, T., Heylen, D., Jóhannsdóttir, K.R., Valgarðsson, G.S.: First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 67–80. Springer, Heidelberg (2012)
7. Cassell, J. (ed.): Embodied conversational agents. The MIT Press (2000)
8. Cassell, J., Gill, A.J., Tepper, P.A.: Coordination in conversation and rapport. In: Proceedings of the Workshop on Embodied Language Processing, pp. 41–50. Association for Computational Linguistics (June 2007)
9. Clausen-Bruun, M., Ek, T., Haake, M.: Size certainly matters–at least if you are a gesticulating digital character: The impact of gesture amplitude on addressees' information uptake. In: Intelligent Virtual Agents, pp. 446–447. Springer, Heidelberg (2013)
10. Cole, T., Teboul, B.: Non-zero-sum collaboration, reciprocity, and the preference for similarity: Developing an adaptive model of close relational functioning. Personal Relationships 11(2), 135–160 (2004)
11. Crowther, W., Woods, D., Black, K.: Colossal cave adventure. Computer Game (1976)
12. Fisher, B.A., Drecksel, G.L.: A cyclical model of developing relationships: A study of relational control interaction. Communications Monographs 50(1), 66–78 (1983)
13. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)
14. Huang, L., Morency, L.-P., Gratch, J.: Virtual rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
15. Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 222–235. Springer, Heidelberg (2010)
16. Reis, H.T., Shaver, P.: Intimacy as an interpersonal process. In: Duck, S.W. (ed.) Handbook of Personal Relationships, pp. 367–389. John Wiley, NY (1988)
17. Tickle-Degnen, L., Rosenthal, R.: Group rapport and nonverbal behavior. Review of Personality and Social Psychology 9, 113–136 (1987)

The Effect of Voice Instruction on the Construction of Mental Model

Restyandito[1], Alan H.S. Chan[1], and Umi Proboyekti[2]

[1] Dept. of Systems Eng. & Engineering Mgt., City University of Hong Kong,
Hong Kong SAR
rrestyand2@mslive.cityu.edu.hk, alan.chan@cityu.edu.hk
[2] Dept. of Information Technology, Duta Wacana Christian Univ.,
Yogyakarta, Indonesia
othie@ukdw.ac.id

Abstract. The goal of this study is to observe the effect of instruction deliverance method in the construction of mental model. A good mental model can help the user's learnability process. There were two methods tested in this study: a step-by-step instruction (SS) and a complete set of whole-steps instructions (WS) to finish a given task. The SS group performed better on the learning process, however they had the least score on both the information retention and transfer process. Their minds were not engaged in the process, as they seemed to simply follow the instructions without being critical. When error occurred, they tended to be less persistent in trying to finish the task. This might be caused by the incomplete mental model as a result of receiving the instruction step by step.

Keywords: voice instruction, mental model, learnability.

1 Introduction

The use of speech technology as an interface to access the internet has been increasing in the past decades [1-4]. It furnishes new chances to enhance the accessibility by providing compensation for limitations of specific user groups. [5]. Speech is also often used in a multimodal interface to augment communication between human and machine. In their user study, Hofmann et al. found that users are willing to use and trust a speech dialog system [2]. Speech as an input method also has advantages such as simpler, faster and more convenient to use than other methods [6]. However, Boufardea et al. [7] pointed out speech interface can never be perfect. There are some factors that can influence the effectiveness of spoken language interface such as acoustics, speaking style, out-of-vocabulary words and understanding gaps as well as technological limitation [5].

The goal of this study is to study the effect of given voice instruction in the construction of mental model. Mental model is defined as the knowledge that the user has about how a system works, its component parts, the processes, their interrelations, and how one component influences another [8]. It is useful for learning, information

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 481–491, 2014.

retrieving and problem solving [9-11]. This study will focus on the use of mental model in the learnability process. Jay Forrester commented "*A mental model changes with time and even during the flow of a single conversation*" (as cited in [12]). Therefore the way an instruction is delivered will have different impacts on how users build and construct their mental models, which will also influence their user experience and effectiveness of the interaction in the learning process.

2 Previous Study

2.1 Auditory Interface

The continued development of technologies and their application open up demands and opportunities toward multimodal interface. Many studies have been conducted on the effectiveness of multimodal interfaces [5],[6],[13]. A study by Rigas et al. [14] suggested that interfaces could be designed in a way that visual metaphors communicate the information that 'needs' to be conveyed to the user and the auditory metaphors (earcons) communicate the other part of information (the interaction part) which is used to perform tasks. However, the use of auditory representations must consider certain factors. Speech is language dependent and often too slow, pure tones are easily confused with each other, while musical instrumentation though easier to listen to it needs learning and abstraction because of the intuitive mapping [15]. It was found that the use of combinations of auditory icons, earcons, speech, and special sound effects helped user to make fewer mistakes in accomplishing their tasks, and in some cases reduced the time taken to complete them [13]. An experiment by Sodnik et al. [16] showed that auditory interfaces were effective but were not faster than the visual interface to use in a mobile environment. Auditory interfaces can be beneficial to reduce the cognitive workload when visual perception is needed to attend to other task, such as driving, or operating machine. Burke et al. [17] conducted a meta-analysis on the effects of multimodal feedback on user performance. They found visual and auditory cue provide advantages in reducing reaction times and improving performance scores, but it does not reduce error rates effectively. Furthermore it was found that visual-auditory feedback was most effective in single task scenarios, because the use of both auditory and visual channels will increased user's workload hence less advantageous in situations where high workload condition are already present. This study will focus only on the auditory interface in the form of voice instruction.

2.2 Mental Model in Learning

Visual imagery has been known as one of the techniques to improve human memory [18]. Greek and Roman orators use this technic to keep track of the many parts of their long speeches. They visualize an imaginary place and as they walk through this place in their mind, they attach objects associated to the parts of their long speeches. When they need to recite their speech, they just need to visualize and walk through that place in their mind [19].

Several studies on the role of mental model in learning to operate a device showed that people who used mental model learned the procedure faster, retained it more accurately and executed them faster [8],[10],[20]. Other studies on the impact of visualization on the learning process have also been conducted. The work of Scwamborn et al. [21] indicated positive main effects of learner-generated pictures on drawing and mental effort on comprehension measurement. Students may learn better from text and pictures than from text alone, because pictures increase appropriate active processing during learning while reducing non relevant cognitive processing. Leutner et al. [23] conducted an experiment to 10th grade students, where they were asked to mentally imagine text content while reading an expository science text. They found that mental imagery increases text comprehension, even though visualization strategies could cause high demands of cognitive load on the learner [22]. They also pointed out "*Decreased cognitive load due to constructing mental images has, in terms of a main effect, no direct impact on reading comprehension and thus, on learning*".

Phillips et al. [11] stated "*Instructions are a common example of communicating models of technological systems and can act as a boundary object between designers' conceptual models and models developed by users*". When delivering the instruction, it is significant to keep the wording of signs and instructions as simple and short as possible [5]. The instruction must be able to communicate a clear decision pathway [24].

3 Method

3.1 Participants

Seventy five first year university Indonesian students (60 male, 15 female) participated in this study. Their mean age was 18.74 years (SD=1.64). All participants received monetary expense allowance for their time. They were familiar with internet and have some experience in using search engine. Table 1 provides a summary of the participants' background.

Table 1. Respondent background experience

		n	%
Gender	Male	60	80.00
	Female	15	20.00
Time spent on the internet daily	< 5 hours	28	37.33
	5-10 hours	32	42.67
	> 10 hours	15	20.00
Using search engine to find information	Always	64	85.33
	Sometimes	11	14.67
	Never	0	0
Using attribute in search engine	Often	9	12.00
	Sometimes	55	73.33
	Never	11	14.67
Using logic operator in search engine	Often	5	6.67
	Sometimes	36	48.00
	Never	34	45.33

3.2 Materials

The experiment was conducted in a university laboratory with intranet access to the student project report repository website. The software used to capture the screen movement was Snagit ver 11.2.0.101. The browser used to access the repository website was Firefox ver 26.0. The computer specification for this experiment was Intel Core 2 CPU 4300 @ 1.80 GHz, with 2014 MB of RAM.

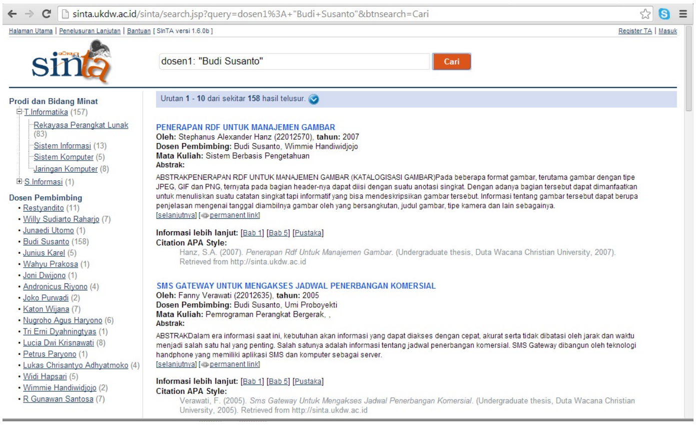
3.3 Design

There are two ways of giving instructions, namely explicit demand and implicit demand [2]. This study will investigate two approaches of delivering instructions: a short one step at a time instruction, and a long instruction explaining the whole series of steps. Participants were randomly assigned to one of the three treatment conditions with 25 participants in each group. The first group is the control group (CG) which received no instruction, the second group received a short step by step instruction (SS), and the last group received long instructions describing the whole set of steps (WS). To minimize technological limitation which may cause ineffectiveness in user's perception as pointed out by Neerincx et al. [5], the instruction was delivered live by a person using decent sound system instead of the speech synthesis and prerecorded messages. The instruction given was clear and using direct vocabulary by take into consideration what the participants knows [24]. All participants were familiar in using the internet and should have some experience of using the search engine. For this between-participant experiment the participants were asked to find some information from the final year project report repository web site of undergraduate students at Duta Wacana Christian University.

Before the experiment, participants received a brief introduction about the experiment which was to study the effect of mental model in the learnability process. The participants from SS and WS were asked to construct mental model in their mind as they listened to the instruction. Participants were assured that it is ok if they could not complete the task and the result would be anonymous. They also received short training on how to use Snagit to record their activities during the experiment.

Figure 1 shows a snapshot of the final year project report repository website. Participants can search information by typing the keyword directly on the search bar ("Budi Susanto"), or refine their search in several ways, such as using attribute (dosen1: "Budi Susanto"), click on the categorical result on the left side (grouped by Department, research topic, and supervisor), or scroll manually using scroll bar.

There were four tasks in the experiment. The first task was to find out how many projects were supervised by *Gloria Virginia* on *Genetic Algorithm*. This task can be completed first by typing the attribute (*supervisor1*) and the keyword (*Gloria Virginia*) on the search bar, and further refine the result with the title of the project (*Genetic Algorithm*). The first task is the learning process in which participants received instructions on how to retrieve that information using attribute. The control group (CG) had to ascertain information by trial and error. The second task was to find supervisors who supervised projects on *Multi Criteria* by *Ferry Himawan*. This time, no instructions



sinta.ukdw.ac.id/sinta/search.jsp?query=dosen1%3A+"Budi+Susanto"&btnsearch=Cari

Halaman Utama | Penelusuran Lanjutan | Bantuan [SInTA versi 1.6.0b] Register TA | Masuk

dosen1: "Budi Susanto" Cari

Prodi dan Bidang Minat

- T.Informatika (157)
  - Rekayasa Perangkat Lunak (83)
  - Sistem Informasi (13)
  - Sistem Komputer (5)
  - Jaringan Komputer (8)
- S.Informasi (1)

Dosen Pembimbing

- Restyandito (11)
- Willy Sudiarto Raharjo (7)
- Junaedi Utomo (1)
- Budi Susanto (158)
- Junius Karel (5)
- Wahyu Prakosa (1)
- Joni Dwijono (1)
- Andronicus Riyono (4)
- Joko Purwadi (2)
- Katon Wijana (7)
- Nugroho Agus Haryono (6)
- Tri Erni Dyahningtyas (1)
- Lucia Dwi Krisnawati (8)
- Petrus Paryono (1)
- Lukas Chrisantyo Adhyatmoko (4)
- Widi Hapsari (5)
- Wimmie Handiwidjojo (2)
- R Gunawan Santosa (7)

Urutan **1** - **10** dari sekitar **158** hasil telusur.

PENERAPAN RDF UNTUK MANAJEMEN GAMBAR
Oleh: Stephanus Alexander Hanz (22012570), **tahun:** 2007
Dosen Pembimbing: Budi Susanto, Wimmie Handiwidjojo
Mata Kuliah: Sistem Berbasis Pengetahuan
Abstrak:

ABSTRAKPENERAPAN RDF UNTUK MANAJEMEN GAMBAR (KATALOGISASI GAMBAR)Pada beberapa format gambar, terutama gambar dengan tipe JPEG, GIF dan PNG, ternyata pada bagian header-nya dapat diisi dengan suatu anotasi singkat. Dengan adanya bagian tersebut dapat dimanfaatkan untuk menuliskan suatu catatan singkat tapi informatif yang bisa mendeskripsikan gambar tersebut. Informasi tentang gambar tersebut dapat berupa penjelasan mengenai tanggal diambilnya gambar oleh yang bersangkutan, judul gambar, tipe kamera dan lain sebagainya.
[selanjutnya] [permanent link]

Informasi lebih lanjut: [Bab 1] [Bab 5] [Pustaka]
Citation APA Style:
Hanz, S.A. (2007). *Penerapan Rdf Untuk Manajemen Gambar.* (Undergraduate thesis, Duta Wacana Christian University, 2007). Retrieved from http://sinta.ukdw.ac.id

SMS GATEWAY UNTUK MENGAKSES JADWAL PENERBANGAN KOMERSIAL
Oleh: Fanny Verawati (22012635), **tahun:** 2005
Dosen Pembimbing: Budi Susanto, Umi Proboyekti
Mata Kuliah: Pemrograman Perangkat Bergerak, ,
Abstrak:

ABSTRAKDalam era informasi saat ini, kebutuhan akan informasi yang dapat diakses dengan cepat, akurat serta tidak dibatasi oleh jarak dan waktu menjadi salah satu hal yang penting. Salah satunya adalah informasi tentang jadwal penerbangan komersial. SMS Gateway dibangun oleh teknologi handphone yang memiliki aplikasi SMS dan komputer sebagai server.
[selanjutnya] [permanent link]

Informasi lebih lanjut: [Bab 1] [Bab 5] [Pustaka]
Citation APA Style:
Verawati, F. (2005). *Sms Gateway Untuk Mengakses Jadwal Penerbangan Komersial.* (Undergraduate thesis, Duta Wacana Christian University, 2005). Retrieved from http://sinta.ukdw.ac.id

Fig. 1. Sinta, the final year project report repository website

were given to all of the groups. Participants were expected to rely on their previous experience to finish the task. This process is known as the retention process. Participant could retrieve the information using the same step they did on the first task, but with different attribute (*title*) and keyword (*Multi Criteria*) and further refined the result by the student's name (*Ferry Himawan*). For the third task, participants were asked to find information on the abstract of projects supervised by *Budi Susanto* on the topic of *resource description framework*. The purpose of this task was to learn the usage of logic operator. Participants could retrieve the information by using 'AND' operator in addition to search attribute. Similar to the first task, they received instructions on how to do it except that the control group received no instruction at all. The last task was aimed to examine the transfer process of using previous knowledge and experience to carry out task in new condition different from the previous tasks. They did not receive any instruction. Participants were asked to find a project based on incomplete information. They were only given the first name of the supervisor (*Umi*), they did not have information whether *Umi* is the first supervisor or the second supervisor, and they only knew the topic of the project instead of the title. Hence they needed to make use of the operator and attribute in slightly different ways. They had to use three attributes (*supervisor1, supervisor2, topic*) and both operators (OR and AND) in the search bar. The purpose of the second and fourth task is to discover how mental model help participants in completing similar task and new task. For each task, the number of errors (NE), success rate (SR), and completion time (CT) were recorded. The performance of each group will be compared.

4 Results and Discussion

Table 2 presents the summary result of the experiment. A one way between subjects ANOVA was conducted to compare the effect of instruction delivery on success rate, average number of errors and average completion time. There was a significant effect on instruction delivery methods on success rate only on Task 1 [F(2,72)=10.031, p = 0.0001]. There was a significant effect on instruction delivery methods on completion time, on Task 1 [F(2,72)=4.227, p = 0.019] and Task 3 [F(2,72)=3.597, p = 0.034]. There was also a significant effect on number of error, on Task 1 [F(2,72)=4.777, p = 0.012] and Task 3 [F(2,72)=7.064, p = 0.002]. However, there was no significant effect on instruction delivery methods on success rate, completion time and number of error on Task 2 and Task 4.

Table 2. Summary of participant's performance for each task

T1	CG	SS	WS	**T2**	CG	SS	WS
Success Rate (SR)	4	20	10	Success Rate (SR)	19	14	20
Avg. Num. of Error (NE)	1.88	2.02	2.72	Avg. Num. of Error (NE)	2.84	2.18	2.72
Avg.Compl. Time (CT)	288s	212s	253s	Avg.Compl. Time (CT)	156s	184s	156s

T3	CG	SS	WS	**T4**	CG	SS	WS
Success Rate (SR)	19	21	21	Success Rate (SR)	3	1	6
Avg. Num. of Error (NE)	0.80	2.06	2.52	Avg. Num. of Error (NE)	4.80	4.56	4.24
Avg.Compl. Time (CT)	146s	192s	173s	Avg.Compl. Time (CT)	276s	293s	286s

Task 2 was meant to see the retention process based on the experience of completing Task 1. The results here show the group which received Step-by-Step instruction (SS) had the lowest performance, while the best performance was achieved by the group which received Whole-Steps instructions (WS). The same performance can be seen on the transfer process in Task 4, where SS had the lowest SR and WS achieved highest SR. The SS participants might not be able to build the mental model of the overall system when they received the step by step instructions, while the WS participants had to construct the mental model as they received the series of instructions before executing them. Hence WS participants had a more complete mental model which they can explore to finish a task. A complete series of instructions enable participants to visualize a decision pathway in their mind [24]. However, this observation must be tested further, as the ANOVA analysis for this experiment did not show statistical difference.

As expected, there is an increase in the success rate of Task 3 compared to Task 1. This might be the result of the learnability effect [9] from the previous two tasks. Both Control Group (CG) and WS show a meaningful improvement on SR, 475% and 210% respectively. While for SS, improvement in SR was noted from one participant only (105%). Based on the observation of recorded screen activities, there were some factors contributing to the failure of following instructions such as spelling error, preconception based on past experience or knowledge, interpretation and familiarity. An example of preconception is when the participants were asked to type *title: “genetic algorithm”*, one of the participants was typing *title= “genetic algorithm”*, which might be influenced by his/her experience using query language. Other examples of mistakes caused by interpretation are when participants were asked to type in the search bar *title: “multicriteria” AND supervisor: ”budi susanto”,* there were several interpretations of the instruction, one of participant typed *title: "multicriteria" & supervisor: "budi susanto"*, other participant typed *title: "multicriteria" N supervisor: "budi susanto"*. There was a participant who did not type the query in the search bar of the site, instead he typed it in the search bar of the browser. An example of mistakes caused by familiarity was when the participants were asked to search a final project supervised by Mrs.Umi of the Management Department, there were some participants who automatically typed Mrs.Umi Proboyekti who is a lecturer at the Information Technology Department where the participants came from and familiar with. Neerinx et al. [5] recommended the use of instruction as short as possible, however this might lead to ambiguity. In this example punctuation mark matters, as well as case sensitive. The instruction given, did not explicitly inform participants how to type the searching query, resulting some of them made mistakes in the first try. Yet, giving very detail instruction might not be effective either, as the instruction will be long, and human's memory ability is limited [18]. Too much detail instruction may cause participants to forget what they need to do [25]. These examples show the challenges of making a clear instruction as there are many factors that can influence users' comprehension [7].

Step-by-Step instruction did not seem to stimulate users to think actively as they listen to the instruction; hence they just typed whatever they thought they had heard. Furthermore, when they encountered an error, they were less likely to try to solve the problem. As seen in Table 2, they had less number of errors compared to the WS participants, even though they did not perform better on completing the task. This could indicate their lack of determination to finish the task. They might conclude that the instruction given to them was wrong.

The performance of all groups was worst for Task 4. This might be caused by the fact that participants needed to transfer their knowledge to complete a new task. Had they more exposure using the system, they might perform better. Supportive information is essential in the process of acquiring cognitive skills, but so is practice [26], whether those skills are recurrent (performed the same way on each incidence) or non-recurrent (performed differently according to conditions managed by complex rules or contextual features) [27].

It can also be seen in Table 2 that the high number of errors does not necessarily mean low success rate. For example, even though number of errors for the CG and WS groups were higher than SS, it turned out they also had higher success rate. Users were versatile and got adapted to the system operation easily. If they ran into errors when using an application, they would naturally try to solve them, by adapting, improvising or negotiating [6]. Most likely participants would have to rely on their mental model to predict the step needed to complete the task. Marchionini [28] found that the efficiency of mental model building depends on the level of detail transferred. Table 3 shows the summary of the regression analysis based on the instruction delivery method and past knowledge or experience in using the search engine. The R Square value is 6%-23% indicating there might be more variables influencing participants' performances, such as age, gender, intelligence level, etc.

Table 3. Summary of regression analysis

SUCCES RATE (SR)

	T1		T2		T3		T4	
	Coeff.	*P-val*	*Coeff.*	*P-val*	*Coeff.*	*P-val*	*Coeff.*	*P-val*
Step by Step (SS)	0.276	0.070	-0.172	0.235	-0.024	0.870	-0.107	0.232
Whole Steps (WS)	0.050	0.746	0.153	0.301	0.123	0.417	0.065	0.482
Search Engine	0.221	0.268	0.080	0.675	0.011	0.953	-0.123	0.299
Attribute	-0.055	0.631	0.046	0.679	-0.148	0.192	0.025	0.715
Logic	-0.084	0.498	0.215	0.073	0.177	0.149	-0.089	0.229
	R Square = 0.139		R Square = 0.166		R Square = 0.070		R Square = 0.089	
	R = 0.372		R = 0.407		R = 0.264		R = 0.298	

AVERAGE COMPLETION TIME (CT)

	T1		T2		T3		T4	
	Coeff.	*P-value*	*Coeff.*	*P-value*	*Coeff.*	*P-value*	*Coeff.*	*P-value*
Step by Step (SS)	-16.094	0.041	26.284	0.245	45.376	0.016	12.18	0.164
Whole Steps (WS)	12.945	0.107	-6.347	0.783	22.792	0.229	4.769	0.593
Search Engine	-9.186	0.371	-7.271	0.807	18.243	0.454	21.07	0.071
Attribute	7.802	0.191	-13.383	0.438	-8.309	0.555	-1.201	0.857
Logic	1.981	0.756	-7.199	0.698	-13.75	0.366	1.238	0.863
	R Square = 0.238		R Square = 0.067		R Square = 0.150		R Square = 0.097	
	R = 0.487		R = 0.258		R = 0.380		R = 0.311	

AVERAGE NUMBER OF ERRORS (NE)

	T1		T2		T3		T4	
	Coeff.	*P-value*	*Coeff.*	*P-value*	*Coeff.*	*P-value*	*Coeff.*	*P-value*
Step by Step (SS)	1.273	0.032	-0.309	0.702	2.510	0.001	0.023	0.976
Whole Steps (WS)	1.582	0.010	-0.242	0.771	1.708	0.018	-0.955	0.227
Search Engine	0.730	0.346	-0.557	0.603	-0.431	0.636	-0.977	0.337
Attribute	0.234	0.600	-0.767	0.218	0.073	0.890	0.656	0.266
Logic	0.400	0.406	0.086	0.897	0.259	0.648	-0.171	0.786
	R Square = 0.178		R Square = 0.042		R Square = 0.207		R Square = 0.082	
	R = 0.422		R = 0.206*		R = 0.455		R = 0.286	

The minimum correlation coefficient R for 70 subjects and more with 95% confidence level is at least 0.232. There is only one instance where R is < 0.232, on the average number of errors (NE) of Task 2. Consequently, it can be concluded that these variables contribute to the performance of the participants. The SR of Task 1 was affected mostly by the instruction delivery method (SS) as it has the highest coefficient of 0.276, while Task 2 and Task 3 were affected mostly by previous experience using Logic (0.215 and 0.177 respectively) while Task 4 was affected mostly by the instruction method (WS) with coefficient of 0.065. Prior experience can help participants to understand the system better. They could construct a mental model based on their experience in using search engine, and use the same reasoning to complete the task. However, having prior experience may lead them not to pay attention to the instruction. There were three participants who used the advance search facility, even though they were instructed to type manually the attribute and keywords on the search bar.

The participants of this study were mostly first year students, while the tasks given were related to finding final year project information usually needed by final year students. Therefore, they might not be really determined to complete the task as if they were final year students. However they were purposely invited to participate in this study, because as first year students, they might not have any prior experience in using the repository system.

5 Conclusion

This study has tried to look at the effect of given voice instruction in the construction of mental model, especially during the retention and transfer process. One way ANOVA analysis has showed that instruction delivery method plays an important role in participants' performance during the learning process. There was evidence that whole steps instruction enables participants to perform better during the retention and transfer process, although the result was not statistically significant.

Further study can be conducted with more participants and task related to participants' interest or need. The tasks given to the participants can be concretized using the Bollywood Method, where a task will be given a dramatic and exaggerated backstory to excite the participants into believing the urgency of the problem [29]. The study can also consider more variables for inclusion in the experiment.

Based on the findings, some recommendation can be made in the interface design using voice instruction.

- Instruction should be clear and precise to avoid ambiguity.
- Instruction given should fit user knowledge and experience
- Step-by-Step instruction ensure higher success rate in the learning process
- Instruction should not be too long, so user will not forget the instruction

Acknowledgment. The first author would like to thank the Directorate General of Higher Education of Indonesia for supporting his study at City University of Hong Kong.

References

1. Lin, D., Bigin, L., Bao-zong, Y.: Using chinese spoken-language access to the WWW. In: 5th International Conference on Signal Processing, Proceedings of the WCCC-ICSP 2000, vol. 2, pp. 1321–1324 (2000)
2. Hofmann, H., Ehrlich, U., Berton, A., Minker, W.: Speech interaction with the internet - A user study. In: Proc - Int. Conf. Intelligent Environ., pp. 323–326 (2012)
3. Sazbon, M., Haddad, Y.: Advanced vocal web browser. In: IEEE Conv. Electr. Electron Eng. Israel, pp. 732–735 (2010)
4. Werner, S., Wolff, M., Eichner, M., Hoffmann, R.: Integrating speech enabled services in a web-based e-learning environment. In: Internat. Conf. Inf. Tech. Coding Comput., pp. 303–307 (2004)
5. Neerincx, M.A., Cremers, A.H.M., Kessens, J.M., van Leeuwen, D.A., Truong, K.P.: Attuning speech-enabled interfaces to user and context for inclusive design: Technology, methodology and practice. Univers Access Inf. Soc. 8(2), 109–122 (2009)
6. Knudsen, L.E., Holone, H.: A multimodal approach to accessible web content on smartphones. In: Miesenberger, K., Karshmer, A., Penaz, P., Zagler, W. (eds.) ICCHP 2012, Part II. LNCS, vol. 7383, pp. 1–8. Springer, Heidelberg (2012)
7. Boufardea, E., Garofalakis, J., Plessas, A.: A dynamic voice portal for delivery of cultural content. In: Proc. - Int. Conf. Internet Web Appl. Serv., pp. 186–191 (2008)
8. Fein, R.M., Olson, G.M., Olson, J.S.: A mental model can help with learning to operate a complex device, pp. 157–158. ACM (1993)
9. Rouse, W.B., Morris, N.M.: On looking into the black box: Prospects and limits in the search for mental models. Psychol. Bull. 100(3), 349 (1986)
10. Kieras, D.E., Bovair, S.: The role of a mental model in learning to operate a device. Cogn Sci. 8(3), 255–273 (1984)
11. Phillips, R., Lockton, D., Baurley, S., Silve, S.: Making instructions for others: Exploring mental models through a simple exercise. Interactions 20(5), 74–79 (2013)
12. Smith, R.: Game impact theory: The five forces that are driving the adoption of game technologies within multiple established industries. Games and Society Yearbook (2007)
13. Rigas, D.I., Alsuraihi, M.: A toolkit for multimodal interface design: An empirical investigation. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 196–205. Springer, Heidelberg (2007)
14. Rigas, D., Yu, H., Klearhou, K., Mistry, S.: Designing information systems with audio-visual synergy: Empirical results of browsing E-mail data, pp. 960-7620 (2001)
15. Wersényi, G.: Auditory representations of a graphical user interface for a better human-computer interaction. In: Ystad, S., Aramaki, M., Kronland-Martinet, R., Jensen, K. (eds.) CMMR/ICAD 2009. LNCS, vol. 5954, pp. 80–102. Springer, Heidelberg (2010)
16. Sodnik, J., Dicke, C., Tomažic, S., Billinghurst, M.: A user study of auditory versus visual interfaces for use while driving. Int. J. Hum. Comput. Stud. 66(5), 318–332 (2008)
17. Burke, J.L., Prewett, M.S., Gray, A.A., Yang, L., Stilson, F.R.B., Coovert, M.D., Elliot, L.R., Redden, E.: Comparing the effects of visual-auditory and visual-tactile feedback on user performance: A meta-analysis, pp. 108–117 (2008)
18. Baddeley, A.D.: Your memory: A user's guide. Sidgwick & Jackson, London (1982)
19. Carpenter, S., Huffman, K.: Visualizing psychology. Wiley. com (2009)
20. Ujita, H., Yokota, T., Tanikawa, N., Mutoh, K.: Computer aided instruction systems for plant operators. Int. J. Hum. Comput. Stud. 45(4), 397–412 (1996)

21. Schwamborn, A., Thillmann, H., Opfermann, M., Leutner, D.: Cognitive load and instructionally supported learning with provided and learner-generated visualizations. Comput. Hum. Behav. 27(1), 89–93 (2011)
22. Ainsworth, S.: The functions of multiple representations. Comput. Educ. 33(2-3), 131–152 (1999)
23. Leutner, D., Leopold, C., Sumfleth, E.: Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. Comput. Hum. Behav. 25(2), 284–289 (2009)
24. Harvey, G.: Designing procedural instructions: 5 key components. Inf. Des. J. 16(1), 19–24 (2008)
25. Deo, S., Nichols, D.M., Cunningham, S.J., Witten, I.H., Trujillo, M.F.: Digital library access for illiterate users (2004)
26. Darabi, A.A., Sikorski, E.G., Nelson, D.W., Palanki, S.: Efficient, motivational, and effective strategies for complex learning: Computer-based simulation experiments in troubleshooting. Technology Instruction Cognition and Learning 3(3/4), 233 (2006)
27. Darabi, A.A., Nelson, D.W., Seel, N.M.: The role of supportive information in the development and progression of mental models. In: Learning and Instruction in the Digital Age, pp. 101–115. Springer (2010)
28. Marchionini, G.: Making the transition from print to electronic encyclopaedias: Adaptation of mental models. Int. J. Man Mach. Stud. 30(6), 591–618 (1989)
29. Sherwani, J., Palijo, S., Mirza, S., Ahmed, T., Ali, N., Rosenfeld, R.: Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In: Int. Conf. Inf. Commun. Technol. Development, pp. 447–457 (2009)

Discourse Particles and User Characteristics in Naturalistic Human-Computer Interaction

Ingo Siegert[1], Matthias Haase[2], Dmytro Prylipko[1], and Andreas Wendemuth[1]

[1] Institute for Information Technology and Communications,
Otto von Guericke University Magdeburg, Germany
{firstname.lastname}@ovgu.de

[2] Department of Psychosomatic Medicine and Psychotherapy,
Otto von Guericke University Magdeburg
{firstname.lastname}@med.ovgu.de

Abstract. In human-human interaction (HHI) the behaviour of the speaker is amongst others characterised by semantic and prosodic cues. These short feedback signals minimally communicate certain dialogue functions such as attention, understanding or other attitudinal reactions. Human-computer interaction (HCI) systems have failed to note and respond to these details so far, resulting in users trying to cope with and adapt to the machines behaviour. In order to enhance HCI, an adaptation to the user's behaviour, individual skills, and the integration of a general human behaviour understanding is indispensable. Another issue is the question if the usage of feedback signals is influenced by the user's individuality. In this paper, we investigate the influence of specific feedback signals, known as discourse particles (DPs), with communication style and psychological characteristics within a naturalistic HCI. This investigation showed that there is a significant difference in the usage of DPs for users of certain user characteristics.

Keywords: human-machine-interaction, discourse particles, personality, user characteristics.

1 Introduction

Verbal human to human communication consists of several information layers, going beyond the pure textual information and transmitting relevant information such as self-revelation, relationship and appeal [31]. These details are normally provided by humans to enhance human-human interaction (HHI) and to increase the likelihood of a positive interaction outcome. Human-computer interaction (HCI) systems have failed to note and respond to these details so far, resulting in users trying to cope with and adapt to the machines behaviour [25]. This adaptation of the user leads to the typical machine-like interaction patterns resulting in a loss of information and lowering the chance of a successful HCI. To obtain a more human-like and more successful interaction with technical systems, those have to be adaptable to the users' individual skills, preferences, and user characteristics. This includes

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 492–501, 2014.

both, the ability to understand the user's capabilities and a proper reaction towards him [35].

In HHI the behaviour of the speaker is characterised by semantic and prosodic cues, given as short feedback signals. These so-called discourse particles (DPs) e.g. "hm" or "uhm" minimally communicate certain dialogue functions such as attention, understanding, or other attitudinal reactions. Thus, these signals play an important role in the progress and coordination of the interaction. They allow the conversational partners to inform each other of their behavioural or affective state without interrupting the ongoing dialogue. As a further advantage, these feedback signals can be easily inferred from the speaker's intonation, which is in the case of DPs not influenced by semantic and grammatical information [27].

Two previous studies investigated necessary prerequisites. The first study investigated the occurrence of DPs within HCI and the relation between DPs and predefined pitch contours [29]. Furthermore, the DPs served as features for complex emotion detection [28]. More information about the meaning of DPs can be found in [5,27]. Our previous work investigated the correlation of DP-usage with different age and gender groups. Thereby, we revealed that the variations within the different groups are quite substantial. This indicates that there must be other factors influencing the individual use of DPs. This paper now investigates the correlation of DP-usage and specific psychological characteristics of the subjects within a naturalistic HCI.

1.1 Discourse Particles in HCI

During HHI several semantic and prosodic cues are exchanged among the interaction partners and used to signalize the progress of the dialogue [1]. The intonation of feedback signals transmits the communicative relation of the speakers and their attitude towards the current dialogue. The occurrence of different intonation-meaning relations are depending on the conversation type. In conversations of narrative or cooperative character confirmation signals are dominating, whereas turn holding signals dominate argumentative conversations [24].

As intonation is influenced by semantic and grammatical information, it is advisable to investigate the intonation of so-called DPs [1]. These speech fragments cannot be inflected, but emphasised. The incorporation of DPs in HCI systems will allow a detection of crucial points within the dialogue and help to initiate proper system reactions. Furthermore DPs are uttered in situations of a higher cognitive load [5].

As DPs have a specific function within the conversation (indicate thinking, conformation or request to respond, cf. [27]), the use of these particles requires the conversational partners to understand the meaning. Hence, it may be assumable that DPs do not occur in HCI. The investigation in [10] showed that while the number of partner-oriented signals are decreasing during HCI, the number of signals indicating a task-oriented, or expressive function are increasing. These findings could be confirmed with our previous study, cf. [29].

The so far presented studies demonstrated that DPs are used within HCI [10] and also tried to explain the broad variety of occurrences between different users [29]. The utilized distribution in young vs. elder users and male vs. female speakers revealed that elderly female speakers using DPs twice as often than elderly male speakers. But the mean variation within the different groups is still quite large. Thus, we assume that other factors influence the use of DPs.

1.2 User Characteristics in HCI

Research on communication and personality dispositions has a distinguished history. Today, it is agreed that personality is a rather complex entity containing different aspects. Thus, many user characteristics are discussed having an influence on the interaction towards technical system. Among others, these variables cover personality traits (attributional style, anxiety, problem solving), which are important for the user's behaviour in both HHI and HCI [8].

In personality psychology and psychological research the "Big Five" factors of human personality were widely confirmed and represent the most influential personality model nowadays [18,23]. Furthermore, the "Big Five"-model had a great impact on research about a certain sequence in natural communication: the initial dyad. Initial dyadic interaction refers to the first contact between humans, i.e., the situation in which two people get to know each other for the first time. A lot of researchers report on strong relations between factors of personality and the communication with another person in this certain situation. In contrast to the "Big Five" model, other theories of personality focuses more on interpersonal relationships. The author in [30] opposed his inter-psychic model to predominant intra-psychic models of personality .

Personality plays an important role in HCI, too (e.g., [7,12]). Former research identified personality traits as well as interpersonal relationship as relevant aspects in the field of HCI [33]. Summarising, there is some evidence suggesting that Extroversion is related to computer aptitude and achievement [32].

In addition, also socio-demographic aspects as age and gender, or affinity to information and communications technology (ICT) are discussed to play an important role [13,19,21]. In the case of ICT-aspects especially the knowledge and skills as well as the anxiety in dealing with technical systems, the user's problem-solving behavior, and thus the whole work style is seen to have an impact [2,3,4]. Furthermore, the user's domain knowledge, and language skills are pointed out in this context [22]. Until now, however, only a few empirical studies investigate the impact of user properties to interaction with a technical system, cf. [22].

2 Dataset

The conducted study utilizes the LAST MINUTE corpus (cf. [25]) as naturalistic HCI database that is already object of examination regarding affective state recognition [11] and linguistic turns [26]. The utilized corpus contains 133 multimodal recordings of German speaking subjects during Wizard-of-Oz (WOZ) experiments. The setup revolves around a journey to the unknown place "Waiuku", which the subjects have won. Each experiment takes about 30 minutes. Using voice commands, the subjects have to prepare the journey, pack the suitcase, and select clothing. Most of the experiments are transliterated, enabling the automatic extraction of speaker utterances. Details can be found in [25].

The experiment is distinguished into two modules, with two different dialogue styles: personalisation and problem solving module [25]. The personalisation module, being the first part of the experiment, has the purpose of making the user familiar with the system and to make his behaviour more natural. In this introduction (IN) the users are encouraged to talk freely. We furthermore located the same dialogue style at the end of each experiment, when the system asks further questions about the satisfaction with the user's solution and denote this as closure (CL).

During the problem solving module the user is expected to pack the suitcase for his journey. The dialogue follows a specific structure of specific user-action and system-confirmation dialogues. This conversation is task focused and the subjects talk more command-like. Thus this part or the experiment has a much more regularized dialogue style. The sequence of these repetitive dialogues is interrupted by pre-defined barriers (Bx) for all users at specific time points. These barriers are intended to increase the stress level of the users.

B1 the task is introduced, no details about commands and target location
B2 the user gets familiar with the system, first excitement gone
B3 the content of the current suitcase is listed verbally
B4 the system refuses to pack items because the weight limit is reached
B5 details about the target location are given
B6 user can repack items but with time pressure

In addition to the WOZ experiment itself, socio-biography and psychometric parameters are collected using validated questionnaires. Psychological questionnaires are established methods for the collection of specific variables. They can thus be used, to determine social and political characteristics, opinions, interests, or psychological characteristics such as personality factors, attributional style, motivation, and many different constructs.

The NEO-FFI [6] is designed to assess the constellation of traits defined by the Five Factor theory of personality. The model assumes that behaviour in situations (state) is influenced by steady characters (traits). The "Big Five" factors are extroversion, agreeableness, conscientiousness, neuroticism, and openness.

Another questionnaire utilizes Sullivan's model of personality and focuses on interpersonal relationships. The inventory of interpersonal problems (IIP) [14] is a model for conceptualizing, organizing, and assessing interpersonal behaviour, traits, and motives. Eight scales mark the interpersonal circumplex by selecting items (domineering, vindictive, cold, socially avoidant, nonassertive, exploitable, overly nurturant and intrusive). As the experiment is conducted with German speaking subjects, the German version is used, cf. [15].

The stress-coping questionnaire (SVF) [17] includes 20 scales (e.g. deviation, self-affirmation, control of reaction) for different types of response to an unspecific selection of situations that impair, adversely affect, irritate, or disturb the emotional stability or balance of the subject.

Additionally to this psychometric instruments socio-demographic variables like age, gender, educational level, experience with computers (e.g. years overall, hours per day/week), and in what context the subjects use the computer are collected. This corpus is designed to have an equal distribution of gender and age of the subjects. The younger group ranges from 18-28 years. the elder group consists of subjects being over 60 years.

3 Results

We used a subset of 89 subjects with a total duration of approx. 45 hours. The group distribution of age and gender is as follows: 21 young male and 23 young female subjects and 19 old male and 27 old female subjects. As the experiment is transliterated, we conducted an automatic alignment with a manual correction phase for the DP-extraction. Within our subset of 89 subjects, only 3 subjects do not utter any DP. The overall number of DPs is 1975, the mean is 28.77 particles per conversation with a standard deviation of 25.15. One subject uses 107 particles in an experiment, which is the maximum. To analyse the DP-usage, we set the DPs in relation to the total number of user's acoustic utterances of any kind like words, see Fig. 1. As statistical test, we use a one-way ANOVA, to compare means of our two mean-splitted samples, cf. [16].

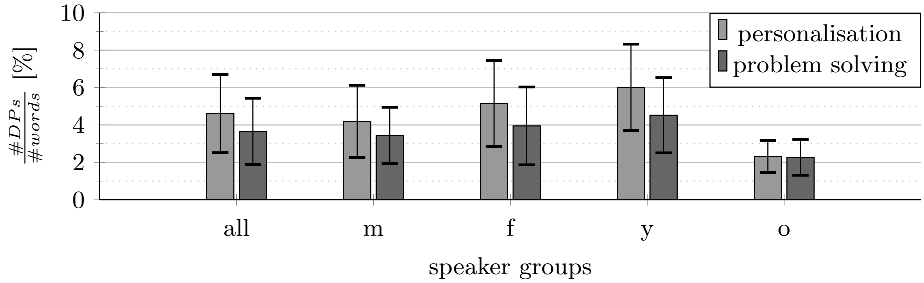


Fig. 1. Mean and standard deviation for the DPs divided into the two dialogue styles regarding different speaker groups in the case of gender (male, female) and age (young and old). For comparison the group independent frequency (all) is given, too.

We further notice, that the usage of DPs is not equally distributed among the gender and the age of the subjects, see Fig. 1. This difference is largely determined by the speaker's age. The difference between the young and old speakers is significant for both personalization ($p < 0.002$) and problem solving ($p < 0.027$). This means that young and old users do not only different by their age, but also in relation to the type of communication (personalization or problem solving). Hereby the group differences are from special interest, while the usage of DPs for elderly does not reveal big differences in the both phases, young users on the other hand have distinct differences between personalisation and problem solving. Regarding the other groups, only substantial differences can be noticed, this may be mostly due to the small sampling size.

From this investigations, it can be seen that the standard deviation is quite high. This indicates a high individuality of the users' DP-usage and we assume that additional criteria, as specific psychological characteristics, are inferring the usage of DPs. Therefore, we further analyse the DP-usage depending on specific user characteristics. Hereby, we again set the DPs in relation to the total number of user's acoustic utterances. of user's acoustic utterances. We furthermore differentiate between user traitsbelow the mean (low trait) and those at or above the mean (high trait). As statistical test, we use a one-way ANOVA, to compare means of our two mean-splitted samples, cf. [16]. The results can be found in Fig. 2. We only depict results with provide substantial results nearly the significance.

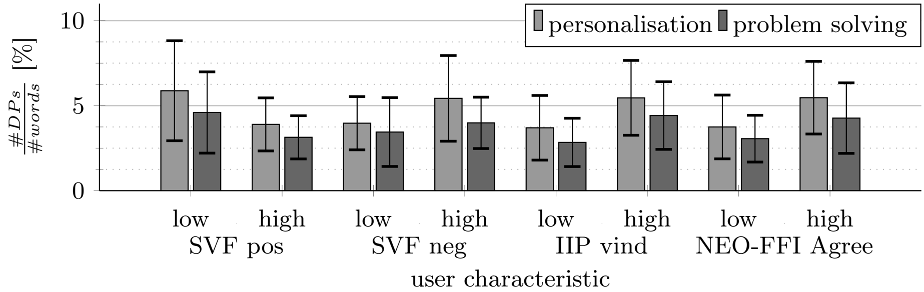


Fig. 2. Mean and standard deviation for the DPs divided into the two dialogue styles regarding different groups of user characteristics

Considering the psychological characteristics, no significant differences are noticeable on the distinction between the two dialogue styles personalisation and problem solving. This is mostly due the fact, that we compare very few users within a very heterogeneous sample.

As the influence of psychological characteristic heavily depends on the situation in which the user is located. The distinction in a free dialogue and regulated dialogue may not be sufficient to describe the user's situation. Especially in the regulated problem solving module very different situations are induced by the experimental design, which also produce partly contradictory user reactions. But to make at least substantial statements, the number of samples is not sufficient, as stated before.

For interpreting the SVF positive strategies distraction (SVF pos), we could state that subjects having better skills in stress management with regard to positive distraction use substantial less DPs. Especially in the personalization they showed less DPs. The finding on SVF negative strategies (SVF neg) confirms the previous one. Subjects who do not have a good stress management and unlike even have negative stress management mechanisms (i.e. stress management mechanisms increasing the stress) also use more DPs.

Evaluating the IIP personality trait vindictive competing (IIP vind), we can state that subjects using DPs more frequently, Volunteers, more likely to have problems trusting others or rather towards others are suspicious and are rather quarrelsome showed more DPs.

Also the interpretation of the NEO-FFI confirms the IIP-findings because the subjects having less DPs show less confidence in dealing with other people.

Thus, it can be assumed that the usage of DPs is accompanied by “negative” psychological characteristics. This supports the findings that DPs are uttered in situations of a higher cognitive load [5].

4 Discussion

Th presented investigation on the use of specific back-channel signals in HCI and their correlation with psychological characteristics allows us to investigate HCI from a new perspective. First, the verified use of DPs in HCI prove the assumption that HCI and HHI are comparable, which has long been presupposed for investigating HCI, cf. [9,34]. Our investigation furthermore indicates that humans tend to use mechanisms from HHI they are familiar with also when interacting with technical systems, although they are aware that these systems do not have the same capabilities than human conversational partners [20].

The precise analysis of DP-occurrence within the dialogue styles reveals that the use of DPs is more likely when the subject is encouraged to talk freely than during structured dialogues. Furthermore, the age of the speakers influenced the usage of DPs, when taking the verbalisation into account. IN our analysis young and old users do not only different by their age, but also in relation to the type of communication (personalization or problem solving). This could be interpreted that young users are more confident when using a machine-like interaction than elderly users. Anyway, young users seem to be familiar with this kind of conversation.

Other factors that influence the usage of DPs are the user’s psychological characteristics. Hereby our investigations reveal that the usage of DPs corresponds with specific psychological characteristics that describe the user’s interpersonal relationship, attributional style, and technological affinity.

Our investigations reveal that the occurrences of DPs could provide hints of specific psychological characteristics in pre-known situations of the interaction. Especially in situations of a higher cognitive load [5], when the user is not able to deal with this “negativity”.Thus, if these characteristics are already known, than the usage of DPs can be seen as stress indicators, which have to be taken into account for an appropriate reaction of the system.

For appropriate reactions, the system should also take into account the different communicative functions the DPs have, cf. [28]. This investigation indicates that technical system can be enabled to easily differentiate the DP-intonation of "thinking". In cases where the user utters a DP having this meaning, the system should wait for the user input, in cases of a more competent user in dealing with technical systems. In contrast, only for users do not having this competence, the system should offer explanations.

5 Conclusion

Our investigations show that DPs are also utilized within a HCI, although, the users know that these feedback signals cannot be interpreted by the technical system. However, it should be noted that one can not draw certain conclusions from the purely presence of DPs. The revealed age-bias as well as psychological characteristics have to be taken into account, especially if the interaction makes certain demands on the user, as higher problem solving abilities or specific language skills. It has also been shown that the current situation in which the user with its specific psychological characteristic is located has a significant impact on the use of DPs.

However, it remains to be clarified to what extent such studies can be transferred to other corpora and other situative interactions. This includes an in-depth study of the used DP-functions with respect to the experimental situation and psychological characteristics. Unfortunately, for this purpose the number of DPs in actual material is too low.

Acknowledgement. This work was supported by the Transregional Collaborative Research Centre SFB/TRR 62 (www.sfb-trr-62.de) "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Allwood, J., Nivre, J., Ahlsn, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9(1), 1–26 (1992)
2. Anderson, A.A.: Predictors of computer anxiety and performance in information systems. Computers in Human Behavior 12(1), 61–77 (1996)
3. Beckers, J.J., Rikers, R.M., Schmidt, H.G.: The influence of computer anxiety on experienced computer users while performing complex computer tasks. Computers in Human Behavior 22(3), 456–466 (2006)
4. Ceyhan, E.: Computer anxiety of teacher trainees in the framework of personality variables. Computers in Human Behavior 22(2), 207–220 (2006)
5. Corley, M., Stewart, O.W.: Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*. Language and Linguistics Compass 2, 589–602 (2008)
6. Costa, P., McCrae, R.: NEO-PI-R Professional manual. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI). Psychological Assessment Resources, Odessa (1992)

7. Cuperman, R., Ickes, W.: Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts disagreeables. J. Pers. Soc. Psychol. 97(4), 667–684 (2009)
8. Daily, J.: Personality and Interpersonal Communication. In: Handbook of Interpersonal Communication, pp. 133–180. Sage, Thousand Oaks (2002)
9. Elliot, C.: The application of pragmatics in Human-Computer interaction. Ph.D. thesis, Sheffield Hallam University (1993)
10. Fischer, K., Wrede, B., Brindöpke, C., Johanntokrax, M.: Quantitative und funktionale Analysen von Diskurspartikeln im Computer Talk. International Journal for Language Data Processing 20(1-2), 85–100 (1996)
11. Frommer, J., Michaelis, B., Rsner, D., Wendemuth, A., Friesen, R., Haase, M., Kunze, M., Andrich, R., Lange, J., Panning, A., Siegert, I.: Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus. In: Proc. of the Eight LREC 2012, ELRA, Istanbul (May 2012)
12. Funder, D.C., Sneed, C.D.: Behavioral manifestations of personality: An ecological approach to judgmental accuracy. J. Pers. Soc. Psychol. 64(3), 479–490 (1993)
13. Hermann, F., Niedermann, I., Peissner, M., Henke, K., Naumann, A.: Users interact differently: Towards a usability- oriented user taxonomy. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4550, pp. 812–817. Springer, Heidelberg (2007)
14. Horowitz, L., Alden, L., Wiggins, J., Pincus A.: Inventory of Interpersonal Problems Manual. The Psychological Corporation, Odessa (2000)
15. Horowitz, L., Strau, B., Kordy, H.: Inventar zur Erfassung Interpersonaler Probleme (IIPD) (Inventory of interpersonal problems-German version), 2nd edn. Beltz, Weinheim (2000)
16. Howell, D.: Statistical Methods for Psychology, 7th edn. Cengage Learning (2009)
17. Jahnke, W., Erdmann, G., Kallus, K.: Stressverarbeitungsfragebogen mit SVF 120 und SVF 78, 3rd edn. Hogrefe, Göttingen (2002)
18. John, O., Hampson, S., Goldberg, L.: Is there a basic level of personality description? J. Pers. Soc. Psychol. 60(3), 348–361 (1991)
19. King, J., Bond, T., Blandford, S.: An investigation of computer anxiety by gender and grade. Computers in Human Behavior 18(1), 69–84 (2002)
20. Lange, J., Frommer, J.: Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion (Subjective experience and intentional setting within intervies of User-Companion-Interaction). In: Informatik 2011: Informatik schafft Communities, Beitrge der 41. Jahrestagung der GI. Lecture Notes in Informatics, vol. 192, pp. 240–254 (2011)
21. Lithari, C., Frantzidis, C., Papadelis, C., Vivas, A., Klados, M., Kourtidou-Papadeli, C., Pappas, C., Ioannides, A., Bamidis, P.: Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. Brain Topography 23(1), 27–40 (2010)
22. Naumann, A., Hermann, F., Peissner, M., Henke, K.: Interaktion mit Informations- und Kommunikationstechnologie: Eine Klassifikation von Benutzertypen (Interaction with information and communication technology: A classification of user types). In: Herczeg, M., Kindsmller, M.C. (eds.) Mensch und Computer 2008: Viel Mehr Interaktion, pp. 37–45. Oldenbourg Verlag, München (2008)
23. Ozer, D.J., Benet-Martinez, V.: Personality and the prediction of consequential outcomes. Annu. Rev. Psychol. 57(3), 401–421 (2006)
24. Paschen, H.: Die Funktion der Diskurspartikel HM (The function of discourse particles HM). Master's thesis, University Mainz (1995)

25. Prylipko, D., Rösner, D., Siegert, I., Günther, S., Friesen, R., Haase, M., Vlasenko, B., Wendemuth, A.: Analysis of significant dialog events in realistic human computer interaction. Journal on Multimodal User Interfaces (2013)
26. Rösner, D., Kunze, M., Otto, M., Frommer, J.: Linguistic analyses of the LAST MINUTE corpus. In: Jancsary, J. (ed.) Proceedings of KONVENS 2012, pp. 145–154. oral presentations, Main track (2012)
27. Schmidt, J.E.: Bausteine der Intonation (Components of intonation). In: Neue Wege der Intonationsforschung, Germanistische Linguistik, vol. 157-158, pp. 9–32. Georg Olms Verlag, Hildesheim (2001)
28. Siegert, I., Hartmann, K., Philippou-Hübner, D., Wendemuth, A.: Human Behaviour in HCI: Complex Emotion Detection through Sparse Speech Features. In: Salah, A.A., Hung, H., Aran, O., Gunes, H. (eds.) HBU 2013. LNCS, vol. 8212, pp. 246–257. Springer, Heidelberg (2013)
29. Siegert, I., Prylipko, D., Hartmann, K., Böck, R., Wendemuth, A.: Investigating the form-function-relation of the discourse particle "hm" in a naturalistic human-computer interaction. In: Bassis, S., Esposito, A., Morabito, F.C. (eds.) Recent Advances of Neural Network Models and Applications, Smart Innovation, Systems and Technologies, vol. 26, pp. 387–394. Springer (2014)
30. Sullivan, H.: The interpersonal theory of psychiatry. Norton, New York (1953)
31. von Thun, F.S.: Miteinander reden 1 – Störungen und Klärungen. Allgemeine Psychologie der Kommunikation (Talking to Each Other 1 - faults and clarifications. General Psychology of Communication). Rowohlt, Reinbek (1981)
32. Van der Veer, G.C., Tauber, M.J., Waem, Y., Van Muylwijk, B.: On the interaction between system and user characteristics. Behaviour & Information Technology 4(4), 289–308 (1985)
33. Weinberg, G.M.: The psychology of computer programming. Van Nostrand Reinhold, New York (1971)
34. Weiss, A., Mirnig, N., Buchner, R., Förster, F., Tscheligi, M.: Transferring human-human interaction studies to HRI scenarios in public space. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part II. LNCS, vol. 6947, pp. 230–247. Springer, Heidelberg (2011)
35. Wendemuth, A., Biundo, S.: A Companion Technology for Cognitive Technical Systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)

The Effects of Working Memory Load and Mental Imagery on Metaphoric Meaning Access in Metaphor Comprehension

Xiaofang Sun[1,2], Ye Liu[1,*], and Xiaolan Fu[1]

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China
{sunxf,liuye,fuxl}@psych.ac.cn
[2] Universersty of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Metaphor is a cognitive process that enables people to make mental mapping across distinct conceptual domains. The present study investigated metaphorical and literal meaning access in metaphor comprehension, and the effects of working memory load and mental imagery on metaphor comprehension. Three sentence priming experiments were conducted and the results showed that the literal meaning of a metaphor was accessed faster than the metaphorical meaning, but metaphorical meaning could be accessed as quickly as literal meaning if there was more cognitive resource involved. These findings indicated that the literal meaning of a metaphor is accessed first in the early stage of metaphor comprehension, and working memory load plays an important role in the process. The study didn't find any significant effect of ima-geability on metaphor comprehension; however, the results implied the metaphors with low imageability need more working memory load to understand. The implication for natural language processing of the computer science was discussed.

Keywords: metaphor comprehension, working memory load, mental imagery, imageability.

1 Introduction

Metaphors are very prevalent in everyday spoken and written language. For example, speakers used approximately one unique metaphor for every 25 words in an analysis of television programs [1]. Another study found 20 metaphors were used per 1000 words for college lectures, 50 metaphors in ordinary discourse, and 60 metaphors in discourses by teachers [2]. Metaphors establish correspondences between different concepts from disparate domains of knowledge [1]. For example, the lawyer is a shark. The topic (the first term) of this metaphor refers to a people, and the vehicle (the second term) refers to a fish. An increasing number of studies have revealed that metaphors are essentially involved in not only our communication but also everyday thought. Hence, metaphor is a way of thinking and cognitive style as well as a common rhetorical device in language [1].

[*] Corresponding Author.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 502–510, 2014.

1.1 The Time-Course of Literal Meaning and Metaphorical Meaning Accesses

The prevalence of metaphor in language and thought has motivated a considerable number of researches on cognitive mechanism of metaphor comprehension and computational methods by machine in natural language processing [2]. How do people understand metaphoric expressions and comprehend the meaning of sentences that differ in their literal and nonliteral interpretations?

Recent metaphor research has revealed that there could be two kinds of cognitive mechanisms involved in metaphor comprehension. One is categorization [3] and the other is comparison process [1]. According to the temporal property of metaphor comprehension, these two kinds of mechanisms belong to two different models [4, 5]: (a) the direct model hypothesizing the metaphorical meaning can be accessed directly by class-inclusion without the access of literal meaning, and (b) the indirect model stating that the literal meaning of a metaphor is necessarily accessed first by comparing the vehicle with the topic. Recently, researchers proposed the third view hypothesizing metaphor comprehension is context-dependent, i.e. when there is a relevant context, the metaphorical meaning is the only one accessed [5]. A recent study using event-related potential supported the notion of dual access to metaphorical meaning and literal meaning in metaphor processing and the literal meaning as a subordinate meaning was activated during the early metaphor comprehension stage [6]. Hence, the time-course of literal and metaphorical meaning access is still controversial. Using a semantic prime paradigm, the present study aimed to examine the temporal property of literal and metaphorical meaning access in metaphor comprehension.

1.2 Working Memory Load

The link between working memory and metaphor comprehension has attracted researchers for years. The recent studies agreed the point of view that working memory capacity plays an important role in metaphor processing [7] and the early processing of metaphors is controllable by executive mechanisms [8]. Because fluid intelligence is considerably associated with working memory capacity, researchers believed working memory is important to the interpretation of conventional and creative metaphors [9]. However, there is no research to focus on working memory load in metaphor comprehension so far. Based on the previous research, the present study hypothesized metaphor comprehension would be inhibited under conditions of high working memory load.

1.3 Mental Imagery and Metaphor Comprehension

Since Paivio proposed there were nonverbal imagery and verbal symbolic processes in cognitive system in 1969 [10], there have been a number of researches focusing on the role of mental image in memory, learning, and language. Generally, we use more concrete concepts to express more abstract concepts, for example, the mind is a computer. The topic of this metaphor refers to an abstract entity, and the vehicle refers to a complex electronic device. However, there are few studies on the role of mental image in metaphor comprehension. Gibbs et al. found that people can form coherent mental images for metaphorical actions and many abstract concepts are partly understood in terms of enduring embodied metaphors [11]. Hence, it is interesting to investigate the role of mental imagery in metaphor comprehension.

Based on the previous research, the present study aimed to find out the temporal property of literal and metaphorical meaning access in metaphor comprehension and the effects of working memory load and mental imagery on the processes. We hypothesized that literal meaning of metaphors were accessed first, high working memory load would inhibit the meaning access, and mental imagery would facilitate metaphor comprehension. Three experiments were conducted to test our hypotheses using a classic priming paradigm similar to Blasko and Connine's study [4]. In the three experiments, metaphoric and literal sentences were used as priming stimuli, and two-character words and nonwords were used as targets. There were three kinds of words used as targets in Experiments 1-3: a word related to the metaphorical meaning, i.e. metaphor target (MT); a word related to the literal meaning of the vehicle of the metaphor, i.e. literal word target (LWT); and a control target unrelated to either the metaphorical meaning or literal meaning, i.e. control target (CNT). Subjects were required to make a lexical decision to judge if the target is a true word or not. All the words were responded as correct in the task, and there were nonwords used to responding as fault. If the metaphorical meaning is accessed first, subjects should response faster to MTs than LWTs. We changed the presenting duration of the priming sentence in Experiments 1 and 2 to manipulate working memory load. In Experiment 3, subjects were required to judge if the priming sentences make sense and this task need the most cognitive effort in the three experiments.

2 Experiment 1

2.1 Participants and Stimuli

Participants consisted of a total of 20 undergraduate and graduate students in the experiment. All participants had normal or corrected-to-normal vision, and they were all native speakers of Chinese and unaware of the purpose of the experiment.

All metaphoric sentences used in Experiments 1, 2, and 3 were in the form of X is Y, and all the terms of X and Y were two-character words. The imageability, aptness and familiarity of the sentences were rated by 64 participants, aged 18-25 years. All the sentences were randomly ordered on three scales that ranged from 1 to 7 on imageability (1 = very easy to arouse a mental image and 7 = very hard to arouse a mental image), aptness (1 = not at all apt and 7 = very highly apt) and familiarity (1 = not at all familiar and 7 = very highly familiar). Based on the rating data, 20 metaphors with high imageability and 20 metaphors with low imageability were chosen. The rating scores of these metaphors were varied in imageability with significant difference but comparable in aptness and familiarity (As shown in Figure 1).

The metaphor vehicles in isolation were randomly presented to 72 additional participants to determine the LWTs. They were asked to write down the most common or central feature of the words. Similarly, the MTs were selected by presenting 181 additional participants with a randomized list of the metaphoric sentences and asking them to choose the single word that is the most central feature of the metaphoric meaning of the metaphor.

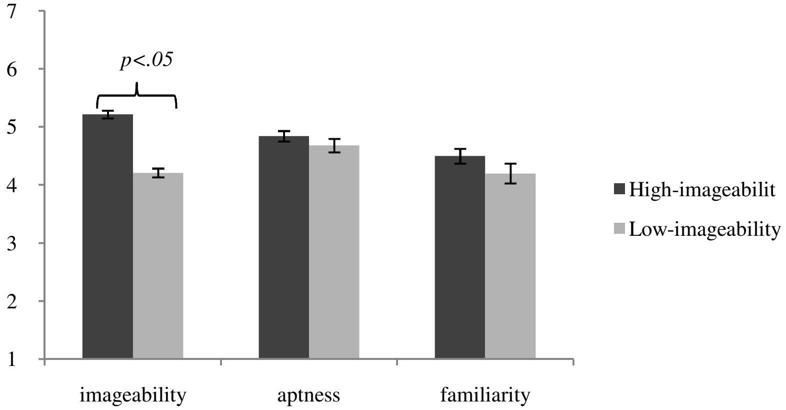


Fig. 1. The mean rating scores with standard errors of means (SEMs) of the metaphors

So that the priming sentences contained 40 metaphoric sentences and 40 filler sentences, all the filler sentences were literal and had the same structure as metaphor. Each subject read all the sentences six times, for a total of 480 trials. Half of the times were with word targets and the other times were with nonword targets.

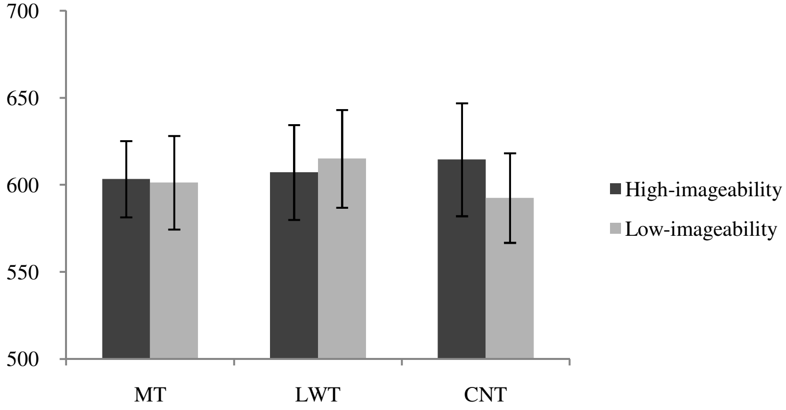


Fig. 2. The mean RTs with SEMs of the lexical decision task without priming sentences by different conditions

An additional lexical decision task was made to ensure that all the targets were comparable without priming sentences by 14 participants. The results of a two-way analysis of variance (ANOVA)—Imageability (high or low) × Target type (literal, figurative, or control)—showed that the groups did not differ significantly from each other ($F < 1$) as shown in Figure 2.

2.2 Procedure

The experiment was programmed with E-Prime 2.0 software. At the beginning of the experiment, participants were provided with an instruction for the experiment. And then they completed a practice set of trials.

Each trial began with a cross in the center of the screen, and it remained for 750 ms, after which the screen went blank for 250 ms. And then the priming sentence appeared in the center of the screen for 750 ms. After the priming sentence disappearing, the target appeared immediately and remained until participants responded. The participants were instructed to indicate whether the target was a word or nonword by pressing the "F" key or "J" key. For half participants, "F" was for true words and "J" was for nonwords, for the other half participants the keys were reversed. Participants were asked to response as quickly and as accurately as possible. The latency between target presentation and the participant's response was recorded as the participant's response time.

2.3 Results and Discussion

All the data were analyzed using SPSS 19.0C (SPSS China). Four participants were excluded from the analyses because their accuracy was low (80.0%, 65.8%, 67.5%, 40.8%). The remaining 16 participants had accuracy of 98.4%, so the accuracy was not analyzed. Reaction times more than 3 standard deviations above or below the mean were excluded from analysis. Totally, 0.9% trials were excluded.

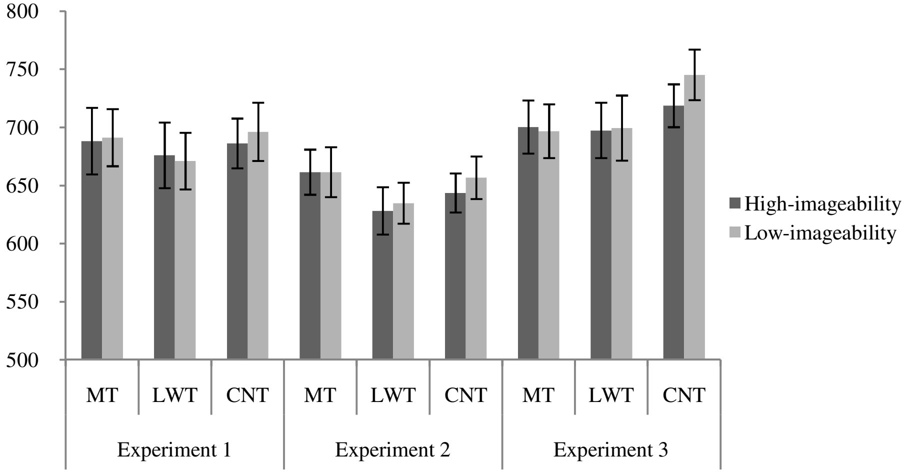


Fig. 3. The mean RTs with SEMs of the lexical decision task by different conditions in experiments 1，2 and 3

The RT data were analyzed using a 2 (imageability: low or high) × 3 (target type: LWT, MT, or CNT) ANOVA (As shown in Figure 3). The results in the present study were considered statistically significant at $p < 0.05$. The analysis revealed a significant main effect of targets type [$F(2, 30) = 3.45$, $p < 0.05$], the responses to LWTs were faster than those to MTs and CNTs. But the main effect of imageability was not significant [$F(1, 15) = 0.71$, $p = 0.09$]. The interaction effect of target type× imageability was not significant [$F(2, 30) = 0.28$, $p = 0.78$].

The results showed that the responses to LWTs were primed by the priming sentences, but the responses to MTs were not primed by priming sentences with a short presenting duration (750 ms). This finding supports the point of view that the literal meaning of a metaphor is accessed first. And, there were no effect of imageability on metaphor comprehension.

3 Experiments 2 and 3

3.1 Participants, Stimuli and Procedure

In Experiment 2, participants consisted of a total of 20 undergraduate and graduate students. In Experiment 3, participants consisted of a total of 32 undergraduate and graduate students. All participants had normal or corrected-to-normal vision, and they were all native speakers of Chinese and unaware of the purpose of the experiment.

In Experiment 2, all the experimental stimuli were the same as those used in Experiment 1. The procedure was also similar to experiment 1. In this experiment, the difference of procedures was that the priming sentences displayed 2000 ms. In Experiment 3, the stimuli included all the metaphors used in Experiment 1 and additional 40 false sentences. The targets were same as those used in experiment 1. In this experiment, when participants were represented the priming sentences, they were asked to judge whether the sentence was true or false, and then they also were asked to do the lexical decision task of targets just like Experiment 1.

3.2 Results and Discussion

In Experiment 2, 20 participants had accuracy of 98.5%. Reaction times more than 3 standard deviations above or below the mean were excluded from analysis. Totally, 3.4% trials were excluded.

The RT data were analyzed using a 2 (imageability: low or high) × 3 (target type: LWT, MT, or CNT) ANOVA. The results of RTs are shown in Figure 3. The analysis revealed a significant main effect of targets types [$F(2, 38) = 14.99$, $p < 0.05$], the responses to LWTs were faster than those to MTs and CNTs. The main effect of imageability was not significant [$F(1, 19) = 3.15$, $p = 0.41$]. The interaction of target × imageability was not significant [$F(2, 38) = 0.47$, $p = 0.63$]. The results were similar with Experiment 1. The responses to LWTs were primed by priming sentences, but the responses to MTs were not primed by the priming sentences even with a much longer duration (2000 ms). No significant effect of imageability was found in the experiment, too.

In Experiment 3, 13 participants were excluded from the analyses because their accuracy of the priming sentence judge task or the target lexical decision task was lower than 50%. The remaining 19 participants had target's accuracy of 99%, so the accuracy was not analyzed. Reaction times of target more than 3 standard deviations above or below the mean were excluded from analysis. The RT data were analyzed using a 2 (imageability: low or high) × 3 (target type: LWT, MT, or CNT) ANOVA. The results of RTs are shown in Figure 3. The analysis with participants revealed a significant main effect of targets types [$F(2, 36) = 3.95$, $p < 0.05$], the responses to LWTs and MTs were faster than those to CNTs. The main effect of imageability was not significant [$F(1, 18) = 1.13$, $p = 0.30$]. The interaction of target × imageability was not significant [$F(2, 36) = 1.34$, $p = 0.28$]. The results showed that both the responses to LWTs and MTs were primed by priming sentences when the prime task need more working memory resource involved.

4 The Results of all the 3 Experiments

Trends in data of different experimental conditions at aggregate level are similar. The RT differences (all the RTs in Experiments 1-3 minus the baseline RTs of the isolated lexical decision experiment) were analyzed using a 2 (imageability: low or high) × 3 (target type: LWT, MT, or CNT) × 3 (experiment condition) ANOVA (As shown in Figure 6).

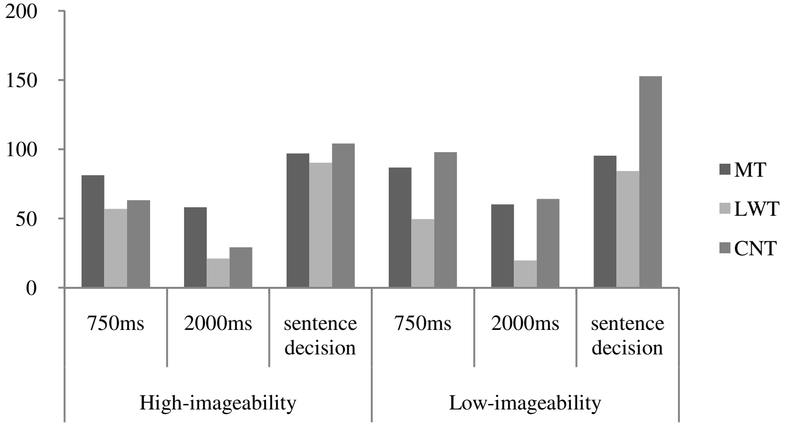


Fig. 4. The mean RTs of three different experimental conditions

The analysis revealed a significant main effect of experimental condition [$F(2, 76) = 95.59$, $p < 0.05$], a significant main effect of target type [$F(2, 76) = 7.74$, $p < 0.05$], a significant target type × imageability two-factor interaction [$F(2, 76) = 4.62$, $p < 0.05$], a significant condition × target type two-factor interaction [$F(4, 152) = 3.82$, $p < 0.05$], but no significant condition × imageability × target type three-factor interaction [$F(4, 152) = 1.04$, $p = 0.39$].

The results showed that the responses in Experiment 2 (the priming sentences displayed 2000 ms) were faster than those in Experiment 1 (the priming sentences displayed 750 ms) and 3 (sentence decision) on the whole. There was no significant difference among the responses to MTs, LWTs, and CNTs with high-imageability priming sentences. The responses to LWTs with low-imageability priming sentences were slightly faster than those to CNTs and MTs with low-imageability priming sentences in the conditions of 750 ms and 2000 ms. However, the responses to LWTs and MTs were faster than those to CNTs with low-imageability priming sentences in the condition of sentence decision task. The contrast of the three experiments indicated that mental imageability of metaphors modulated the effects of working memory load on metaphor comprehension, although it had no significant effects on the cognitive process directly. For the metaphors with low imageability, metaphorical meaning was harder to access than literal meaning when the working memory load was not very high, whereas all of them were hard to access when the working memory load was very high. These results implied the metaphors with low imageability need more working memory load to understand.

5 General Discussion

The results of the present study indicated that the literal meaning of a metaphor was accessed faster than the metaphorical meaning, and metaphorical meaning could be accessed as quickly as literal meaning if there was more cognitive resource involved. These findings support the point of view [3] that the literal meaning of a metaphor is accessed first in the early stage of metaphor comprehension. These results were partly consistent with Blasko and Connine's study [4], in which only the figurative meaning of high-familiar metaphors and low-familiar metaphors with high aptness was available in the priming paradigm. The difference between the present study and Blasko and Connine's results may be due to that the familiarity and aptness of the metaphors used in the present study was moderate. The results of the present study partly supported the theory Bowdle and Gentner proposed [1, 12]. This theory hypothesized that there was a shift in mode of mapping from comparison to categorization as metaphors are conventionalized [1]. According to this theory, the literal meaning of a metaphor is accessed first when it is not familiar, however, the metaphorical meaning of a metaphor would be processed easily when it is very familiar and conventionalized.

The present study also demonstrated that working memory load plays an important role in metaphor comprehension. This finding is consistent with the studies on working memory capacity [7-9] in metaphor comprehension. The present study didn't find a significant effect of mental imageability on metaphor comprehension; however, the results implied the metaphors with low imageability need more working memory load to understand. The results indicated that mental imagery of metaphors modulated the effects of working memory load on metaphorical meaning access, although it had no significant effects on the cognitive process directly. When the working memory load was very high, both of literal meaning and metaphorical meaning was hard to access, whereas literal meaning was easier to access than metaphorical meaning for the metaphors with low imageability when the working memory load was not very high. These results partly demonstrated that mental imagery plays a role in metaphor comprehension; and the mechanism underlying this process need further researches to explore.

The findings of the present study implicate that comparison between the topic and vehicle of a metaphor is necessary because its literal meaning is accessed first especially when the metaphor is unfamiliar. This point is very helpful in the research area of natural language processing of artificial intelligence. The intelligent system could compute a plausible figurative meaning for a metaphor by comparing the literal meaning of the topic and vehicle. To complete the process of metaphor comprehension, the conceptual representations of the topic and vehicle are necessary for the intelligent system. Mental imagery of a metaphor would be a potential factor on the process.

Metaphors also act as a bridge between the human thinking activity and designs of human-computer interaction. Metaphors are used to designing and selecting good user interfaces such as the desktop of computer, and play an important role on user experience. On the other hand, metaphors of computer science and information sciences help psychologists and brain scientists see human thinking from a new perspective.

Acknowledgments. This research was supported in part by grants from 973 Program (2011CB302201) and the National Natural Science Foundation of China (61375009).

Reference

1. Bowdle, F., Gentner, D.: The Career of Metaphor. Psychological Review 112(1), 193–216 (2005)
2. Utsumi, A.: Computational Exploration of Metaphor Comprehension Processes Using a Semantic Space Model. Cognitive Science 35, 251–296 (2011)
3. Glucksberg, S., Keysar, B.: Understanding Metaphorical Comparisons: Beyond Similarity. Psychological Review 97(1), 3–18 (1990)
4. Blasko, D.G., Connine, M.: Effects of Familiarity and Aptness on Metaphor Processing. Journal of Experimental Psychology: Learning, Memory, and Cognition 19(2), 295–308 (1993)
5. Pynte, J., Besson, M., Robichon, F., Poli, J.: The Time-Course of Metaphor Comprehension: An Event-Related Potential Study. Brain and Language 55, 293–316 (1996)
6. Lu, A., Zhang, J.X.: Event-Related Potential Evidence for the Early Activation of Literal Meaning During Comprehension of Conventional Lexical Metaphors. Neuropsychologia 50, 1730–1738 (2012)
7. Chiappe, D.L., Chiappe, P.: The Role of Working Memory in Metaphor Production and Comprehension. Journal of Memory and Language 56, 172–188 (2007)
8. Pierce, R.S., Maclaren, R., Chiappe, D.L.: The Role of Working Memory in the Metaphor Interference Effect. Psychonomic Bulletin & Review 17(3), 400–404 (2007)
9. Beaty, R.E., Silvia, P.J.: Metaphorically Speaking: Cognitive Abilities and the Production of Figurative Language. Memory and Cognition 41, 255–267 (2013)
10. Paivio, A.: Mental Imagery in Associative Learning and Memory. Psychological Review 76(3), 241–263 (1969)
11. Gibbs, R.W., Gould, J.J., Andric, M.: Imagining Metaphorical Actions: Embodied Simulations Make the Impossible Plausible. Imagination, Cognition and Personality 25(3), 221–238 (2005-2006)
12. Wolff, P., Gentner, D.: Structure-mapping in Metaphor Comprehension. Cognitive Science 35, 1456–1448 (2011)

Natural and Multimodal Interfaces

Human Factors in the Design of BCI-Controlled Wheelchairs

Wafa Alrajhi[1], Manar Hosny[2], Areej Al-Wabil[2], and Arwa Alabdulkarim[3]

[1] College of Computer and Information Sciences, Imam University, Saudi Arabia
[2] College of Computer and Information Sciences, King Saud University, Saudi Arabia
[3] King Abdulaziz City for Science and Technology, Saudi Arabia
wafa7d@gmail.com, {mifawzi,aalwabil}@ksu.edu.sa,
amait64@hotmail.com

Abstract. In this paper, we synthesize research on the type of cognitive commands that have been examined for controlling Brain Computer Interface (BCI) wheelchairs and the human factors that have been reported for the selection of different protocols of BCI commands for an individual user. Moreover, we investigate how different researchers have considered the necessity of sustained movement from a single thought/command, having an emergency stop, and the commands necessary for assisting users with a particular disability. We then highlight how these human factors and ergonomics' considerations were applied in the design and development of an EEG-controlled motorized wheelchair, aiming to emphasize users' requirements for people with severe physical disabilities. In this case study, we propose a brain controlled wheelchair navigation system that can help the user travel to a desired destination, without having to personally drive the wheelchair and frequently change the movement directions along the path to the destination. The user can choose the desired destination from a map of the environment, using his/her brain signals only. The user can navigate through the map using BCI cognitive commands. The system processes the brain signals, determines the required destination on the map, and constructs an optimized movement path from the source to the intended destination. To construct an obstacle-free path with the shortest possible distance and minimum number of turns, a path planning optimization problem is solved using a simple Simulated Annealing (SA) algorithm. The resulting optimized path will be translated into movement directions that are sent to the microcontroller to move the wheelchair to the desired destination.

Keywords: Brain Computer Interaction (BCI), electroencephalography (EEG), Path Planning Optimization, Simulated Annealing, Wheelchair.

1 Introduction

Human factors in the design of assistive technologies are essential to the successful adoption and utilization of devices that provide alternatives to functional limitations imposed by users' physical disabilities. Recent advances in technologies have made it possible for a person to interact with and control devices using only his/her brain

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 513–522, 2014.

waves or Brain Computer Interaction (BCI). Brain–computer interactions/interfaces (BCIs), brain–machine interfaces (BMIs), Direct Brain Interfaces (DBIs), and neuroprostheses, all refer to the same concept. According to [1], a BCI interface was defined formally in the first international meeting for BCI research in June 1999 as: "A communication system that does not depend on the brain's normal output pathways of peripheral nerves and muscles" [1].

There are some available techniques to detect the brain activity such as electroencephalography (EEG) and magnetoencephalography (MEG), where EEG is considered to be the most common way to detect the electrical activity in the brain for the context of wheelchair designs [2]. In EEG systems, the sensors are placed on the brain scalp without surgical intervention. Nowadays, unobtrusive wireless headsets are available that can be used to detect EEG signals (e.g. Emotiv's EPOC and Neurosky's Mindwave [5-6]). The available EEG headsets are relatively inexpensive, easy to wear and control. Furthermore, the temporal resolution of EEG which represents the ability to detect changes within a certain time interval is relatively good; a millisecond or even better. However, the spatial resolution - a measurement of the accuracy of a graphic display - and the frequency range are limited, This consequently limits the amount of information that can be extracted [3-4]. One of the popular EEG headsets is EPOC which is made by Emotiv Systems.

The proliferation of BCI-oriented assistive technologies have the potential to improve the quality of life of people with severe motor disabilities with increased independence and less reliance on caregivers. Among the promising devices that have been developed for this purpose, is an EEG based brain controlled wheelchair, which the user can move using his/her brain signals only; hence, alleviating the need for any physical movement to control the device [7]. This wheelchair can be used to serve people who cannot move their limbs or people living with spinal cord injury. Nevertheless, a person with a disability may face difficulty in controlling the brain controlled-wheelchair for long periods of time, since the procedure usually requires non-trivial concentration by the person with a disability throughout the navigation process from the source to the destination. Accuracy of BCI-controlled systems remains a concern and using brainwaves to drive a wheelchair may not effectively lead the user to the required destination.

Taking such difficulty into account, we developed a brain controlled wheelchair system, which we called Brain-Wheel, in a way that will relieve users from the task of planning the path to the destination. To avoid the inaccuracy of existing BCI tools, we are restricting the use of BCI to the selection of the destination. Hence, BCI is not utilized in this context for guiding the wheelchair step-by-step as the user is navigating to the destination. The system was designed so that users of this system can choose a target destination, which they would like to navigate to, from the 2-D environment map using their brain signals. In the system, Emotiv's Epoc is used to detect the brain signals for selecting the required destination from the presented room map. In the Brain-Wheel system, we used the Emotiv cognitive suite, where the headset can understand the user's intent to perform specific actions. Based on the user's intention to move, the detected brain signals will determine whether or not to start the navigation system. The navigation system will then decide the optimal path that the wheelchair should follow using a metaheuristic algorithm, which has been specifically designed for this problem. The output of the algorithm will be fed back to the microcontroller where we used an Arduino UNO Rev3 [8]. The circuit, which the Arduino controls, consists of two servos [9]. Once the signals from the software are received, the Arduino directs the two servos to rotate accordingly, to push the wheelchair's joystick shaft forward, backward, left, or right. Thus, allowing the wheelchair to move to the desired location via the user's command. Insights from this project and reflections on the design of related systems are discussed in this paper.

In this paper, a review of related work is presented in the next two sections. Then, we discuss the Brain-Wheel system that we developed with an emphasis on the human factors related to BCI control and motion modules.

2 Human Factors in the Design of Powered Wheelchairs

In this section, we describe the human factors in the design of wheelchairs that support independent movement of users with a range of disabilities. Innovative designs for wheelchairs have emerged in recent years that address a wide spectrum of ergonomics ranging from the seats, motor controls, and head support to the interaction modalities that facilitate freedom of navigation and movement with configurable controls.

Innovations in wheelchair design are intended to improve the ergonomics of wheelchairs and independence of wheelchair users, thereby saving the cost of additional treatment or assistance in daily living. Complexities in the interaction between wheelchairs and their users have risen in recent years that are in-line with advancements in computing power, decrease in cost of microcontrollers, and the emergence of a variety of sensors. Human factors in the design of wheelchairs have been examined extensively with regards to the mechanical components such as the seats, foot rests, hand rims, castors, head supports and arm rests [16]. Several factors influence the energy needed to propel wheelchairs; most notably are as the users' position and the control modules for navigating in the space. Human factors related to the control components of electrical powered BCI wheelchairs have been recognized as key design issues due to the inaccuracy of sensors in BCI modules but have been inadequately examined [e.g. 10]. BCI-controlled wheelchairs have been designed with wired and wireless EEG headsets. Wireless headsets have the advantage of increased freedom of head movement but with less accuracy in interaction/control. On the other hand, wired headsets provide more accuracy but in a more obtrusive setting using the EEG caps and constrained movements. Navigation interfaces have facilitated controlling the movement commands and the selection of destinations in gradual navigation through physical spaces. Virtual environments have been proposed to train users in a safe context-of-use before engaging in the real-time control of the BCI wheelchair in the actual environment [10]. Minimizing the cognitive load of users in interacting with BCI wheelchairs is a key design factor and different control mechanisms have been examined where some interfaces allow users to select the navigation path phase-by-phase while other interfaces facilitate selecting only the destination and handover the path-planning and maneuvering task to the computing and mechanical modules of the wheelchair [10-11, 14]. Computational intelligence has potential for contribution in such scenarios of Human-Computer Interaction (HCI) contexts of research and development to alleviate the cognitive load of users; however, very few attempts been reported in the literature to address this interaction design problem for BCI-controlled wheelchairs. Reducing the mental effort and concentration of users that is required for BCI-controlled wheelchairs has been examined for selection components in [10] and in stopping controls in [11] and [14]. In user acceptance evaluations of BCI-controlled wheelchairs, human factors of response time of BCI, training time of the systems to recognize patterns of user thoughts and interpreting them into commands, and the thresholds of mental effort required to trigger controls (e.g. selection, navigation, sustained attention for recognition of evoked potentials) are key in the effective design of such assistive technologies.

3 BCI-Controlled Wheelchair Designs

BCI-controlled wheelchair prototypes have been developed to provide un-aided control of wheelcahairs for people with disabilities. In this section we present some of the existing BCI applications designed for powered wheelchairs.

A brain controlled wheelchair system was proposed in [10]. The proposed system is composed of three stages: detect the brain signals, classify them into actions, and interfacing to the wheelchair. Firstly, to detect the brain signals the authors used 16-channel 24-bit electroencephalogram (EEG). Sensorimotor rhythms (SMR), which can be produced by imagining the limbs or moving them, are used to produce the desired brain signal. To achieve the second step, which requires understanding and classifying the detected signals, the authors investigated several feature extraction algorithms, such as discrete Fourier transform (DFT) and common spatial patterns (CSP). CSP aims to facilitate the process of differentiating between the two classes of data by increasing the variation between them, which aids in the classification process. Different machine learning algorithms have been used as a classifier. Support Vector Machine (SVM) was used to predict the class of the given input. After the feature selection phase, the authors investigated the optimal sensors number and location. Over 60 sensors, the sensors that produce the most demanding signal that can serve both CPS and the classifier were chosen. The classification performance results show that when the number of sensors is increased, the classification results will be better. The system has been tested in a virtual 3D simulated environment and a modular controller was used as an interface to the wheelchair.

B. Rebsamen et al. [11] also develop a BCI-controlled wheelchair using a hybrid P300and mu-Beta interface. The authors used visual stimuli to invoke P300 signals where the items or destinations that the users can navigate to are presented and flashed sequentially. To select a destination from the presented list, the user needs to focus his/her attention on the destination image. P300 signals were used select the navigation item that the user focuses on. The authors of [11] represented the navigation environment as a graph, where a limited number of destinations through the environment can be selected to navigate to. These destinations are linked via virtual paths where the paths are stored in the memory, such that depending on the destination the required path will be retrieved. Thus, the paths to the destination are not calculated using computational intelligence. While this solution may resolve navigation problems in a relatively static environment, dynamic contexts-of-use would require human intervention to modify the paths for the BCI-controlled wheelchair.

4 Brain-Wheel: A Brain-Controlled Wheelchair

The design of "Brain-Wheel" combines BCI with an optimized wheelchair navigation path that takes into account the context-of-use and physical environment. This section reviews the developed Brain-Wheel system describing its main components, which are: detecting the brain signals, constructing the path which requires the room map to be processed first. Later, the path will be fed to the motion module or control box. The system design is shown in Figure 1.

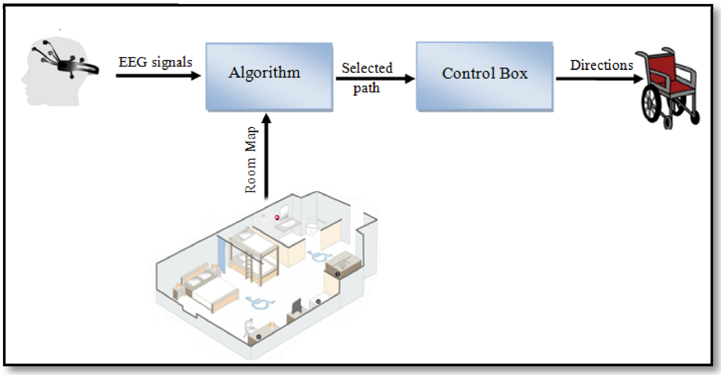


Fig. 1. System Architecture Design

4.1 Detecting Brain Signals

The main goal of the system is to help those who have lost their ability to move their four limbs by assisting them to navigate from one place to another. Thus we developed our system in a way that can respond to the user's cognitive actions using the Emotiv cognitive suite. The Emotiv Epoch headset was considered due to its wireless connection to the computer and its ability to detect different type of signals. A list of the cognitive actions can be found in the Emotiv cognitive suite such as: push, pull, right, left,...etc. The default action, which the user needs to train the Emotiv headset on first, is the neutral state of the user. While training, the user must be focused and avoid any distraction to enable the panel to detect the appropriate signals.

Two cognitive actions were considered in the design of the system: Push and Right. Push is used to simulate pressing the buttons, which in turn is translated to selecting a destination from the map, after reaching it through the Right cognitive command. The Right navigation command is used to move through the map from one cell to another. The user of Brain-wheel must train Emotiv first, before starting to use the system. This is to facilitate allowing the system to save his/her profile and thus recognize his/her brain signals for each cognitive action. After training, the system can respond to the user's push cognitive command, which will move the user from one window to another.

The room, where the user wants to navigate through, has been transformed to a digital 2-D image and has been divided in to cells, i.e., rows and columns that fit the wheelchair size. Figure 2 shows the 2-D map, where the initial location of the user is assumed to be at the location of the entrance.

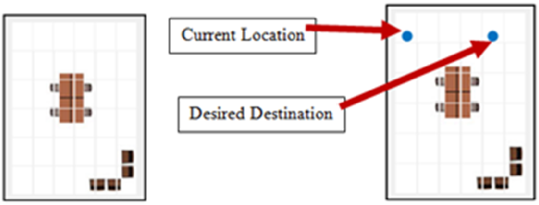


Fig. 2. Selecting a destination using push cognitive commands

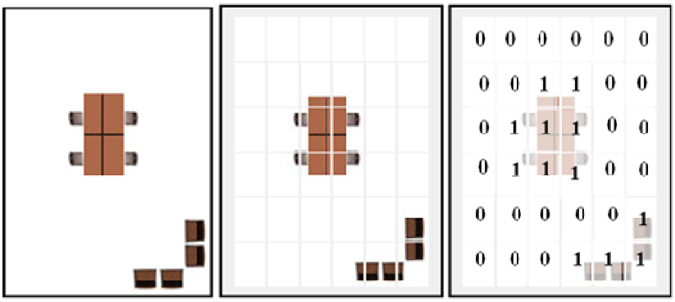
4.2 Constructing the Path to the Destination

Optimizing the path for a robot or a vehicle movement is an interesting field of study. In the last decade, the path optimization problem has received the attention of many researchers, due to its close connection to robot movement applications. The path planning optimization problem can be defined as: trying to find a collision-free path that connects a specific starting location with a specific goal/destination. In the path planning optimization problem, each location in the path is represented as a state, and the transition between those states represents the actions [12]. The path is optimal when the sum of its transitions' costs is less than the cost of all possible paths that lead to the same target. There are several existing approaches for computing the optimal path. Heuristic and metaheuristic-based algorithms are among the common algorithms to solve the path planning problem [12].

Metaheuristics are general purpose algorithms that can be used to solve difficult problems. There are many metaheuristic-based algorithms that have been used to solve the path planning problem, such as Ant Colony Optimization, Genetic algorithms, Swarm optimization, etc. The metaheuristic that we selected to solve the path planning optimization problem is Simulated Annealing (SA) [15]. Using the detected signals and the chosen destination, the system processes the SA algorithm to optimize the path to be followed towards the destination. The path planning optimization process in our BCI wheelchair goes through two phases:

Map Processing

Before planning the path, the system processes a 2D map of the room to define the obstacles' locations so that the generated path would not intersect with them. To define the obstacles' locations, the map will be converted into a digital matrix that contains 0’s and 1’s using an image segmentation algorithm implemented in MATLAB. The output matrix is shown in Figure 3.



0	0	0	0	0	0
0	0	1	1	0	0
0	1	1	1	0	0
0	1	1	1	0	0
0	0	0	0	0	1
0	0	0	1	1	1

Fig. 3. Transform a map into a matrix

Optimization Using Simulated Annealing

The designed algorithm is based on a Simulated Annealing (SA) metaheuristic, where the advantage of using SA over other metaheuristics is that it can overcome the drawbacks of direct local search in terms of being trapped in a local optimum, and thus can produce better quality solutions. In addition, being a single solution based metaheuristic, simulated annealing is also simpler to implement and less expensive than population based methods (like evolutionary and swarm intelligence algorithms), in terms of both time and space, since it does not need to keep track of multiple solutions at the same time.

In a metaheuristic search, there are different solution representations that can be used according to the type of the optimization problem. In our design, we make some simplifying assumptions, such as representing the environment as a grid where the grid cells are numbered from 0 to N. To move from a node (cell) to another, four possible moves are allowed (horizontal and vertical), whereas diagonal movements are prohibited. This is meant to ease the wheelchair movement through horizontal and vertical directions only. For the path representation, a permutation representation is adopted where the path is represented as a sequence of integer numbers. Each number represents the cell number that the wheelchair will move through.

The objective function for our path optimization problem is concerned with the total path distance, the number of obstacles encountered, and the number of turns (twists) in the path. Due to the allowed movements, the path will add one of its adjacent cells at each movement. Thus, the number of nodes that the path passes through represents the total distance of the path. In addition, for every obstacle passed through in the path, i.e. a non-free cell, a penalty value will be imposed. Moreover, to avoid jerky movements of the wheelchair, the number of turns in the path will also be penalized, such that each change between horizontal and vertical movement will be considered as a turn.

In the SA algorithm, the initial path is improved step by step by generating a neighboring path and comparing its quality with the current path. If the new path has a better quality, it replaces the current path. Otherwise, the new path is adopted with a certain probability. To generate a neighboring path, a special neighborhood move was designed to fit the constraints of the path optimization problem presented here.

4.3 Motion Module

After the path has been generated, the software would translate the path into commands of “forward”, “right”, or “left”. The commands are sent to the Arduino UNO microcontroller, the brain of the motion module, and are received as “F” for “forward”, “R” for “right”, and “L” for “left”. The Arduino is programmed to control two servos mounted over the wheelchair’s joystick’s shaft. Given a command signal, from the Arduino to the servo, the servo's motor will turn its own shaft to a specified angle. Each servo controls an axis (X, Y), and their initial setting is 90 degrees (middle value) each. When “F” is received, the Y-axis servo would turn to 180 degree, and the X-axis servo would be in its initial setting, thus moving the joystick forward. When “R” is received, The Y-axis servo would be given its initial setting value, and the other servo a 0 degree, same for when the “L” is received except now the x-axis servo is set to 180 degree. Each command is carried out for one second; then would lock back to its initial setting 90 degrees for each servo, for it to stop. This was made to help avoid collisions. Communication between the Arduino and PC/LAPTOP is made using serial communication, over a USB cable. Once the signals from the software are received, the Arduino will direct the two servos to rotate accordingly, to push the wheelchair’s joystick shaft forward, left, or right. Thus, allowing the wheelchair to move to the desired location given by the user’s command. Because our work was designed as an external component to the powered wheelchair, and we didn’t modify the mechanical components of the wheelchair, it is envisioned that similar power wheelchair models could integrate our system in their design. Our system allows the wheelchair driver to sit comfortably in his/her power wheelchair, only facing the laptop screen on their lap tray, and wearing the EEG headset. A USB cable is connected from the laptop to the motion module.

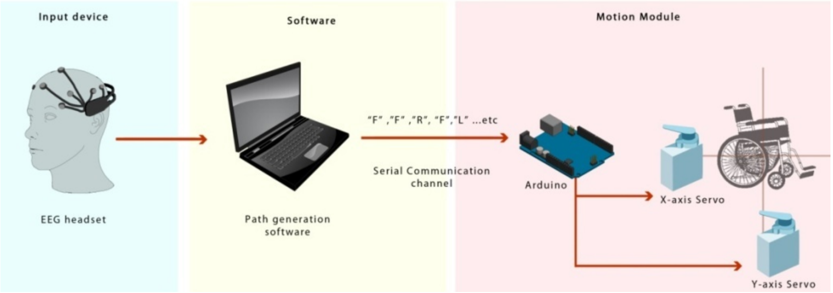


Fig. 4. Motion Module

5 Conclusion

Aiming to facilitate the navigation of the wheelchair and other brain controlled devices, we proposed in the Brain-Wheel project a navigation system that combines Brain Computer Interaction and Path Planning Optimization. Instead of guiding a device to the destination, an interface that contains the environment map will be presented to the user offering various destinations to be reached from the point of navigation. The user has to select a destination from the presented map using his/her brain signals. Two cognitive actions have been used in Brain-Wheel: push cognitive and right cognitive. Using right cognitive, the user can navigate from one cell to another. When the required destination cell is reached a push cognitive is required. The system will then construct a collision-free path to the desired destination using a Simulated Annealing metaheuristic. Finally, the path will be fed to wheelchair using a control box that will transform the path into directions of movements that are connected to the wheelchair's motor modules.

The current system is a prototype at this stage, but in future work, we wish to create an ergonomically designed enclosure for our wheelchair motion module. For example, the speed of the wheelchair was fixed during testing of the system, and cannot be changed. A more flexible system would allow the user to select the speed of their choice. As for safety measures, we plan to add sensors to prevent the wheelchair driver from colliding with unobserved objects in our system.

Acknowledgment. "The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding the work through the research group project number RGP-VPP-157". The authors would also like to thank Omar Abunayyan for his help in designing the motion module's physical parts and coordinating the collaboration with the Advanced Lab of Intelligent System Research (ALISR), King Saud University. In addition, the authors are grateful to Eng. Salim Benyoucef for designing the motion module's mechanical mechanism to house the two servo motors and creating and printing a 3D model of the design

References

1. Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., Donchin, E., Quatrano, L.A., Robinson, C.J., Vaughan, T.M.: Brain-computer interface technology: A review of the first international meeting. IEEE Transactions on Rehabilitation Engineering 8(2), 164–173 (2000)
2. Constantin, A.: A Brain-Computer Interface for the Classification of Motor Imagery. Bachelor thesis, Williams College, USA (2007)
3. Graimann, B., Allison, B., Pfurtscheller, G.: Brain–computer interfaces: A gentle introduction. In: Brain-Computer Interfaces, pp. 1–27. Springer, London (2010)
4. Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo Zhang, N.: Evolving Signal Processing for Brain–Computer Interfaces. Proceedings of the IEEE 100, 1567–1584 (2012)
5. Neurosky (2012) (January 2, 2013) Internet: http://www.neurosky.com/
6. Emotiv Software Development Kit, http://emotiv.com

7. L.L.C. Puzzlebox Productions, Puzzlebox Brainstorms (February 4, 2013), Internet: http://brainstorms.puzzlebox.info/index.php/index.php
8. Arduino, Arduino Uno Board, http://arduino.cc/en/Main/arduinoBoardUno (accessed: December 14, 2013)
9. Platt, C.: Servo motor. In: Encyclopedia of Electronic Components, vol. 1, pp. 201–207. Maker Media (2012)
10. Yazdani, N., Khazab, F., Fitzgibbon, S., Luerssen, M., Powers, D., Clark, C.R.: Towards a brain-controlled Wheelchair Prototype. In: Proceedings of the 24th BCS Interaction Specialist Group Conference, pp. 453–457 (2010)
11. Rebsamen, B., Burdet, E., Zeng, Q., Zhang, H., Ang, M., Teo, C.L., Guan, C., Laugier, C.: Hybrid P300 and Mu-Beta brain computer interface to operate a brain controlled wheelchair. In: Proceedings of the 2nd International Convention on Rehabilitation Engineering & Assistive Technology, pp. 51–55 (2008)
12. Moreno, J.A.: Heuristic algorithm for robot path planning based on a growing elastic net. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 447–454. Springer, Heidelberg (2005)
13. Al-Ghamdi, N., Al-Hudhud, G., Alzamel, M., Al-Wabil, A.: Trials and tribulations of BCI control applications. In: Science and Information Conference (SAI), pp. 212–217 (October 2013)
14. Mandel, C., Lüth, T., Laue, T., Röfer, T., Gräser, A., Krieg-Brückner, B.: Navigating a smart wheelchair with a brain-computer interface interpreting steady-state visual evoked potentials. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 1118–1125. IEEE Press, Piscataway (2009)
15. Kirkpatrick, S.: Optimization by simulated annealing: Quantitative studies. J. Stat. Phys. 34(5-6), 975–986 (1984)
16. Carlson, T., Demiris, Y.: Collaborative Control for a Robotic Wheelchair: Evaluation of Performance, Attention, and Workload. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42(3), 876–888 (2012)

Interface Design and Dynamic Audio

Luiz Roberto Carvalho[1] and Alice T. Cybis Pereira[2]

[1] Federal University of Santa Catarina
semprecarvalho@gmail.com
[2] Federal University of Santa Catarina
acybis@gmail.com

Abstract. In the age of digital devices, text, image, sound, interactivity, blend themselves into a symbiotic and unique media, presenting a multifaceted specie of language called hypermedia. However, since many years ago, we have seen a notable emphasis on visual communication´s interfaces, and due to its limitations, products and services in design can often present inconsistencies when other sensory properties are relevant, as in the case of sound information. This over-emphasis on visual displays has constrained the development of interactive systems that are capable of making better use of the auditory modality. Recognizing the HCI as an integrating element of media and visual, sound and tactile metaphors, this study will demonstrate investigations that contextualize the role of sound into interactive environments by proposing an overview for the term interactive sound, suggesting its classification into direct-interactive and indirect-adaptative sounds, and pointing out its meanings and applications.

Keywords: sound design, game sound, dynamic audio, interactive sound.

1 The Sensorial Multiplicity of the Interface

According to Cavalcante (2010, p.200), “interface is the surface to access and exchange information”. Shneiderman & Plaisant (2009) point out that the usage of sound, tridimensional representations and animations is growing in order to improve the appeal and the possibilities of content presentation in the interfaces, as well to better attend the cognitive characteristics of the users. In order to make a user undergo to an optimized execution of a task, exceptional conditions can be presented to attract their attention. On the definition of Padovani & Moura (2008, p.24) “The attention is a concentration of mental activity. It allows the user to select a perceptual channel, deciding which information should be prioritized in determined contexts”.

Preece (2005, p.86) points out that “the attention consists on the process of selecting one thing to concentrate on a certain moment among the variety of things available possibilities”. The author emphasizes that it involves auditory and visual senses, since it allows users to focus on relevant information for what they´re doing.

Into an interactive environment, it’s important to point out that the graphic interface will influence upon the use of the auditory resources. Both must be congruent, that is, projected as complements of each other. Menzies (2002) stands out that to use the auditory resource we must be cautious with the number of audio objects that will

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 523–531, 2014.

compose the interface, since its excess can cause a negative impact and even noise pollution. The same author says that on his works there was an enlargement on the usage of sounds into interactive systems and their majority is because of dynamic interactions on the environment. For an adequate development of an interactive project, the management of interactive auditory objects has some primary points that should be observed, as:

- The object that will use the auditory resources must be clearly defined;
- The interface must be projected to allow the creation and be able for possible alterations on the interactive acoustic environments.

Shneidermann & Plaisant (2009) explains that audio can be used into interactive systems according to three different ways: narration or dialogue, ambient sound or music, and as sound effects. In narration, the human speech intends to be informative, presenting explanations when necessary. The background music defines the mood and the rhythms of the presentation and it's linked to emotional interpretations that can stimulate reactions on the user. The sound effects are brief and have a function to stand out some point of the message, focusing and enlarging its impact. The lack of sound – the silence – can also transmit some sort of information. A pause can indicate a change into the narrative script, as well as the use of continuous music in different scenes can help to point out the maintenance of a particular theme. On the cognitive point of view, the auditory channel is an excellent way to transmit information and give feedback to the user. While the background music can be used to evocate emotions and define an environment; a specific sonorous effect can be used to transmit specific alert information (NOKIA CORP., 2005 apud COLLINS, 2008, p.78).

2 Dynamic Audio: Direct-Interactive and Indirect-Adaptive Sounds

The applications of sound in interactive platforms differ from the commonly used in music and movies, which are essentially linear. In games, the most important element of interactivity and that offers meaning to the terms, according to Richard Rouse (apud Collins, 2008, p.4) is the non-linearity, because "without the non-linearity the developers of games would be developing films". The term non-linear refers to the fact that the user has choices, and each choice will result on the construction of a different narrative. It is the primary distinction that separates the interactive environments of traditional linear applications, as cinema and television, which the succession of events happens in a fixed and unchangeable way.

In the last years, with the digital devices development, the dynamic behavior of sound offered a significant improvement above the interaction and immersion user´s level. The transitions between graphic spaces and sound assets become more integrated, allowing the system to offer immediate responses to users. According to Bar-B-Q (2003), interactive audio is any type of sound produced by an audio system that is programmed to generate a real-time response to an input stimulus, on the form a pre-determined sonorous expression. This audio system is composed by an "interactive audio engine⁴[1]" that receives users' commands (input stimuli), select the data (interactive data directory), and then send it to an interactive audio output directory. This output will then send an electric impulse thru the speakers, which will result in the generation of sound waves.

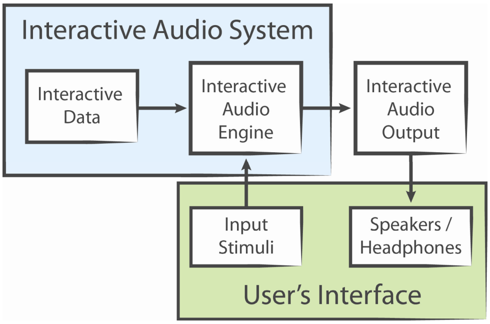


Fig. 1. Interactive sound structure (adapted from Bar-B-Q, 2003)

Not all of the audio systems that react to an input stimulus can be defined as interactive. Bar-B-Q (2003) affirms that an interactive system allows changes on the input command to modify the auditory behavior, while a reactive system simply statically reproduces audio events without any kind of response related to the user's stimuli. This interactive system can be classified in two categories: as direct audio, the user controls the audio responses consciously; as indirect audio, the user is controlling some other parameter that affects the audio behavior. According to Bar-B-Q (2008), the main characteristics of interactive audio are:

- Improves the user´s experience;
- Inspires the user´s involvement;
- Creates a unique personality for the products;
- Allows users to make new kind of activities;
- Creates a participative experience;
- It is potentially cheaper to implement;

Among the possible platforms for interactive audio's application, there are virtual environments, games, web applications, music players, e-books, softwares, smartphones, vehicles, household appliances, toys and others. Since digital devices are in constant growth and development, possibilities for interactive audio´s applications get larger.

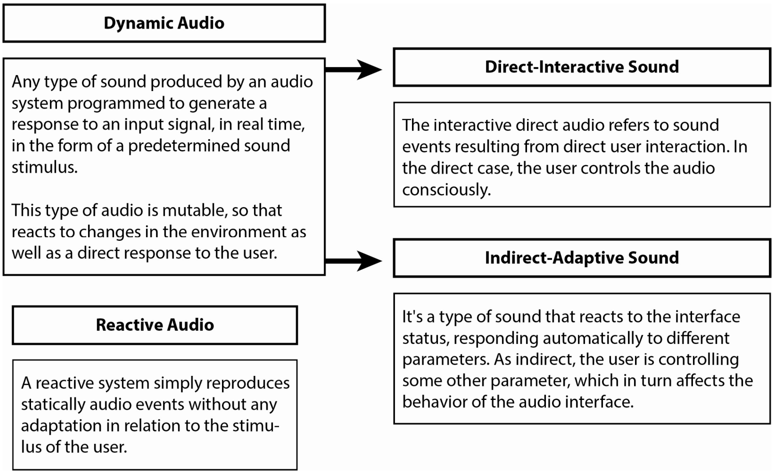
In discussions about audio in interactive environments, the term *interactive* is sometimes used to exchange or amplify the meaning of the terms *reactive* and *adaptive*. In fact, the interactive audio refers to sound events resulting from the user's direct interaction, and the adaptive audio, however, is a type of sound that reacts to the interface's state and its status, reacting to distinctive parameters. To avoid that the

[1] A group of software or hardware of algorithms that process interactive data based on entrance stimuli to process an audio exit.

name *interactive* can result into a technical ambiguity, Collins (2008) suggests the term *dynamic audio* to define sound events occurring on the interface, which is divided and classified into *interactive* and *adaptive sounds*.

Collins (2008) defines the dynamic audio as a mutable audio, a wide concept that embraces interactive and adaptive sounds. It is the audio that reacts to environmental changes and to the user's response. The interactive audio is defined by a sound event that reacts to a direct user's signal, as a sound emitted when a button is pressed. In a different way, the adaptive audio reacts to an interface's status, reacting to pre-established parameters that aren't directly controlled by the user. On the adaptive audio, the emitted sounds are not plainly determined by the user's action, that is, it involves other variables that the user has no direct control of it. On the game *Super Mario Bross* (Nintendo Inc.) the rhythm of the music increases when the stage level timer is about to end, to warn the user that he should get hurry to finish his task.

All types of sounds that are put into an interactive environment, as music, dialogue and sound effects, can be classified according to Bar-B-Q (2003) as *direct* or *indirect interactive* audio, or according to Collins (2008), as *interactive* and *adaptive dynamic* audio, and according to both, as reactive audio. Since these concepts agree with one another, and aiming to minimize the possibility of ambiguity, the categorization of the terms direct-interactive sound and indirect-adaptive sound are assumed, both being part to the dynamic audio group, as proposed on the following image:



Dynamic Audio

Any type of sound produced by an audio system programmed to generate a response to an input signal, in real time, in the form of a predetermined sound stimulus.

This type of audio is mutable, so that reacts to changes in the environment as well as a direct response to the user.

Direct-Interactive Sound

The interactive direct audio refers to sound events resulting from direct user interaction. In the direct case, the user controls the audio consciously.

Indirect-Adaptive Sound

It's a type of sound that reacts to the interface status, responding automatically to different parameters. As indirect, the user is controlling some other parameter, which in turn affects the behavior of the audio interface.

Reactive Audio

A reactive system simply reproduces statically audio events without any adaptation in relation to the stimulus of the user.

Fig. 2. Terminological proposition of dynamic and reactive audio, direct-interactive and indirect-adaptive sounds

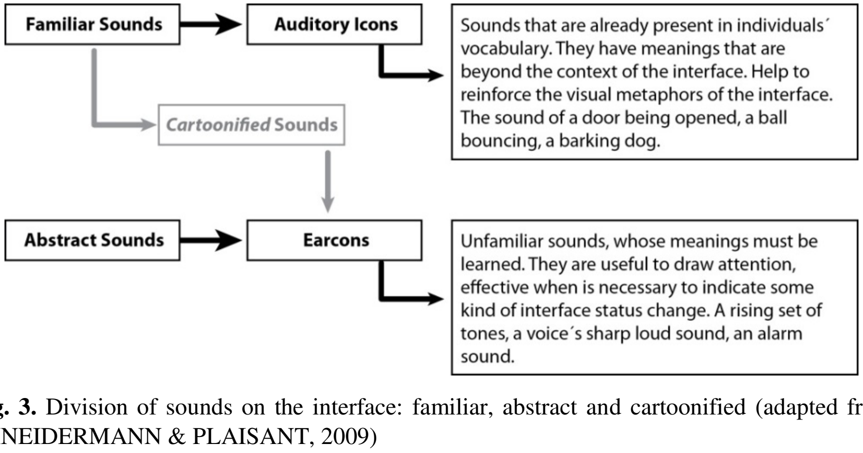
2.1 The Specificity of Dynamic Audio

Koji Kondo (2007, apud COLLINS, 2008, p.139), musical compositor of the Super Mario game series, describes four components of the dynamic music, typical of interactive environments:

- The ability to create music that change according to user's interaction;
- The ability to develop a multilayered production, by creating distinctive themes for the same composition;
- The ability to add surprise elements, enlarging the interactivity;
- The ability to add musical elements with specific characteristics that matches to the interface's condition;

According to the author, dynamic music should evidence the participative character of the interface and indicate changes of its condition. These objectives can be achieved by the use of rhythmic changing, instrument and voice additions, and even by altering the whole music according to distinctive stages of the interface, with possibility of adding or varying the reproduction of the sonorous sentences. The dynamic music must react or interact according to the narrative and fulfill the user's expectations.

Shneiderman & Plaisant (2009) point that a useful distinction of dynamic sounds is to classify them between *familiar sounds* - that are called *auditory icons* - and *abstract sounds*, called *earcons*. Auditory icons (as the sound of a door being opened or a ball jumping) help to reinforce the visual metaphors of the interface. Sonorous icons represent the kinds of sounds that aren't known by the user, and their meanings must be learned. Other category of sound includes the *cartoonified sounds*, which exaggerate the aspects of familiar sounds. Cartoonified sounds must be considered as belonging to the abstract sounds group because they normally don't have any relation with the sounds that are noticed on a physical environment, because of their characteristics of exaggeration and increasing proportions. The following image is an scheme of these concepts:



Sounds that are already present in individuals´ vocabulary. They have meanings that are beyond the context of the interface. Help to reinforce the visual metaphors of the interface. The sound of a door being opened, a ball bouncing, a barking dog.

Unfamiliar sounds, whose meanings must be learned. They are useful to draw attention, effective when is necessary to indicate some kind of interface status change. A rising set of tones, a voice´s sharp loud sound, an alarm sound.

Fig. 3. Division of sounds on the interface: familiar, abstract and cartoonified (adapted from SHNEIDERMANN & PLAISANT, 2009)

The Interactive Audio Special Interest Group (IASIG, 2012) proposes a series of functions attributed to dynamic audio, affirming the existence of a considerable superposition between these categories and pointing that they shouldn't be considered mutually excluding. These functions are divided in semiotics, emotional (intimately linked to the semiotic function), structural, narrative, immersive, esthetical and kinetic.

According to IASIG (2012), the semiotics function of dynamic audio seek to transmit an emotional meaning, aiming to guide the user's attention to identify objectives in a way that it is possible to diminish the learning curve, creating a positive level of inclination with the interface. Sound symbols help to identify objectives and focus on the user's perception in determined objects. In many games, for example, the presence of enemies starts a tense music, and when the user finds benefic elements, as coins and heart shaped life, these end up having the same good sonorous suggestions. In other words, sound symbols are used many times to help users to identify other elements of the interface. These symbols can induce suggestions of humor and feelings when incorporated in the interactive environment, in a way that it is possible to make the interface more comprehensible.

A crucial semiotics role that the dynamic sound performs is its preparatory and anticipatory function. In games, anticipate action is a critical element, especially on adventure titles. Sounds without clear visual indication can incentive the user to look to the direction of a sound. The usage of sounds to add a behavioral and cognitive tendency is as important as the preparatory elements of dynamic audio. They change the user's perspective about the interface.

The emotional functions are intimately related to the semiotics. Here, a distinction must be done between transmitting a message through sounds and its condition of humor induction: the mood changes according to what the user feels, while emotional aspects simply transfer information. The user can understand that a sound exposes sadness without being sad. Considerable quantities of sounds on the interfaces have emotional effects that can enlarge or diminish the degree of difficulty for the execution of a task, as the case of the increasing rhythm of a composition while a task is being executed. In this way, sound can control or manipulate the user's emotions, guiding responses through to the interface.

Structural functions of dynamic audio are linked to the act of creating, reinforcing or masking the interface structure to indicate changes on the narrative and situate the user. As in the movies, music and sounds are used frequently in interfaces to reinforce or improve its continuity structure. A significant example of the utilization of structural functions of dynamic audio is on the game *Vib Ribbon* (SCEI, 1999) in which the music can literally direct the narrative of the content. The game allows users to insert their preferable music to be used as a reference to map the difficulty levels. The system scans the audio signal characteristics, and then executes two obstacle courses for each song (one easy and other hard). The narrative structure of the interface can vary according to the chosen song. Although this is a very singular case, the utilization of songs and audio tracks for the creation of interface structures is a resource that has a meaningful potential. However, the dynamic audio is used with more frequency to improve the general structure of the interface. The inclusion of sonorous clues intermediating two interfaces acts as a linking element for a gradual and continuous transition of the contents. A brief silence can also inform the user that the suggested time for the execution of a task can be over, or that something different is going to happen, indicating possible changes of the interface condition.

In many cases, audio signals can help to situate the user on the interface's narrative context. When listening to distinctive musical compositions in different areas, the user is capable of identifying his whereabouts through the response of sounds. The audio is commonly used to locate the user on the plot, anchoring the user in terms of where, when and what is happening, as well as serving as an anticipation element for what is yet to come. The dialogue can also serve as a big event on the narrative, and it is used to reveal the attribution of objectives and specific tasks of the application. Non-verbal sounds can also reveal details about environment and objects through the usage of ambient moods that are particularly useful to create empathy and familiarization, making the interfaces more immersive.

The immersive function develops a critical role on the dynamic audio. It deals directly with the suspension of interface's incredulity, adding realism through the creation of an illusion of the *real*, which is indispensable for the user's immersion. The IASIG (2012) points out that the illusion of being immerse in a virtual environment is really reinforced by audio. Besides integrating the user on the narrative of the interface, the dynamic audio can also be used to make the sounds of the application direct focus on the user in a way that he won´t be distracted by sonorous stimuli produced by the environment around him; they can be noise, sounds, voices. Dynamic audio can help masking these external sounds, as it progressively enlarges the user's concentration and focus on the interface task.

The kinetic functions connect a sensorial stimulus of audio to a specific motor response of the individuals. Some interfaces are projected so that the users can directly interact with sonorous stimuli and physical moves, like in the game *Dance Dance Revolution* (Konami, 2000). Kinetic dynamic audio works as the main motivator factor for the execution of movements because is the primary element that confirms (or rejects) the correct execution of a task on the interface's context.

The esthetical functions of the dynamic audio deal with the creation of identity and intertextual references of the interface, in which the sound is used to create beauty, generating acceptance and familiarity. The esthetical function offers the possibility to create moods that supply clues about the characteristics of the interface. Slow and smooth introductory songs, for example, normally indicate that the interface has a light rhythm of tasks. Musical compositions that are more accelerated normally indicate action and dynamic. Certain kinds of music adapt well to certain types of interfaces, and knowing that different types of narratives have distinct interactivity requirements, sound elements should also follow these patterns.

However, there is a series of variables that difficult the insertion of dynamic audio. The duration of an interface condition is a complex element to quantify, having in mind that each user manipulates the interface according to his familiarity and knowledge. The user experience will also influence the interactivity with sound elements. Sometimes, sounds are not sufficiently relevant to the interface's context, and become repetitive and boring, generating a listener's fatigue. The concept of *listener's fatigue* must be treated with caution; some interfaces are projected for consecutive use and repetitive sounds can be tiring, especially if the user spends a lot of time on a particularly specific area of the interface.

To solve this difficulty, some games started to incorporate timings for the audio clues, in a way that if the user stays in a determined environment, the music won't repeat endlessly; instead, it just stops being played. Marty O'Donnell (apud COLLINS, 2008, p.14) argues about the game *Halo*, that has a command called *I'm bored* that – if the user haven't completed a determined objective and 5 minutes are gone –the song just disappears, in fadeout[2].

According to Shneiderman & Plaisant (2009), since the origin of the desktop's interfaces, a series of sounds was used to indicate tasks, warnings, or even to point out the conclusion of an action, as the sound that is emitted when an archive is put on the trash can for elimination into an operational system. The effect for most of the users is a satisfactory confirmation of actions; on the other hand, after a few hours the sounds can become a distraction instead of a contribution, especially where there are many machines and many users on the same area. Some applications ring a bell or the sound of a tune when an error occurs. This alarm can be useful if the user could lose the mistake but it can also be embarrassing if other people are on the same area. Sound designers must find a way between drawing attention to a problem and avoid embarrassment to the user. Considering the ample and distinctive range of experience and temperament of the users, the most appropriate solution is offer to user some control over the sounds, making the dynamic audio approach consonant with the principles of *user´s experience design*.

3 Final Considerations

There are many examples of a growing tendency pointing to dynamic audio beyond the traditional computer desktop applications: videogames, smartphones, tablets, domestic devices, vehicle systems. The most advanced dynamic audio system that exists nowadays is found on videogames platforms. As games become more sophisticated, instances of audio that are reproduced in response to a user's stimuli are also becoming more intelligent. That said, the development of dynamic audio into any kind of interactive environment requires tracking innovations that are found in video games.

Interactive spaces offer a wide range of possibilities to intersect different modalities of language in a never before offered way. On the digital environment, there is no construction of meaning only by a unique semiotics system. It deals with systems that allow navigation through distinct groups of information in a multi linear way, involving many integrated language modalities - as verbal, image, sound, animation, the use of colors, typography and other resources to produce meanings. From this point of view, it´s important not to privilege a certain type of language upon another.

Since the effects that each element involved into an interactive production is noticed, the interface development process become more conscious, contributing for the elaboration of better virtual spaces. It´s important to figure out the meanings that each element can produce, considering its relation and integration as a whole inside the interactive system. The dynamic audio intensifies the processes of immersion and cognitive processing of the user, making the interaction experience more evolving. This kind of experience, enclosed by image and sound, is more complex and complete because it reaches the user in distinctive senses through a unique communication

[2] Technical term used to indicate gradual diminishing of a sound until it becomes inaudible.

object: the interface. Background songs, music, oral and writing language, images, animations and texts; all of these forms of expression – languages – get mixed on the same message.

There is a tendency in direction of the interactivity that was already documented in many areas. Consumers – especially the young – demand dynamic activities in place of passive devices. Dynamic audio has the potential to feed this demand, but if not well applied can suffocate innovation and disappoint this new and important public. Establishing a solid structure for dynamic audio can guarantee that the next generation of applications based in audio will produce results that are closer to the human cognitive model, in a way that the information presented on a interactive context will better accommodate the content´s absorption structures. This way, when systematizing the dynamic audio structures, the sounds used in the interfaces will be made in a more efficient way, and the applications would be plainly developed.

References

1. BAR-B-Q, Project. Group Report: What is Interactive Audio? And What Should It Be? The Eighth Annual Interactive Music Conference PROJECT BAR-B-Q 2003, San Antonio, USA, section 5 (December 2003), `http://www.projectbarbq.com/bbq03/bbq03r5.htm` (accessed November 20, 2011)
2. BAR-B-Q, Project. Group Report: Group Report: Providing a High Level of Mixing Aesthetics in Interactive Audio and Games. The Thirteenth Annual Interactive Music Conference PROJECT BAR-B-Q 2008, San Antonio, USA (December 2008), `http://www.projectbarbq.com/bbq08/bbq08r8.htm` (accessed November 20, 2011)
3. Collins, K.: Game Sound: An introduction to the history, theory, and practice of video game music and sound design. MIT Press, Massachusetts (2008)
4. Cavalcante, M.C.B.: Mapeamento e Produção de sentido: Os links no hipertexto. In: Marcuschi, L.A., Xavier, A.C. (eds.) Hipertexto e Gêneros Digitais: Novas Formas de Construção de Sentido, 3rd edn., Cortez, São Paulo (2010)
5. IASIG; Interactive Audio Special Interest Group. Functions of Game Audio [s;l] (2012). Disponível em: `http://www.iasig.org/wiki/index.php?title=Introductios` (acesso em January 12, 2012)
6. Menzies, D.: Scene Management for Modelled Audio Objects in Interactive Worlds. In: Proceedings of the 8th International Conference on Auditory Display, July 2-August 5, Advanced Telecommunications Research Institute (ATR), Kyoto (2002)
7. Padovani, S., Moura, D.: Navegação em Hipermídia. Moderna, Rio de Janeiro (2008)
8. Preece, J.: Design da Interação. Bookman, Porto Alegre (2005)
9. Shneiderman, B., Plaisant, C.: Designing the User Interface: Strategies for Effective Human-Computer Interaction, 5th edn. Addison Wesley (March 2009)

A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters

Birgit Endrass[1], Lynne Hall[2], Colette Hume[2], Sarah Tazzyman[2],
and Elisabeth André[1]

[1] Human Centered Multimedia, Augsburg University,
D-86159 Augsburg, Germany
{endrass,andre}@hcm-lab.de
[2] University of Sunderland,
Sunderland, SR6 0DD, UK
{lynne.hall,colette.hume,sarah.tazzyman}@sunderland.ac.uk

Abstract. In this paper, we outline the creation of an engaging and intuitive pictorial language as an interaction modality to be used by school children aged 9 to 11 years to interact with virtual characters in a cultural learning environment. Interaction takes place on a touch screen tablet computer linked to a desktop computer on which the characters are displayed. To investigate the benefit of such an interaction style, we conducted an evaluation study to compare the pictorial interaction language with a menu-driven version for the same system. Results indicate that children found the pictorial interaction language more fun and more exciting than the menus, with users expressing a desire to interact for longer using the pictorial interaction language. Thus, we think the pictorial interaction language can help support the children's experiential learning, allowing them to concentrate on the content of the cultural learning scenario.

Keywords: Interaction Design, Interaction Modality, Virtual Agents, Culture, User Experience.

1 Introduction

While traditional learning systems provide conventional interaction devices such as mouse and keyboard, especially for child users intuitive interaction is important to provide an engaging experience. Menus provide bound and restricted interaction choices, possibly limiting the user's perceived freedom in their interactions. Free text input can be desirable, however, due to technical constraints such as limited support of vocabulary and grammar this is hard to realise. Further, children's keyboard skills are often not fully developed compared to adults, reducing children's abilities to express themselves. Recent paradigm shifts towards more natural user interfaces, based on either direct touch or three-dimensional spatial interaction [1] provide interesting alternatives to increasing user engagement, particularly for children. One of the most often stated benefits is the

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 532–543, 2014.

view that interacting with an application through directly touching graphical elements is a more "natural" or "compelling" approach than working indirectly with a mouse or other pointing devices [2].

In this paper, we investigate a game play interaction modality designed for, and with children. We present a pictorial interaction language using touch-based gestures on a tablet computer that allows children to interact freely with characters displayed on a different screen. The interaction is developed for use in playing a serious game in which children communicate with a virtual character to learn about resolving a cultural conflict.

2 Background

This paper focuses on the development and evaluation of a pictorial interaction language for children aged 9 to 11 years. This interaction modality is part of the eCute project [3], which aims to create and encourage cultural awareness among children.

In the MIXER showcase [4], the user plays the role of an invisible friend to provide advice and support to a virtual character, called Tom. The narrative of the MIXER application centres on Tom visiting a summer camp where he meets a group of characters that he knew before. With this group, Tom plays a game called Werewolves (see [5] for a description of the rule set). In the game, each player is assigned a role, as either a werewolf or a villager. The aim of the game is to deduce which character in the group is the werewolf, before the werewolf kills all of the villagers. Several times during the game, Tom asks for the user's advice. After playing for a while, Tom leaves the first group of children and meets a different group that he has not met before. In this group, Tom and the user are confronted with crucial changes to the rule set by which the game is played; this leads to a critical incident and a potential conflict situation.

To create a novel, engaging and effective learning experience we aim to develop an interaction modality that is both intuitive and engaging for children of the target age group. A pictorial interaction language was identified as a solution to the problem of creating a novel and universal interaction experience.

3 Related Work

We think that finding novel and intuitive interaction modalities for educational systems is a crucial task. Sali and colleagues [6] investigated three different dialogue approaches for game interfaces. They found that users prefer a natural language interface over interfaces that allow users to select sentences and interfaces that make use of an abstract response menu interface. However, some users had problems with the natural language interface because they found it hard to figure out what to say in a particular situation. Compared to adults, this problem may be magnified for children. We encountered similar issues in former work [7] where children interacted with a virtual learning environment using typed text input. The interaction choice for natural language input resulted in several problems including recognition problems for the software coupled with the difficulty and time required for children to express themselves in typed text. We think that by using a pictorial interaction language the disadvantages of text input are reduced, whilst retaining a large degree of interaction freedom.

Pictorial languages are commonly used with children in the field of augmentative and alternative communication, e.g. in communication training for autistic children [8], [9]. Widget symbols (e.g. [10]) also find their usage on websites that provide understanding and communication for people who find reading text difficult, e.g. [11]. Their potential for intuitive communication is gaining ground for non-disabled children as well. For example, a pictorial language is used on CBBC (children's television channel) in the UK to facilitate communication. We thus see great potential in using a pictorial language as an intuitive interaction modality to communicate with virtual characters in learning environments as well.

Researchers have found that in the field of human computer interaction using a visual style of expressing oneself helps to motivate children to complete creative and challenging tasks, such as telling stories [12], or learn computer programming using storytelling environments [13].

To overcome the language barrier in inter-cultural communication, Takasaki and Mori [14] describe a communicator that was developed for children of different cultural backgrounds to be able to talk to each other using pictogram communication. This was reported to be an effective and practical user interface design method with children. In a similar manner, we aim to design a pictorial language for children to enable communication with a virtual character.

4 Design and Realisation

With the intention of improving both engagement and user experience for 9-11 year olds, we use an Apple iPad as the interaction device in combination with a pictorial language as interaction modality, provided as an extension to a desktop-based system connected via Wi-Fi.

4.1 System Setup

Figure 1 shows an overview of the setup including a child using it. The user can observe what happens in the virtual environment on the screen of the desktop while interaction takes place on the tablet computer. In this way, information that should only be visible to the user is shown on the tablet computer, while the environment with the virtual characters is visible for everyone. This supports the impression of being an invisible friend whose actions cannot be seen by the other characters involved in the gameplay.

As the focus of the present study was to test the suitability of the pictorial interaction language, it was tested with an early version of the MIXER game holding a virtual friend character that is involved in a fictive game with a group of characters (running in the AAA application [15]). During the game, the friend character asks the user for advice several times. In each case, the character leaves the group, comes closer to the screen and updates the user on what happened in the game. Depending on the context of the question, different icons are provided on the tablet computer to construct the answer message in a pre-structured "grammar", by e.g. combining an action and an emotion. After hitting the send-button, the friend character reacts to the message and returns back to the group of other characters.

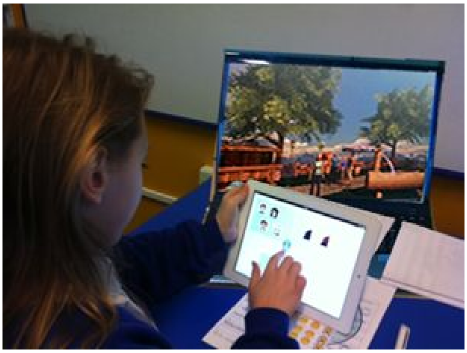


Fig. 1. Example setup, with a child using the pictorial interaction interface on an iPad

4.2 Interaction Modes

We designed two different advisory modes for interaction with the virtual friend character in the MIXER game:

- "During game advisory mode" to support the friend character during game-play;
- "Critical incident advisory mode" to deal with the critical incident after playing with a different group of characters that play the same game with different rules.

In this paper, only the advisory modes that occur during the game were investigated. Therefore, we identified four different advisory modes that describe standard situations for the Werewolf game: (1) Questioning who the werewolf is, (2) Reasoning why somebody is the werewolf, (3) Reacting if somebody else is being accused, (4) Reacting if oneself is being accused.

Depending on the context of the game, the advisory modes can either be used alone or combined to simulate a longer conversation between the user and the friend character. For example, after a character has been accused of being a werewolf (3) the friend character could ask who else could be a candidate (1) and why the user thinks so (2). Interaction is managed in a question and answer style, with the friend character asking for advice and the user answering by constructing a message.

4.3 Vocabulary Selection

We had to create a language that would fit our purpose of communicating with a virtual character that was playing a game of werewolves with other virtual characters. As this was a very specific requirement, we could not, for example, acquire a set of validated open source icons. It was necessary to create and test our own icon set. The first stage in the creation process was to investigate the language used while playing the werewolve game. Taking into account the rules of the game, some words were obvious, such as 'You', 'They', 'Accuse', 'Defend', 'Werewolf' etc. We recruited a total of 70 children (aged 9 to 11) who played the real world werewolf card game in small groups. The games were recorded and transcribed. In total, we identified 60 frequently used words, such as "I **accused** her because she **looks suspicious**" or "he's being too **quiet**". These words were later grouped, e.g. emotions or actions and structured in a way to match the identified interaction modes.

4.4 Icon Design

Besides following the design standards of ISO/IEC 11581 by using a consistent size, icon behaviour, and a similar design for all of our icons, a challenge was to design the icons to be sufficiently intuitive for children to construct meaningful messages.

In total, a set of over 60 icons was required for the pictorial interaction language based on the study mentioned above. The majority of the icons show a small green character. This character was shown in different positions to convey the different action states that were identified. For example, for the word 'calmly' the character was shown in a meditative pose. The colours green and red were used in the icons to convey positive and negative respectively. The icons were designed intuitively, by using what seemed to be the most appropriate visual representation for children of each word. However, what is obvious to a team of researchers is clearly not always going to be obvious to a child. Thus, we conducted a small study with 30 children to test their comprehension of the icons. We began by introducing and playing a short game of werewolves, which gave the children a context in which to discuss the meaning of the icons. The children were given activity sheets with pictures of the icons, and then asked to think about the game they had played and to try and work out what each of the icons meant. Following the game we held small focus group activities during which the children were shown the icons again and asked to discuss what the icons represented. This gave the research team useful qualitative information about children's views of the icons and their design. The icons that were not easily understood by the children became part of the next activity in which the children themselves helped to redesign the icons. These were used to develop the final icon set. The focus group activity was repeated with the final icon set with a further 25 children at a different school, during which all icons were successfully identified by the children. Figure 2(left) shows a small subset of the icons designed for our pictorial interaction language along with their intended meanings.

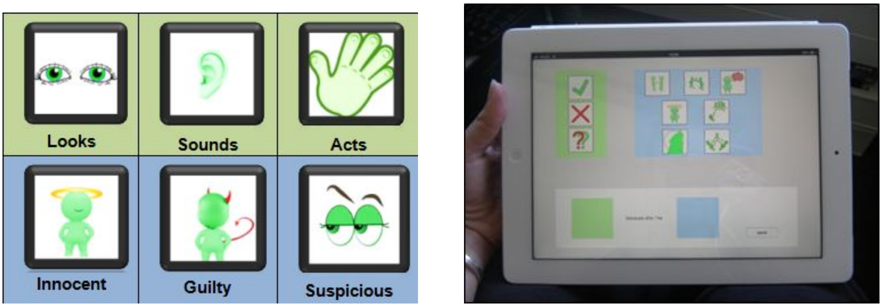


Fig. 2. Left: Example icons with intended meanings; Right: Interaction interface on the iPad

4.5 Interface Design

For the interface shown on the iPad, groups of icons are provided, while one icon of each group can be selected to form a sentence in a pre-structured grammar, e.g. by combining an action with an emotion. Figure 2 (right) shows the iPad with an interactive screen of the third advisory mode. Different coloured views contain the different groups of icons. Icons are moved by touching and dragging them across the screen. The white area on the lower part of the screen holds the message that the user constructs, providing empty views of the same colour of the group of icons that can be selected. The simple colour code helps the user understand that an icon of each provided group should be selected and moved to the corresponding area. Additionally, icons are automatically attached to the correct position (centre of same coloured area) as soon as they are moved into the user sentence area. In case an icon was selected for that area before it is replaced and the former is popping back to its initial position. Thus, only well formed sentences can be produced by the child, not allowing grammatically incorrect or nonsense sentences that would be uninterpretable for the system.

An example of a standard situation in the Werewolf game includes questioning why another player might be a werewolf (2). To answer this question, an action and a reason can be combined by the user. To help the user understand what kind of answers can be created, the message is initialised by the words "because he / she", followed by two different coloured views relating to actions and reasons respectively. Using the icons shown in Figure 2 (left), messages such as "because he/she looks guilty" or "because he/she acts suspicious" could, for example, be constructed.

5 User Study

To test the possible benefit of a pictorial interaction language over traditional interaction modalities, a user study was conducted with 9-11 year olds.

5.1 Interaction Modes

For the present study, we implemented two interactive versions of our system both using the touch-based interaction on the iPad: icon-based vs. menu-based. The icon-based version contains the pictorial language described above. The menu-based version was implemented to provide a set of choices in text form, representing choices that could also be constructed with the pictorial interaction language, which can be selected by the user by clicking on them (see Figure 3 for comparison of the two iPad interfaces).

The setup of the game is constant for both versions. In each case the friend character repeats the choice of the user, comments on it and returns to the group. However, the characters' comments are limited to the number of choices in the menu-based version to ensure that users are not influenced by the wider variety of the system in the icon-based version.

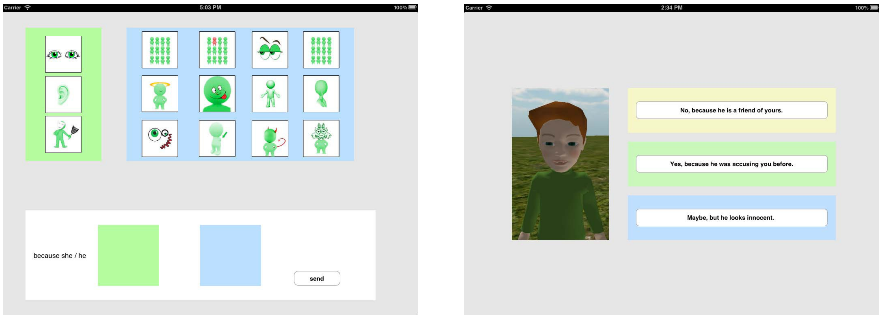

Carrier
5:03 PM
100%
because she / he
send
Carrier
2:34 PM
100%
No, because he is a friend of yours.
Yes, because he was accusing you before.
Maybe, but he looks innocent.

Fig. 3. Screenshots of the iPad showing icon-based interaction (left) and menu-based interaction (right)

5.2 Expectations

With the pictorial interaction language, we provide interpretational freedom to the users by offering many opportunities to construct sentences. In addition we want to inspire children's curiosity and exploration. For our study, we hypothesized that an icon-based interaction style would be perceived as more engaging and interesting compared to a traditional menu-based interaction.

However, a possible advantage of the menu-based version might be that it is more intuitive to use for inexperienced users, since fewer, clearer choices are provided and there is no need for children to construct their own sentences.

5.3 Procedure

To investigate which interaction modality the children preferred, an evaluation study was conducted with the target age group with each child having a PC, iPad and headphones. Children were introduced to the study activities, before playing both versions of the system. After using each of the versions, the children completed a questionnaire, then used the other version and completed the same questionnaire again.

For evaluation, a 4-part questionnaire was developed:

Part 1 provided requested descriptive data, e.g. the children's age, gender and previous exposure to tablets.

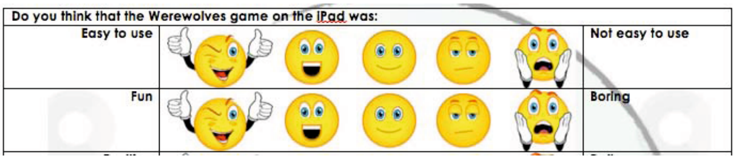
Part 2 included questions focused at gaining the child's response to their first interactive experience. Each question was provided as bi-polar adjectives using a 5-point facial scale, see Figure 4. Facial scales have been shown to be well suited for evaluation with children in school environments,see [16]. The questions included:

- Ease of use: was the application easy to use or not, and could the child achieve what they intended with the interaction
- Engagement: was the experience fun, was it exciting, would they want to play again, would they have liked to play longer
- Visual appeal and interaction comprehension: did children like the appearance of the interface and could they understand the meanings provided in the interaction dialogue (e.g. the menu items or the icons)
- Open questions asking what children liked most and least about the game

Part 3 repeated the questions in Part 2 for the second interaction experience.

Part 4 asked the child to compare the two interaction modalities and decide which had been more fun, exciting and interesting.

Finally children were given a gold star sticker and were asked to put the sticker on a picture of the version they liked the best. The gold star sticker was chosen as children recognise stars and stickers as tokens that are awarded for something that is very good i.e. stars and stickers are often given in class for a good piece of work.



Do you think that the Werewolves game on the iPad was:

Easy to use		Not easy to use
Fun		Boring

Fig. 4. Example questions from the questionnaire

71 children aged 9-11 years (M = 9.65, SD: .56) living in the UK participated in the study. 59.2% (n = 42) of the sample was boys, and 40.8% girls (n = 29). Most children had used an iPad before (84.5%, n = 60).

35 of the children used the icon-based application first, followed by the menu-based version, while the remaining 36 children played the versions ordered the other way round (i.e. the procedure was counterbalanced to avoid order or practice effects).

6 Results

Mean values of the children's ratings are summarized in Figure 5. It shows that overall the icon-based version was rated more positively compared to the menu-based version. There was one exception to this for ratings of the pictures on the iPad being easy to understand / hard to understand. For this question, children rated the menu-based interaction slightly more favourably (i.e. the menus were easier to understand).

Figure 5 also clearly illustrates that children rated all the questions positively for both the menu and icon-based interaction, with no mean ratings above 3 (scale ranged from 1 to 5, with 1 being most favourable and 5 the least favourable). The highest (i.e. least favourable) mean response of 2.61 was for ratings of how exciting / dull the menu-based interaction was.

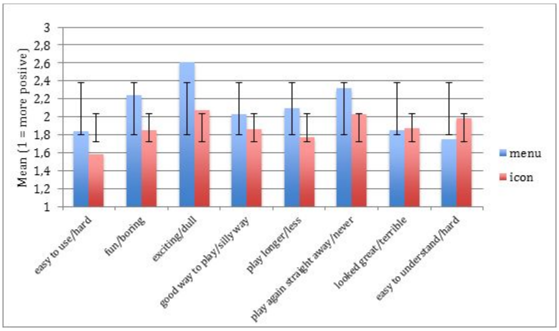
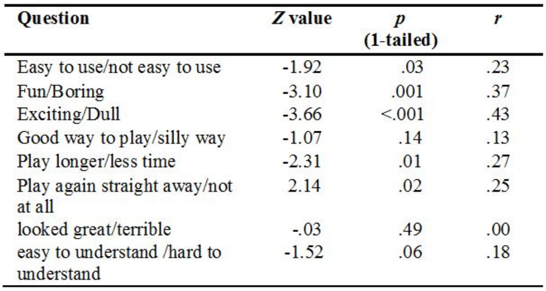


Fig. 5. Mean ratings (error bars 1 SD) of icon-based vs. menu-based interaction

The facial scale that children rated could not be assumed to follow an interval scale, but rather ordinal. Therefore, non-parametric Wilcoxon signed-rank tests for related samples were calculated to determine whether there were any significant differences in mean ranks between children's ratings of the menu-based versus icon-based interactions. Figure 6 illustrates the Z statistic, associated



Question	***Z* value**	***p* (1-tailed)**	***r***
Easy to use/not easy to use	-1.92	.03	.23
Fun/Boring	-3.10	.001	.37
Exciting/Dull	-3.66	<.001	.43
Good way to play/silly way	-1.07	.14	.13
Play longer/less time	-2.31	.01	.27
Play again straight away/not at all	2.14	.02	.25
looked great/terrible	-.03	.49	.00
easy to understand /hard to understand	-1.52	.06	.18

Fig. 6. Interaction questions for the menu-based and icon-based interaction, associated Z values, significance (1-tailed), and effect size (r)

significance for one-tailed tests, and effect sizes (r) for each of the questions in our questionnaire.

Children rated the icon-based version (median = 1) as slightly easier to use then the menu-based interaction (median = 1). The icon-based interaction (median = 2) was also rated as more fun to use compared to the menu-based interaction (median = 2). Children rated that the icon-based interaction (median = 2) was more exciting compared to the menu-based interaction (median = 3) and wanted to play longer with the icon-based (median = 1) interaction compared to the menu-based (median = 2) interaction. The icon-based interaction (median = 2) was rated more favourably by children for wanting to play again straight away compared to the menu-based interaction (median = 2).

No significant differences were found between the icon (median = 2) and menu-based (median = 2) interactions for ratings of whether it was a good way to play the game. Children rated the design of the interface (looked great/looked terrible) on both the menu-based (median = 2) and icon-based interaction (median = 2) favourably, with no significant differences reported between the two interactions. Children also found the interface for both the menu-based (median = 1) and icon-based (median = 1) version easy to understand, with no significant differences. Figure 7 illustrates the positive (icon ¿ menu) and negative (icon ¡ menu) ranks derived from the Wilcoxin tests for each of the eight questions that children rated using the facial scale. The figure clearly demonstrates that children favoured the icon-based interaction over the menu-based interaction for all questions, with the exception of whether the interaction interface on the iPad was easy/hard to understand.

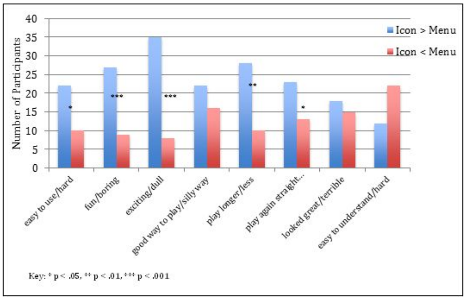


Fig. 7. Number of participants who rated the icon-based interaction more favourably than the menu-based interaction (icon ¿ menu), and the menu-based interaction more favourably than the icon-based interaction (icon ¡ menu)

Binomial tests (0.50) were carried out to determine whether children found the icon-based or menu-based interaction more fun, exciting and interesting. Children reported finding the icon version more fun compared to the menu version [fun (Z = 5.93, p ¡ .001), menus n = 10 (.14), icons n = 61 (.86)], and the icon version more exciting [exciting (Z = -2.85, p ¡ .01), menus n = 23 (.32), icons n = 48 (.68)]. No preference for menus or icons was reported by children for levels of interest [interesting (Z = .000 p = 1.0), menus n = 36 (.51), icons n = 35 (.49)].

Each child was given one sticker and asked to place this on his/her favourite version of the interaction interface. 92% (n = 55) of children placed their sticker on the icon-based version, leaving just 8% (n = 5) of children who placed it on the menu-based version. A one-sample binomial test revealed that significantly more children said that the icon-based version was their favourite compared to the menu-based version [favourite (Z = 6.33, p ¡ .001), words n = 5 (.08), pictures n = 55(.92)].

7 Conclusions and Future Work

This paper discusses the development and evaluation of a pictorial interaction language to test the suitability of such an interaction modality for 9-11 year old children for a cultural learning scenario. We compared the experience of the pictorial interaction language with a more traditional menu-driven interaction. Through using the same application with the same interaction device we have established that children preferred the pictorial interaction, considering it to be more fun and exciting than a menu-driven approach. In line with our expectations, the children rated the pictorial interaction as harder to understand compared to the menu-driven approach but at the same time more fun. We thus think that the pictorial language provides a more challenging interaction that positively influences the overall user experience. Our focus was to investigate the potential for a pictorial interaction approach to engage children in a games-based learning experience. Through directly comparing pictorial interaction and menu-driven interfaces, even though children were positive about both approaches, results indicate that:

- Children found the pictorial interaction to be more fun than the menu-driven version.
- Pictorial interaction was perceived as more exciting than menus
- Children would have liked to play longer with the pictorial interaction than with the menu-driven system
- Pictorial interaction is well suited for, and preferred by, the age group with just 8% of the children preferring the menu-driven version.

The pictorial interaction language is now integrated into the complete cultural conflict learning experience. Studies conducted in Germany and UK with the full system will further establish the benefits of an intuitive pictorial interaction language for supporting children in developing cultural understanding and awareness.

Currently the whole system is prepared for usage with Japanese children, to investigate whether the pictorial interaction language and the serious game are understood in an Asian culture as well. Therefore, the provided grammatical structure of the interface had to be slightly adapted.

Acknowledgments. This work was funded by the European Commission within the 7th Framework Program under grant agreement eCute (education in cultural understanding, technologically enhanced).

References

1. Bowman, D.A.: 3D User Interfaces. In: The Encyclopedia of Human-Computer Interaction. The Interaction Design Foundation, Aarhus (2013)
2. Forlines, C., Wigdor, D., Shen, C., Balakrishnan, R.: Direct-touch vs. mouse input for tabletop displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2007 (2007)
3. Nazir, A., Ritter, C., Aylett, R., Krumhuber, E., Swiderska, A., Degens, N., Endrass, B., Hume, C., Hodgson, J., Mascarenhas, S.: ECUTE: DIFFERENCE IS GOOD! In: IADIS International Conference on e-Learning, pp. 425–429 (2012)
4. Aylett, R., Lim, M.Y., Hall, L., Endrass, B., Tazzyman, S., Ritter, C., Nazir, A., Paiva, A., Hofstede, G.J., Andre, E., Kappas, A.: Werewolves, cheats, and cultural sensitivity. In: Proc. of 13th Int. Conf. on Autonomous Agents and Multiagent Systems, AAMAS 2014 (2014)
5. Plotkin, A.: Werewolf: A Mind Game (2010), `http://www.eblong.com/zarf/werewolf.html`
6. Sali, S., Wardrip-Fruin, N., Dow, S., Mateas, M., Kurniawan, S., Reed, A.A., Liu, R.: Playing with words: From intuition to evaluation of game dialogue interfaces. In: Proceedings of the Fifth International Conference on the Foundations of Digital Games, FDG 2010, pp. 179–186. ACM, New York (2010)
7. Aylett, R.S., Vala, M., Sequeira, P., Paiva, A.C.R.: Fearnot! - An emergent narrative approach to virtual dramas for anti-bullying education. In: Cavazza, M., Donikian, S. (eds.) ICVS-VirtStory 2007. LNCS, vol. 4871, pp. 202–205. Springer, Heidelberg (2007)
8. Bondy, A.S., Frost, L.A.: The picture exchange communication system. Focus on Autism and Other Developmental Disabilities 9(4) (1994)
9. Mirenda, P.: Toward functional augmentative and alternative communication for students with autism. Language, Speech, and Hearing Services in Schools 34, 203–216 (2003)
10. Symbols Worldwide Ltd T/A Widgit Software: Widgit, `http://www.widgit.com/index.php` (last viewed: April 2, 2014)
11. SymbolWorld, `http://www.symbolworld.org/` (last viewed: April 2, 2014)
12. Antle, A.: The design of cbc4kids storybuilder. In: Interaction Design and Children (IDC 2003), pp. 59–68 (2003)
13. Kelleher, C., Pausch, R., Kiesler, S.: Storytelling Alice Motivates Middle School Girls to Learn Computer Programming. In: CHI 2007, pp. 1455–464. ACM (2007)
14. Takasaki, T., Mori, Y.: Design and Development of a Pictogram Communication System for Children Around the World. In: Ishida, T., R. Fussell, S., T. J. M. Vossen, P. (eds.) IWIC 2007. LNCS, vol. 4568, pp. 193–206. Springer, Heidelberg (2007)
15. Damian, I., Endrass, B., Huber, P., Bee, N., André, E.: Individualized Agent Interactions. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 15–26. Springer, Heidelberg (2011)
16. Read, J.: Validating the Fun Toolkit: An instrument for measuring childrens opinions of technology. In: Cognition Technology and Work (2007)

Tangible or Not Tangible – A Comparative Study of Interaction Types for Process Modeling Support

Albert Fleischmann[1], Werner Schmidt[2], and Christian Stary[3]

[1] Metasonic AG, Münchner Str. 29, D-85276 Pfaffenhofen, Germany
[2] Technische Hochschule Ingolstadt, Esplanade 10, D-85049 Ingolstadt, Germany
[3] University of Linz, Altenbergerstraße 69, A-4040 Linz, Austria
Albert.Fleischmann@metasonic.de, Werner.Schmidt@thi.de,
Christian.Stary@jku.at

Abstract. Many organizations loose potential for optimizing their operation due to limited stakeholder participation when designing business processes. One of the reasons is that traditional modeling methods and (interactive) tools are not suitable for domain experts who neither want to struggle with complex or formal notations, nor with the respective modeling tool. Tangible modeling interfaces are a significant move towards stakeholder inclusion. We review their respective capabilities not only with regard to modeling, but also to implementation and execution of business processes, setting the stage for improving the effectiveness of interactive Business Process Management support, and thus, stakeholder participation in organizational development.

Keywords: Tangible user interface, process modeling, model documentation, model execution, Subject-oriented BPM, multi-modal interaction.

1 Introduction

Modeling is a crucial activity for successful Business Process Management (BPM). Eliciting process knowledge of stakeholders by modeling requires adequate methods and (interactive) tools. To meet this requirement tangible modeling user interfaces have been developed, e.g., [12], complementing traditional intangible ones, e.g., [11]. Having a mix of interaction modalities allows supporting target groups with different capabilities. In this contribution we present different approaches to interactive process modeling support, and discuss their impact on the suitability for the task.

As process models serve as means for documentation, blue print for work behavior and origin of computer-based workflows we look at the consequences different styles of interaction and user interfaces have for modeling, persistent documentation, implementation and execution of business processes. Validation and optimization of business processes can be subsumed by execution, as they also require executable, thus intangible model representations. For intangible interaction several approaches have been developed, in particular for structuring the user interface of Workflow Engines (controlling the execution of process instances at runtime), e.g., [3]. However, the potential of tangible interaction styles and their recognition in terms of multimodal interactive or collaborative modeling support still needs to be explored.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 544–555, 2014.

The paper is structured as follows: After this introduction we present the framework we have used to analyze different interaction approaches. In section 3 we document the evaluation according to this framework. We conclude and sketch future work in section 4.

2 The Framework for Description and Analysis

In order to analyze different approaches to support process modeling the framework refers to their BPM background, usefulness and usability:

Way of Modeling and Methodological Background. This aspect refers to how the approach works and modeling is supported. The partially interdependent items to that respect are:

- *Nature of interaction*: What is the origin of the approach justifying the nature of interaction? It refers to the elements used for modeling, how they are utilized in the course of the modeling, and whether the interaction is driven by a specific methodical approach, such as Business Process Model and Notation (BPMN), Petri-nets, or Subject-oriented Business Process Management (S-BPM).
- *Ease of Use and Learnability*: How difficult is it to handle? What is the learning effort? The questions allow reflecting on how difficult it is to grasp the modality of interaction in the context of the supported BPM method(s). The latter could depend on the degree of formality and complexity of the notation, and expressed in terms of parameters like number of symbols, or similarity to natural language structures.
- *Stakeholder Participation*: How intense is stakeholder participation in the course of modeling? The quality of process design depends on how the approach supports involving stakeholders, leveraging their knowledge, expertise and creativity.
- *Collaboration and Distribution*: Is collaborative and distributed modeling possible? This question aims on how the approach supports joint designing of processes by modelers at different locations at a time.

Sustainability of Documentation. Process models are subject to change - created models need to be stored and accessible in a straightforward way, as in agile environments they need to be adapted continuously. It is of interest how a modeling support feature, by its nature, supports preserving or converting models for future revision and reuse. This aspect is closely related to the following.

Implementation and Execution. Process modeling no longer aims at merely depicting processes graphically enabling communication and optimization. It is rather of importance to execute models minimizing transformations, whereby execution refers to both completely manual work procedures, and automatically generated workflows. The latter align IT systems with work tasks and stakeholder needs. However, approaches might differ in transforming permanent formats used for modeling into executable ones to run generated workflows.

3 Evaluation

Using the description scheme presented in section 2 we have analyzed approaches that refer to established tools, or are increasingly used in BPM practice, thus promising candidates to become common use.

3.1 Brown Paper

Way of Modeling and Methodological Background

Nature of Interaction. A brown paper approach usually consists of a pin board covered by brown paper and a set of cards of different shapes and colors, being attached to the board during modeling and complemented by hand-written annotations, arrows etc. (for an example see http://www.metaplan.us/approach/ID/34). The elements can be used for many purposes like brainstorming, domain structuring or modeling of business processes.

The brown paper approach is not bound to a specific process modeling method or language. Modelers can use any notation for which they define the semantics of cards or drawn symbols representing the language elements, e.g., Event-driven process chains (EPC), Business Process Model and Notation (BPMN), Subject-oriented Business Process Management (S-BPM).

Learnability and Stakeholder Participation. Brown paper modeling is technically easy. There is no tool overhead, people just need to label symbol cards, pin them to the board and eventually use a marker pen to add information. The selected modeling language determines the learning effort users as business domain experts need to invest to be able to design processes. The effort increases with the number of language elements and the freedom of use. A list of modeling conventions might help, but cause overhead for modelers. Providing different cards representing key symbols also could help, but it does not reduce complexity. The latter can be achieved by limiting the language vocabulary to subsets of elements. However, they could cause interoperability problems with computer-based modeling tools which process different subsets, and with the transformation to executable procedures.

Due to its ease of use the brown paper approach applied by a well-trained moderator enables intensive stakeholder participation, and is only limited by applying it in specific method context, depending on the complexity of the method.

Collaborative and Distributed Modeling. Collaborative modeling is possible. 5-7 participants work together at one pin board. There are examples for virtual and digital moderation with virtual boards via video or web conferencing (see for example http://metaplan.de/moderation/), the applicability of the traditional brown paper approach for distributed modeling is poor though.

Sustainability of Documentation. The common way to save results from modeling is a photo protocol, with the models being stored as images (sometimes the brown papers are kept, too). Reuse is only possible by displaying or printing the images.

Alternatively, the wall papers can be transformed to computer systems, using the graphical user interface of modeling software (see section 3.4), taking the risk of errors when redrawing them.

Implementation and Execution. Brown paper models neither can be executed automatically, nor serve as work descriptions for manual execution of processes. In order to implement and execute them they need to be redrawn using modeling software (see section 3.4).

3.2 Tangible BPM

Way of Modeling and Methodological Background

Nature of Interaction. T-BPM stands for Tangible BPM (see [7, 8], www.t-bpm.de). It is based on the Business Process Model and Notation (BPMN) standard 2.0. The major elements of the interface are four different building blocks representing activities, events, documents and gateways as notation elements of BPMN. The modelers label building blocks using erasable whiteboard markers and put them on a table in order to lay out a process. Edges between elements are also drawn with markers. Elements can easily be removed and relabeled.

Learnability and Stakeholder Participation. Modeling with T-BPM is technically easy, but complexity increases according to the extent BPMN language elements are used. Although there are only four shapes to remember for modeling, the user needs to know a lot more about BPMN to use them correctly, e.g., the standard offers 8 types of gateways and more than 60 types of events (see http://www.bpmb.de/index.php/BPMNPoster and [11]). Hence, modelers first need to identify the right type and mark the building block, respectively, e.g., as a throwing event, caused by a message, before they can label it with process-specific information, e.g., invoice sent. If they do not specify the type precisely, the process description might be not sufficiently detailed for later use, e.g., for developing software. As a consequence, domain experts need at least basic know how of BPMN, and likely the support of a method expert to capture complex situations.

Similar to the brown paper approach, stakeholders can easily participate in modeling because the technique is technically easy-to-use. The haptic experience with the movable building blocks motivates and helps lowering barriers. However, the complexity of BPMN, if used comprehensively, causes higher cognitive effort for users, or requires a method expert guiding the design process.

Collaborative and Distributed Modeling. T-BPM offers several possibilities to collaboratively model processes in a group of 5-7 people around a table. Distributed design of process parts can be organized using several tables at different locations. Integrating the parts and coordinating the interfaces afterwards may cause high effort though.

Sustainability of Documentation. T-BPM models laid out on the table can be photographed and stored as images like brown paper processes. In order to have them stored in formats that make further electronic processing possible they need to be put into modeling software (see section 3.4).

Implementation and Execution. Implementation and execution conditions of T-BPM are similar to those of the brown paper approach (see section 3.1.3).

3.3 Comprehand

Way of Modeling and Methodological Background

Nature of Interaction. Comprehand is a tabletop interface that provides a digitally augmented modeling surface and graspable color-coded building blocks [9] [12]. People can model a process by placing them onto the table. Different from T-BPM, all movements and positions of the elements are filmed by a video camera from below. Using ReacTIVision and JHotDraw, the results are instantly interpreted, projected by a video beamer from below and displayed on an auxiliary screen. Building blocks can be labeled via a computer keyboard and connected by just touching each other. Again the system immediately shows the results of the respective user interaction on the table screen, like labeled blocks and arrows linking them. While most of the modifications are enabled by physically repositioning, removing or adding elements, other tangible tools like an 'eraser' serve to remove connecting lines from the screen. Figure 1 depicts the Comprehand interface.

The technology in general is open for any modeling language once their semantics are assigned to the building blocks. For capturing processes it has been configured to support modeling according to the subject-oriented approach. This approach captures both, the interaction of process participants, which orchestrates their collaboration, and their individual behavior, which describes the way they contribute to accomplishing a process. The modeling method is based on natural language structure with subject, predicate and object. Its graph-based notation gets on with only five symbols for representation: subject, message (including business objects), and the three action types do, send and receive [4]. Once modeling blocks are available for these concepts, no additional type specification is necessary, e.g., compared to T-BPM. One more reason to tailor the interface for S-BPM is the method's capability to automatically generate executable code from the model (see section 3.3.3).

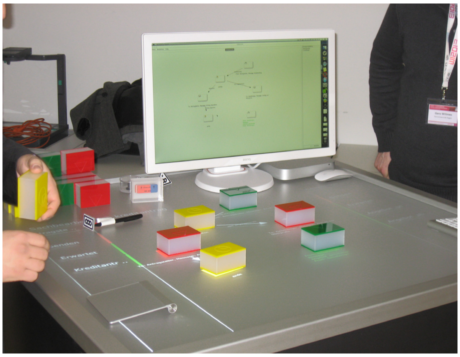


Fig. 1. Comprehand tabletop interface (see also `http://www.metasonic.de/touch`)

Learnability and Stakeholder Participation. Modeling with the interface is easy and does not require lengthy introduction. Typically, users quickly figure out how they need to apply the building blocks and the whole setting. In combination with the method properties described above the environment empowers domain experts to intuitively express their process knowledge in a straightforward way, without being hampered by some tool or method overhead.

S-BPM, and thus, the S-BPM instance of Comprehand, explicitly includes the stakeholders as it starts describing the process from their perspective when modeling their work (subject behavior) as a sequence of actions (function state), sending messages to other participants (sending state) or receiving messages from others (receiving state). The graspable modeling tools increases their motivation and fosters the focused elicitation of their knowledge [5]. In this way, stakeholder participation is facilitated, an objective often articulated in the context of Social BPM [1, 2, 10].

Collaborative and Distributed Modeling. The digitally augmented tabletop interface allows subject representatives to jointly model subject behaviors and interactions. It supports a variety of scenarios for collaborative and distributed modeling, such as the one detailed in the following (for further details see [9]): Initially, all participants involved in a process are assigned to a part of the surface for modeling subject behavior. To specify an interaction a message element is placed on the subject space, named, eventually annotated, and then moved to the area of the receiving subject (see left part of figure 2). Once the message exchange is completely captured, each representative of the different subjects model his/her respective (individual) behavior step-by-step, using the entire surface and placing state elements for function, sending or receiving states as required for task accomplishment (see right part of figure 2). Message ports for all other subjects serve to create receiving or sending states by placing a state element on incoming or outgoing messages shown in the ports and then dragging them to the desired position in the behavior model. The set of available messages has been defined in the previous step, i.e. interaction design. The system always tracks what happens on the table and displays the representation on an additional screen. Similar to the brown paper and T-BPM approach 5-7 people are a reasonable size for groups working at a single table.

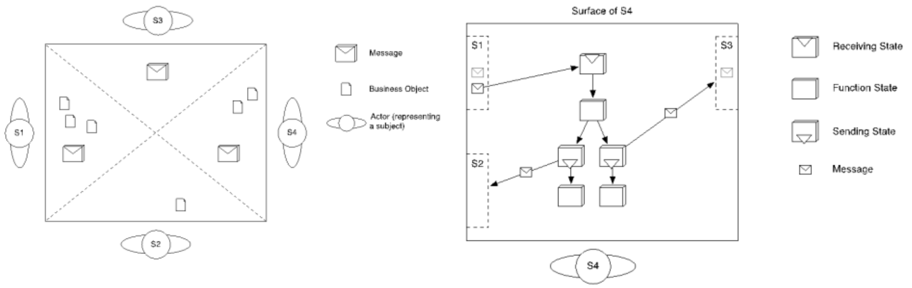


Fig. 2. Modeling surface for subject interaction (left) and behavior (right) – see [9]

Another set of typical use cases is based on distributed or co-located multiple tabletops. Using multiple tabletops facilitates both the modeling of subject interaction and the simultaneous or asynchronous behavior development for various subjects, either at the same place or spatially distributed (see fig. 3). For that purpose tables can be interconnected via a communication network. Synchronization is handled by an XMPP server, connecting to model clients and/or computer-based communication channels (like social networks, chat or videoconferencing) (see section 3.4). In this scenario representatives of each subject work at a specific table and can instantly notice incoming messages from subjects being modeled at other tables. They then can design the subject behavior at hand according to the actions they consider adequate following the receipt of those messages. Crosswise they can include sending states into their behavior specification causing message transfer via the ports to other subjects in order to trigger their reactions.

In any scenario the participants can discuss and negotiate while designing interaction and behavior, either face-to-face or via electronic channels. Due to the auxiliary display they can instantly follow their modeling procedure, and thus, check the results of their own design work in line to those of other subject representatives. In addition, they might check the overall coherence of the model based on the overall set of specified subject interactions. As each stakeholder can experience his/her behavior locally and from the overall organizational perspective at the same time, such an approach increases the probability to come up with an agreed-upon and well-accepted process model. In this way, effects like spamming co-workers (subjects) through broadcasting information become transparent immediately, and can be handled at design time. In case of S-BPM design time encapsulates build and runtime when executing SBDs automatically.

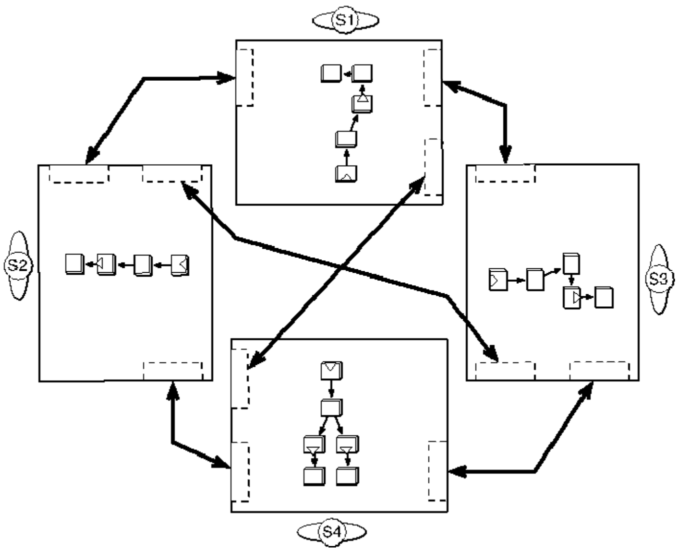


Fig. 3. Multi-surface setup for distributed modeling – see [9]

Sustainability of Documentation. For documentation the interactive system offers two options, as described in the following. As the system records all model states filmed and interpreted during the design phase in a repository, it facilitates user support for physically reconstructing former versions of current or previous sessions. The software, e.g., indicates where to put shapes on the surface, in order to rebuild a recorded state of the tangible model.

In order to persistently save models and make them available for electronic processing the system provides an automatic transfer facility. It transforms model information into the internal format of available modeling software following S-BPM, e.g., the Metasonic suite (www.metasonic.de). This transformation does not lead to any reduction or loss of information. Hence, there is no (manual) effort required for processing, as for the brown paper and the T-BPM approach.

Implementation and Execution. The models built with the tangible user interface described above and automatically transferred to the S-BPM software environment can be elaborated, validated, implemented as a workflow and brought to execution without programming. Enabler of the latter feature is the correspondence of the S-BPM modeling language to a process algebra with a precise formal semantics. It allows automated code generation and makes subject-oriented process descriptions executable and in this way, empowers process stakeholders to instantly validate the model and model changes without having IT specialists involved [4].

Variations. Technically less sophisticated tangible modeling support tools, also based on S-BPM, are Rural Comprehand and Buildbook. The first works with (magnetic) cards to be laid out on a surface and drawing lines to connect them when constructing diagrams. Buildbook consists of a letter case, representing a subject, and color-coded plug-ins encoding function, receiving, and sending states of the S-BPM notation as well as edges [6]. In both cases modeling results can be saved as photos, and transformed to the modeling software via image processing. From that point on, further processing is possible as described in section 3.3.3.

3.4 GUI-Based Modeling

Way of Modeling and Methodological Background

Nature of Interaction. Graphical User Interfaces (GUIs) are part of software tools that provide modeling features according to the implemented BPM method. The variety ranges from drawing tools like MS Visio, e.g., mainly offering stencils for EPCs or BPMN, to systems tailored ones for one or more particular methods, sometimes part of a business process management suite, also including a workflow engine. Examples are the ARIS Toolset (EPC, org charts etc.), Tibco, bizagi, Signavio (all mainly focused on BPMN), and Metasonic Build (S-BPM). In the following we do not further consider pure drawing tools.

In most cases design tools are used by method and tool experts who model processes according to information they have obtained before, either in interviews and workshops with process participants (domain experts). The results are presented to others, discussed and modified in follow-up sessions.

Learnability and Stakeholder Participation. Modeling requires sound literacy of the implemented method and software tool environment. As mentioned before, ease-of-use is closely related to the complexity of the method, a fact that often limits active participation of stakeholders [1, 10].

Collaborative and Distributed Modeling. As modeling is performed at a computer-based work place usually a single user is involved. Jointly working on models is limited to workshop settings where the modeling screen is projected on a wall, so that participants can discuss what they see, and guide the modeler developing the process on the computer system. GUI-based approaches only support distributed modeling when using corresponding virtual communication and information sharing spaces. There are approaches to leverage social software (communities, micro blogs, chat etc.) in order to jointly design processes (e.g., ARIS Connect), a model can only be manipulated at a specific location at a time.

Sustainability of Documentation. Software-based modeling tools support storing models and artefacts in databases and repositories, according to internal formats, and thus provide access for future (re-)use.

Implementation and Execution. In many cases modeling results can be printed out and exported in HTML format, and in this way, serve as paper-based or online process guidelines for manual execution. It depends on the applied method and utilized tool, in how far implementation and execution as computer-controlled workflow is possible and a straightforward task. For instance, bringing ARIS EPCs to execution requires programming, e.g., using Business Process Execution Language (BPEL). The extent to which BPMN can be automatically mapped to BPEL and interpreted by a process engine at runtime depends on the match of the BPMN subsets of the modeling and execution component. There might differences from case to case, although BPMN is defined as a standard. Many software vendors only support a subset of the standard, mainly not identical, causing interoperability problems and additional transformation effort.

S-BPM models are executable on the fly (see section 3.3.3) - the Proof component of the S-BPM suite (www.metasonic.de) enables stakeholders experiencing and iteratively improving the modeled work procedures in a computer-based role play without having IT experts involved so far. Once the validation is finished the model can be implemented into an organization, and finally, executed by the Flow component to handle process instances in daily business. Programming is only necessary if it comes to the integration of existing applications into the workflow, such as ERP systems.

3.5 Comparative Analysis

Although the presented approaches reveal a variety of formats and modalities, they are closely coupled to the method context, either encoded in software or tangible hardware, or being part of skill deployment or social facilitation. Figure 4 depicts the main results of the comparison. The table structure takes into account the type of interaction and relevant phases of BPM and required process support (application context).

The table entries reveal media discontinuity of the tangible modeling approaches with respect to seamless processing of modeling results, except for those based on the S-BPM approach. The latter allows the combined use of different modalities to create a coherent business model. The model representations, even when accomplished in distributed sessions, can be integrated, elaborated if necessary, and brought to execution, all with minimal manual effort.

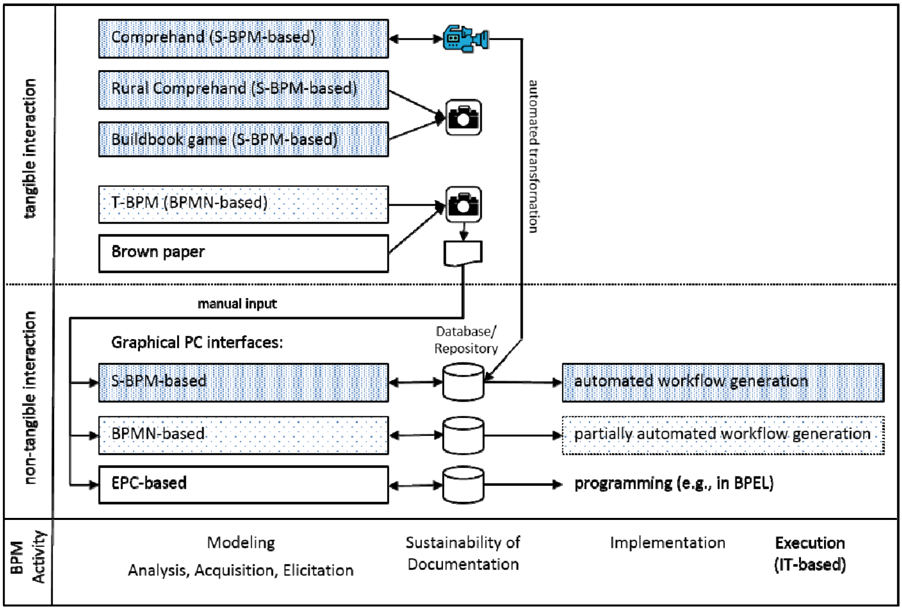


Fig. 4. Tangible and non-tangible modeling support

4 Conclusion and Future Work

As stakeholder participation turns out to be essential for business processes, modeling methods and (interactive) tools need to become suitable for domain experts. They should keep notations and utilization of features simple. In this contribution we have questioned the usefulness and usability of tangible modeling approaches as they promise a significant move towards stakeholder inclusion. We have reviewed their respective capabilities with regard to modeling, implementation and execution of business processes.

When contrasting tangible interfaces with non-tangibles they turn out to be beneficial in case the major modeling purpose is the description, explanation, discussion, and development of mutual understanding of business procedures. In case the main objective is to bring a model to execution, a seamless method approach, such as

S-BPM causes the least effort and workload. Corresponding tangible features not only offer the benefits of attracting stakeholder engagement, but also ensure coherence of representations on the organizational process level.

The support of collaborative and distributed modeling besides immediate an automatic generation of executable models reduces iterations that might be required due to different models of work. Hence, the acceptance of results, in particular achieved through coupling modeling with execution ('what you model is what you execute') can better leverage the potential for continuous improvement provided by concerned stakeholders.

As the user experience of tangible systems seems to be essential, next steps in research should be field studies of collaboration support. In terms of enhancing the palette of interaction features it might be worthwhile to look at ways to model processes also with gestures. Hereby, the S-BPM approach with its lean yet expressive notation seems to be a promising candidate for a respective approach.

Acknowledgements. The authors are grateful to the development teams of Comprehand and Touch, in particular Stefan Oppl, Matthias Neubauer, Florian Krenn, Dominik Wachholder, and Thomas Rothschädl.

References

1. Bruno, G., Dengler, F., Jennings, B., Khalaf, R., Nurcan, S., Prilla, M., Sarini, M., Schmidt, R., Silva, R.: Key challenges for enabling agile BPM with social software. Journal of Software Maintenance and Evolution: Research and Practice 23, 297–326 (2011)
2. Erol, S., Granitzer, M., Happ, S., Jantunen, S., Jennings, B., Johannesson, P., Koschmider, A., Nurcan, S., Rossi, D., Schmidt, R.: Combining BPM and social software: contradiction or chance? Journal of Software Maintenance and Evolution: Research and Practice 22, 449–476 (2010)
3. Fleischmann, A., Schmidt, W., Stary, C.: Semantic execution of subject-oriented process models. In: Kurosu, M. (ed.) Human-Computer Interaction, Part I, HCII 2013. LNCS, vol. 8004, pp. 330–339. Springer, Heidelberg (2013)
4. Fleischmann, A., Schmidt, W., Stary, C., Obermeier, S., Börger, E.: Subject oriented Business Process Management. Springer, Berlin (2012)
5. Fleischmann, A., Schmidt, W., Stary, C.: Subject-oriented BPM = Socially executable BPM. In: Proceedings of the 15th IEEE Conference on Business Informatics (CBI 2013), Workshop on Social Business Process Management (SBM 2013), pp. 399–406. IEEE Computer Society, Vienna (2013)
6. Fleischmann, C.: Subject-oriented Process Survey – An approach and modeling tool for executing subject-oriented process surveys, Diploma thesis, Vienna University of Technology, Vienna (2013)
7. Grosskopf, A., Edelman, J., Weske, M.: Tangible business process modeling – Methodology and experiment design. In: Rinderle-Ma, S., Sadiq, S., Leymann, F. (eds.) BPM 2009. LNBIP, vol. 43, pp. 489–500. Springer, Heidelberg (2010)

8. Lübbe, A.: Tangible Business Process Modeling: Design and evaluation of a process model elicitation technique, Dissertation, University of Potsdam (2011), http://ecdtr.hpi-web.de/static/books/tangible_business_process_modelling/ (last access October 12, 2013)
9. Oppl, S.: Subject-oriented elicitation of distributed business process knowledge. In: Schmidt, W. (ed.) S-BPM ONE 2011. CCIS, vol. 213, pp. 16–33. Springer, Heidelberg (2011)
10. Schönthaler, F., Vossen, G., Oberweis, A., Karle, T.: Business processes for communities. Springer, Heidelberg (2012)
11. Silver, B.: BPMN Method and Style, BPMN implementer's guide: A structured approach for business process modeling and implementation Using BPMN 2, Cody-Cassidy (2011)
12. Wachholder, D., Oppl, S.: Stakeholder-Driven Collaborative modeling of subject-oriented business processes. In: Stary, C. (ed.) S-BPM ONE 2012. LNBIP, vol. 104, pp. 145–162. Springer, Heidelberg (2012)

Body Image and Body Schema: Interaction Design for and through Embodied Cognition

Ozgun Eylul Iscen, Diane Gromala, and Maryam Mobini

School of Interactive Arts and Technology Simon Fraser University 250-13450 102nd Avenue
V3T 0A3 Surrey, B.C., Canada
{oiscen,gromala,mma99}@sfu.ca

Abstract. The interdisciplinary literature on body image/body schema (BIBS), which is within the larger realm of embodied cognition, can provide HCI practitioners and theorists new ideas of and approaches to human perception and experience. In very brief terms, body image consists of perceptions, attitudes and beliefs pertaining to one's own body, whereas body schema is a system of sensory-motor capabilities that function, usually without awareness or the necessity of perceptual monitoring. The dynamic relationality and plasticity of BIBS open up different avenues for interaction design. An overview of six main ideas deriving from BIBS literature are enumerated, followed by a discussion of projects designed for chronic pain patients that demonstrate how these ideas can be adopted in interaction design processes as a perspective or attitude rather than a mere application of traditional methods. Through bridging HCI and BIBS theories and research, we can develop a holistic framework in which we design for and through embodied cognition.

Keywords: Embodied cognition, body image, body schema, interaction design, virtual reality, chronic pain.

1 Introduction

The interdisciplinary literature on body image/body schema (BIBS), which lies within the broader realm of embodied cognition, can provide HCI practitioners and theorists new ideas of human body, perception and experience. Embodied cognition challenges the mind-body dualism and highlights how the body shapes our cognitive processing from perceiving and thinking to linguistic and emotional processing. In this context, the growing literature and research on BIBS, which draws upon diverse fields including neuroscience, psychology, and philosophy, investigates numerous aspects of embodied cognition, and directly addresses the crucial question of how mind is embodied.

The BIBS research addresses issues—such as embodiment, consciousness, awareness, attention, and agency—that play crucial roles in shaping the direction of current research on embodied cognition. In brief terms, body image consists of perceptions, attitudes and beliefs pertaining to one's own body, whereas body schema is a system of sensory-motor capabilities that function, usually without awareness or the necessity

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 556–566, 2014.

of perceptual monitoring [7]. For instance, BIBS can be described as the difference between having a perception (belief) regarding one's body (such as conscious monitoring of one's movement, or a belief about one's body's capacity to move), and having a capacity to move (the actual accomplishment of the movement). Since body schema primarily controls the interaction between one's body and environment, it enables us to function in an integrated way with our environment. Similar to J.J. Gibson's theory of affordance [8], BIBS research investigates how the body acts, moves or interacts through opportunities and constraints shaped by the dynamic interaction of body and environment. BIBS research also reveals that plasticity is involved in both body image and body schema, and that both are interdependent systems rather than mere exclusive categories. Such an insight into how the body functions in human experience reshapes our ways of thinking about our body, mind, interaction, technology and design. Therefore, the question emerges: what are the implications of embodied cognition, specifically of the growing body of BIBS literature for interaction design, including tangible interactions and non-verbal interfaces.

We draw upon six ideas about embodied cognition that are supported by the current BIBS literature: 1. Mind is embodied. We think with our bodies rather than solely relying on our brains, even though we are not always or often consciously aware of it [7][21]. 2. Proprioception plays an important role in embodiment and consciousness [7][21]. 3. The sense of self and perception of others are strongly informed by embodiment [1][7]. 4. Plasticity is involved in both body image and body schema, and they are interdependent systems, not exclusive categories [7][14][20]. 5. Such operations of BIBS do not become apparent to conscious awareness until there is a reflection on our bodily situations brought upon by certain 'limit-situations' such as pain [5][7]. 6. As body image and body schema are also shaped by pre-reflexive process, it is difficult to assess or evaluate BIBS-related issues based solely on reflective methods, such as interviews or surveys. The experimental situations need to be improved to address both the human subject's both reflective (how they express (verbalize) how they feel while moving) and pre-reflective / proprioceptive processes (how their body moves)[7][20].

We discuss some of our, The Transforming Pain Research Group's[1], projects here in order to demonstrate how these six BIBS-related ideas can be adopted in interaction design processes as a perspective or attitude rather than as a practical application of existing methods. Two of our current projects the *Virtual Meditative Walk* and *Mobius Floe* are unique, immersive virtual environments with distinctly different approaches to chronic and acute pain management. The *Virtual Meditative Walk,* for example, includes visual biofeedback and verbally coaches users to learn how to meditate. In addition, the primary navigation interface, a treadmill, was incorporated to address kinesiophia, a fear of or reluctance to move. The system is designed to directly address chronic pain patients' specific embodied conditions, bodily awareness and potentially a sense of agency that they may develop over their persistent pain. We combine mindfulness-based stress reduction (MBSR) with the technologies of VR

[1] The Transforming Pain Research Group (Pain Studies Lab) is affiliated with the School of Interactive Arts and Technology at Simon Fraser University.

and biofeedback as tools that enable patients to develop a greater awareness of their interoceptive or inner processes and share their physical conditions in more expressive ways. In contrast, *Mobius Floe* does the opposite by focusing the patient away from their bodily awareness through continuously engaging distractions, which help reduce the intensity of perceived pain [9]. This type of VR treatment - pain distraction - was developed specifically for acute or chronic pain patients who are experiencing a sudden worsening of their bodily pain, at least for the short periods of time. These and other examples offer a compelling case for how BIBS-related ideas influence and guide researchers' approaches to health-related projects. Insights from BIBS literature can be translated into a set of tangible tools and principles for creating practical HCI and interaction design projects within a holistic perspective.

BIBS research, which enables the investigation of the various aspects and conditions of embodied cognition, can contribute to HCI practitioners and theorists by offering new ideas about embodiment and interaction. If the body, and its interaction with the environment shape how we perceive our own body, self, and others, as well as the artifacts and the environments we interact with, then we need to integrate the insights deriving from BIBS literature into our design theories and practices in order to generate human-computer interactions that may be ultimately more effective, expressive and engaging. However, we also think that such connections between the fields of BIBS and interaction design necessarily imply a shift in the perspective or the attitude towards body, subjectivity, and interaction beyond applications of those ideas that are still shaped by traditional hierarchical or dichotomist understanding of the relations of mind and body. This is no small feat, as our understandings of mind as separate from body are persistent, implicit and deeply inculcated. Additionally, a holistic perspective of BIBS research can address the HCI theories and practices that focus on embodied interaction, and design for diverse communities including groups of people with varied embodied conditions or cultural sensitivities since it motivates adaptive user interfaces that lie beyond assistive technologies. Finally and conversely, we believe that HCI theories and practices can in turn enrich the research methodologies that are employed in BIBS-related theory and research. For instance, affective computing offers BIBS research new contexts for experimentation with our bodily states and interactions because it newly accounts for bodily capacities. By the integration of HCI and BIBS theories and practices, we can develop a holistic framework around specific contexts and needs, and new vocabularies and trajectories for further research. This offers a new, enlarged, lens to think with and create through, and enables concrete examples of design for and through embodied cognition.

2 Body Image/Body Schema: Unfolding Embodied Cognition

The phenomena of the human body and experience are part of complex and long-lasting interdisciplinary inquiry in diverse fields of study including cognitive science, psychology and philosophy. During the 20th century, many thinkers and researchers made the body a central issue in their works (e.g. Maurice Merleau-Ponty). Consequently, even though the dualism of body and mind (disembodied mind) haunt all claims to the contrary, we can say that the contemporary cognitive sciences offer strong examples and arguments, that prove that the mind is embodied [4][7][11][13][21]. The core argument of embodied cognition is that the thoughts and beliefs we have about the world, self and others, are grounded in our perceptual-action experience with things and the environment. It means that our bodily states, actions and interactions will shape how we perceive and think; which implies a necessity for more dynamic and interactive body-mind relations.

For instance, the studies on phantom limb reveal the effects of visual input on phantom sensations and inter-sensory effects (such as the relationship between visual and touch perception): V.S. Ramachandran‘s research show that the phantom hand was felt to move when the patient sees a normal hand being moved in the mirror (which is also known as mirror-touch synaesthesia in the phantom limb) [15]. Such findings strongly ask for a more dynamic and interactive model of brain rather than hierarchical models. On the other hand, there is extensive neurophysiological evidence of a close link between action observation and action execution. Studies on mirror neurons show that the same neurons get fired, when we both act and visually observe the same action being performed by another, since it gets translated into proprioceptive sense of that action. Simultaneously shared modalities of observation (of others) and action capability (of one's self), which occur within brain areas related to language production, emphasize their significance for intersubjective communication and understanding [16]. Furthermore, David Kirsh's study with dancers showed that using one's own body to explore a dance movement is a better way to understand dance movement than watching someone else exploring it [10]. This means that observing someone doing a certain act has a stronger impact than merely mentally thinking or imagining that act for understanding or learning it, but if someone actually performs that act, even in an imperfect ways or without completing it, it has a stronger impact than merely observing someone else doing it. As Kirsh discusses, "to fully make sense of what we are seeing we need to run our motor system simultaneously with watching to get a sense of what it would be like if were to perform the action ourselves" (6). Thus, studies of embodied and situated cognition (e.g. Suchman), which reveals its impact within design-related fields as well, support an understanding of body as embedded within physical and social environments that constitute self and perception of others and surrounding objects, and motivates action, thought, emotion and communication [7][21].

In this regard, the research on body image and body schema constitutes one of the most significant themes within the realm of embodied cognition, as it reveals the variety of ways we have of relating to and becoming aware of our bodies. In the literature, there is almost an agreement that body image and body schema address different aspects of body but are interrelated. Even though there is some variance about the nature and dynamics of these notions across various disciplines, BIBS's contribution to our understanding of embodied cognition has received widespread confirmation as an area of study of valid arguments. In basic terms, body image addresses phenomenal aspects of embodiment, which refers to a first-person awareness of one's own body and its contribution to the content of one's conscious experience [7]. For instance, patients with anorexia nervosa have a disrupted body image that cause them to perceive their own body in a way different than it actually is. Thus, we can say that body image consists of a system of perceptions and beliefs about one's body. Therefore, it can be shaped by perceptual, conceptual, emotional, cultural and interpersonal factors. On the other hand, body schema refers to prenoetic aspects of our perception, which refer to the structuring of consciousness rather than the apparent structure of consciousness [7]. This means that the notion of body schema addresses the question of how one's body shapes or constrains one's perceptual field. Body schema refers to the system of sensory-motor capabilities and tacit performances that function without our awareness or the necessity of our perceptual monitoring. Not surprisingly, the reciprocal interactions between body image and body schema are related to other fundamental notions such as agency, consciousness, attention, ownership, control and intersubjectivity [1][7].

However, as Shaun Gallagher argues, it would be wrong to describe body image and body schema along the lines of manifest versus latent, or conscious versus unconscious realms. It is better to frame the distinction as having a perception (belief) of something and having a capacity to accomplish it (to do something; or conscious monitoring of movements and the actual accomplishment of movement)[7]. Body schema functions in a more holistic way, as it is integrated with its surrounding environments or objects (such as tools and devices in a hand). This shows that body schema can extend beyond the boundaries of body image. Here, it is very crucial to understand that body image and body schema are related to one another. There is a relationality and plasticity involved in both body image and body schema: a transformation in one can derive from or yield a change in the other [7][20]. For instance, some patients compensate the impairment of body schema by employing body image in unique ways. Visual awareness of one's own body (or even visual perception of other's functioning body as in the case of V.S. Ramachandran's work with a mirror to mitigate phantom pain) can override one's body-schematic functions simultaneously. And long-term body image can function prenoetically, that is, without our awareness after learning, and so become part of our body schema. On the other hand, the body image may change based on the operations of the body schema since the body schema controls the interaction of body with the surrounding environment, and negotiates the environmental affordance in Gibson's terms. Therefore, it is important emphasize both the interrelatedness and the plasticity of body image and body schema, and that they are dynamic and relational rather than being fully given or immoveable.

However, the complex interactions between body image and body schema do not become apparent to consciousness until there is a disruption which requires a reflection on our bodily situations brought by certain 'limit situations' with which our bodies react and cope with, such as discomfort and pain [5][7][11]. For instance, patients with chronic pain may perceive and experience body and its surrounding environment differently, since their disrupted body image or impaired body schema requires more attention and effort for actions that can be done by normal subjects without conscious efforts or inattentive guarding of painful areas [5][7]. Furthermore, making an interior experience sharable through externalization or objectification, such as descriptive expressions, is a real struggle for pain patients [17]. Pain experience is very difficult to communicate. This is parallel to the fact that the objectivist or phenomenological methodologies, which rely on one's conscious experiences or reflections, fail to acknowledge the significant mechanisms underlying those experiences. Pre-reflective and proprioceptive processes shape the body image and, especially, body schema; therefore, the assessment and the evaluation of those BIBS-related issues have some difficulties need to be addressed. Furthermore, the body image and body schema are not fixed but open to transformation through new encounters with the body, environment and others, through new experience and cognition of the body, through imagination and learning [20]. This transformative quality reminds us about the significance of developing technologies or adopting existing technologies in order to address those plasticity and immediacy of the body image and body schema. Therefore, design for diverse communities, such as chronic pain patients, is not only widening the scope and inclusiveness of design practice, but also practicing of design to explore the complex phenomena of human body and experience in general; this approach in design in turn can address diverse physical conditions and cultural sensitivities.

This is a very compelling idea for the field of design; as the literature on embodied cognition including BIBS, provide empirical evidence and theoretical support for encouraging interaction design for and through embodied cognition. To cite David Kirsh's paper on the convergence of embodied cognition and interaction design, "If it is true that we can and do literally think with physical objects, even if only for brief moments, then new possibilities open up for the design of tangible, reality-based, and natural computing." (3)[10]. Parallel to this conclusion, we also highlight the significance of interaction design *for* embodied cognition, which may guide further research within the field of BIBS by addressing the complexity, immediacy and plasticity of that embodied cognition.

3 Interaction Design for and through Embodied Cognition

While we discuss about the ideas deriving from BIBS literature may enlarge HCI research, we also acknowledge that the HCI community has already accepted some of those ideas in one form or another. For instance, theories such as situated cognition (Suchman) [19], the extended mind (Clark & Chalmers) [4], and enacted perception (Noë) [13], which have considerable impact on HCI-related fields, emphasize an understanding of distributed and interactive mind rather than a disembodied one. They argue for an understanding of cognition that is continuous with the process in the environment, and with the actions we perform with others or other objects. Therefore our actions and interactions are part of our cognition; the cognition is situated, distributed and interactive. Similarly, Paul Dourish voiced and emphasized the need for an incorporation of embodiment into interaction design by making systems we build more tangible and meaningful in terms of human experience [6]. His works and his notion of 'embodied interaction' have illuminated t a different way of approaching interaction design, which is experientially, cognitively and socially grounded. In this regard, we also acknowledge Xerox PARC's longstanding contributions to the incorporation of those concerns and ideas into computing and HCI-related fields.

The recent research and practice within the subfields such as tangible interaction, haptic robotics, affective computing or wearable technologies have employed various embodied cognition-related ideas and contribute to the knowledge about our embodied capacities. However, we propose further research agendas for not only employing those ideas, or any other empirical and theoretical support from embodied cognition literature, but also for taking those ideas as departure point or end itself rather mere means to certain ends. This comes back to our emphasis on the interrelatedness of design through and for embodied cognition, with a holistic perspective, such as within the realm of body image/body schema in our case.

4 Designing for Chronic Pain Patients and BIBS

We discuss our projects here in order to demonstrate how these ideas, deriving from BIBS literature, can be adopted in interaction design processes as a perspective or attitude - as a framing or lens through which we approach body, mind, interaction, technology and design in a holistic way- rather than mere practical application nailed to more traditional methods. Therefore, it would be more appropriate to discuss our research projects as a trajectory through which we develop a responsibility to build bridges between areas that support such a holistic design perspective and practice. Our projects start with the idea that mind is embodied; and draws upon the significance of bringing the variety of sense information and experience into consciousness, and of addressing the immediacy and plasticity of our body image/body schema. As BIBS literature reveals, we sense the physical states of our body based on a variety of sense information, including proprioception, kinesthesia haptics, nociception (pain-related), temperature, and visceral sense. In our projects, we attempt to address this variety and richness of sensory information and experience through our design practice for pain patients. In this regard, we can say that our focus on BIBS within our design practice grew directly from our works and involvement within the framework of designing for chronic pain patients. As we state earlier, some ‘limit-situations’ of body, such as persistent pain, may reveal unique sense experiences through which we may explore the complexity of the phenomena of human body and experience. Designing for chronic pain patients, who have diverse embodied states, help us to set a research trajectory along the continuum that addresses the various aspects of BIBS as outlined in this paper. As BIBS ideas encourage and help us to build a holistic research trajectory, designing for chronic pain patients requires a similar perspective or attitude based on a transdisciplinary and multi-faceted research agenda, and biopsychosocial methodological framework (which simultaneously takes biological, psychological and social factors into consideration).

For instance, within our VR-based research projects, we acknowledge the significance of ‘pain self-modulation’ where the attention is directed inward rather than pain distraction based on attention being directed outward. It is a different paradigm in which we seek the ways in which we can develop ‘intraface’ by enabling a greater awareness of responsiveness and greater sense of agency based on revealing the plasticity of our body image/body schema. Evidence suggests that degree of mismatch strongly correlates to severe levels of chronic pain, and if they could be trained to better match body image with their body schema, the pain may be mitigated [14].

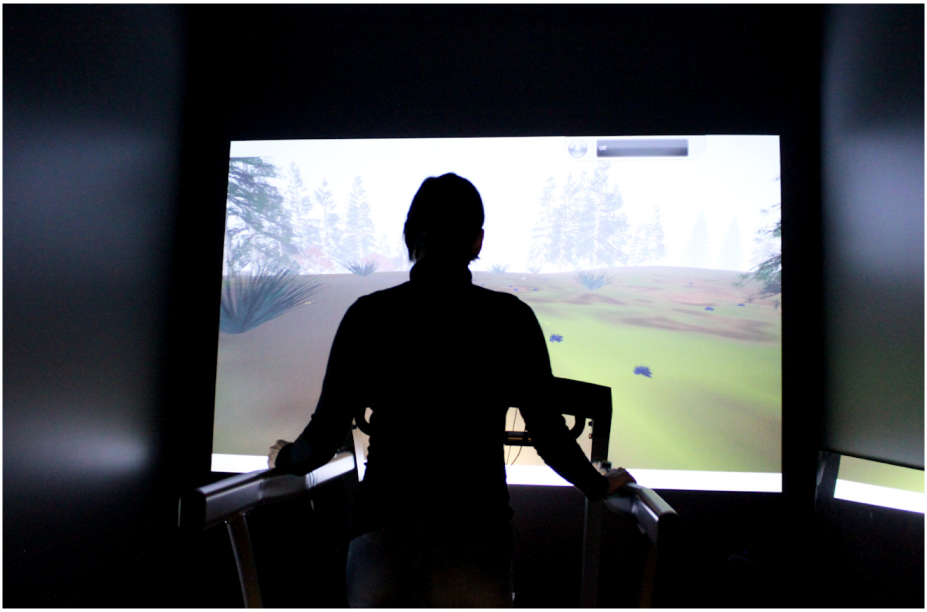


Fig. 1. A scene from Virtual Meditative Walk

For instance, our current projects such as *Virtual Meditative Walk* (Figure 1) and *Sensorium*, which corporate unique virtual environments with biofeedback and meditation, address chronic pain patients' specific embodied conditions and bodily awareness. We take into consideration their proprioceptive and interoceptive senses, which strongly shape human movement, interaction and experience, and bring embodied states -and how they are affected or transformed- into conscious awareness by mapping the changes in one's embodied states (through biofeedback mechanisms such as galvanic skin response and heart rate variability) onto changes in visual and sonic qualities of VR environment. We employ VR technologies for pain mitigation and management by controlling changes in 3D visual & sonic elements based on mindfulness-based stress reduction (MBSR) and biofeedback data in real-time to support their learning of mindfulness meditation techniques. As immersants learn how to meditate while walking, real-time biofeedback technology continuously measures breathing, skin conductance and/or heart rate. For instance, sonic VR project, *Sonic Cradle* enables users to compose a soundscape in real time using their breathing, which is one of the key elements of mindfulness meditation and self-pain-management. By revealing the changes in their bodily states in a peaceful and relaxing way, this approach encourages a calm mental clarity and loss of intention where breath-rate becomes an interface to composing a soundscape. [22]

Once brought into consciousness and learned in VR, body image and body schema can be transformed in relation to one another. For instance, the recent study conducted by William Stephoe, Anthony Steed and Mel Slater [18], which explores immersive embodiment of an 'extended' humanoid avatar featuring a tail and its impact on one's sense of ownership and agency over their body, reveals the importance of visuomotor synchrony in forming convincing perceptions of body ownership and agency, and of the plasticity of the brain's representation of the body for gestural human-computer interfaces. The study's findings show that some participants start to act as if they have a tail, by moving it through hip movements or protecting it as they move. A short-period experience of 'having a tail' in VR was enough for the participants to gain the sense of ownership over the tail, and their body schema was extended and modified through such an experience. From our perspective, such a study strongly confirm the significance of the plasticity of BIBS, and the potentials of VR environment for exploring such impacts, including therapeutic implications within diverse settings. For instance, in our most recent research, we work with teenagers, before, during and after surgery for treating scoliosis, in order to help them to transform their distorted body image and restore its balance with body schema. In this regard, the research studies on phantom limbs and neuron mirrors support the idea that watching another body performing a specific movement, even though they are not available for the person, are actually helping to fire neurons associated with those acts, and enable a psychophysical state associated with those acts. For example, seeing one's or other's body acting or moving perfectly –without any distortion – might support the physical and psychological healing through the compensation or transformation of body image and body schema (based on their plasticity and interrelatedness).

For instance, in another project, we bring soundwalking (exercise in acting listening) and biofeedback together in order to develop a unique non-invasive system for reducing pain anticipation and other effects of kinesiophobia (the fear of movement) by both encouraging and eliciting muscle activity. We examine if listening to a first-person walk through a sound environment elicits covert muscle contractions, as assessed with an electromyogram (EMG). As previous studies have shown, such covert muscle activity is triggered when participants viewed others performing certain physical activities [2] or when they visualized themselves performing a physical activity [3]. We look at whether our system, which presents the sounds of muscle movement through sounds of walking, might also elicit such covert muscle contractions and this could have therapeutic benefits in its own right [12].

Furthermore, our approach for designing for patients who have chronic pain or other chronic conditions that involve pain is not limited to VR. We also work on mobile applications and devices to extend what is learned in VR, as well as with a haptic robot and wearable sensor technologies that allow for different approaches to related problems. Based on affective computing, sensor technologies and biofeedback mechanisms, we help patients to change their affective states in order to support and encourage patients to better manage their pain experience. For instance, the projects *Analgesic Glove*, and *Haptic Creature*[23] (in collaboration with Dr. Karon MacLean and SPIN Lab at the University of British Columbia), draws upon the idea of changing the perception of self and other through the interaction with an object (a glove-like device and an animal-like creature), which have responsiveness based on sensory technologies often in combination with biofeedback. Through interaction with the objects - whether by becoming aware of proprioceptive processes that are not consciously monitored otherwise or by extending the objects or incorporating affective experience into our own body image/body schema - we can transform one's body image and body schema. As highlighted in this paper, the research on neuroplasticity underscores the significance of potentials enabled by our bodily explorations and interactions with the objects/environments to transform our embodied states, perceptions and other cognitive faculties.

The goal of this work is to offer sensorily richer and expressive communication of embodied states to the patients and health practitioners, as it affects both pre-reflexive and reflexive processes underlying those states and transformations. Furthermore, this work also aims at developing noninvasive methods to help chronic pain patients to visualize their pain, express the intensity and variability of their pain experience, and learn to use more relaxing and non-drug based treatments, such as meditation, through their interaction with immersive environments and sensory technologies. This work also addresses the issue of presentation or communication of sensory experience and data, which can go beyond the methods that are either too self-reflective (such as surveys) or too objective (such as medical examination). This confirms once again that we need to develop a holistic understanding and practice of design for bridging all these areas relevant to embodied cognition, interaction design and pain experience. The overview of our research trajectories revealed that designing for and through embodied cognition accompany one another, as we do not only design for chronic pain patients based on BIBS-related research, but those practices help us to explore the phenomena of human body and experience in unique ways that can guide further BIBS-related research.

5 Conclusion

In this paper, we outlined six fundamental ideas, derived from BIBS literature that show how the mind is embodied. We discuss our own research trajectory, within the framework of designing for chronic pain patients, which necessitates and supports a shift in our perspective of approaching interaction design. Based on our experiences, we argue that incorporating those ideas into design practices requires a shift in the perspective or understanding of the human body, perception and its experiences, all of which affect interaction design in unique ways. The dynamic, interactive and distributed understanding of cognition, where the interrelatedness and plasticity of BIBS play a crucial role, guides our approach to interaction design. The integration of BIBS literature is concerned not only with developing more expressive, inclusive or efficient designs, but about approaching design in novel ways, where we can explore the phenomena and mechanism of the human body and experience in rich multi-dimensional ways.This comes back to our emphasis on the interrelatedness of design for and through embodied cognition with a specific awareness of ongoing discussions and needs of both fields; embodied cognition and interaction design.

Acknowledgements. We would like to acknowledge the contributions of previous and current researchers, students and collaborators affiliated with the Transforming Pain Research Group (Pain Studies Lab) at the School of Interactive Arts and Technology, Simon Fraser University.

References

1. Bermúdez, J.L., Marcel, A., Eilan, N. (eds.): The Body and the Sel. MIT Press, Cambridge (1998)
2. Berger, S.M., Irwin, D.S., Frommer, G.P.: Electromyographic Activity During Observational Learning. The American Journal of Psychology 83(1), 86–94 (1970)

3. Bird, E.I.: EMG Quantification of Mental Rehearsal. Perpcetual and Motor Skills 59, 899–906 (1984)
4. Clark, A., Chalmers, D.J.: The Extended Mind. Analysis 58, 7–19 (1998), Reprinted in D. Chalmers (eds.) Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press, Oxford (2002)
5. Coakley, S., Shelemay, K.K. (eds.): Pain and Its Transformations: The Interface of Biology and Culture. Harvard University Press, Cambridge (2007)
6. Dourish, P.: Where the Action Is: The Foundations of Embodied Interaction. MIT Press, Cambridge (2001)
7. Gallagher, S.: How Body Shapes the Mind. Oxford University Press, London/New York (2006)
8. Gibson, J.J.: The Senses Considered as Perceptual Systems. Houghton Mifflin, Boston (1966)
9. Keefe, F., Huling, D., Coggins, M., Keefe, D.F., Rosenthal, M.Z., Herr, N.R., Hoffman, H.G.: Virtual Reality for persistent pain: A new direction for behavioral pain management. PAIN® 153(11), 2163–2166 (2012)
10. Kirsh, D.: Embodied Cognition and the Magical Future of Interaction Design. ACM Trans. Comput.-Hum. Interact. 1(3), 1–30 (2013)
11. Leder, D.: The Absent Body. The University of Chicago Press, London (1990)
12. Mulder, T.: Motor Imagery and Action Observation: Cognitive Tools for Rehabilitation. Journal of Neural Transmission 114, 1265–1278 (2007)
13. Noë, A.: Action in Perception. MITPress, Cambridge (2004)
14. Preester, H., Knockaert, V.: Body Image and Body Schema: Interdisciplinary Perspectives on the Body. John Benjamins Publishing, Phidelphia (2005)
15. Ramachandran, V.S., Rogers-Ramachandran, D.: Synaesthesia in Phantom Limbs Induced With Mirrors. Proc. R. Soc. Lond. B. 262(1369), 37–386 (1996)
16. Rizzolatti, G., Craighero, L.: The Mirror-neuron System. Annu. Rev. Neurosci. 27, 169–192 (2004)
17. Scarry, E.: The Body in Pain: The Making and Unmaking of the World. Oxford University Press, New York (1985)
18. Stephoe, W., Steed, A., Slater, M.: Human Tails: Ownership and Control of Extended Humanoid Avatars. IEEE Transactions on Visualization and Computer Graphics 19(4), 583–590 (2013)
19. Suchman, L.: Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press, New York (1987)
20. Tiemersma, D.G.: Body Schema & Body Image: An Interdisciplinary and Philosophical Study. Swets & Zeitlinger, Amsterdam/Lisse (1989)
21. Varela, F., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge (1991)
22. Vidyarthi, J., Riecke, B.E., Gromala, D.: Sonic Cradle: Designing for an Immersive Experience of Meditation by Connecting Respiration to Music. In: Proceedings of the Designing Interactive Systems (ACM DIS) Conference, pp. 408–417 (2012)
23. Yohanan, S., MacLean, K.E.: The Role of Affective Touch in Human-Robot Interaction: Human Intent and Expectations in Touching the Haptic Creature. Int. J. Soc. Robot 4(2), 163–180 (2012)

Exploring Initiative Interactions on a Proxemic and Ambient Public Screen

Huiliang Jin, Bertrand David, and René Chalon

Université de Lyon, CNRS
Ecole Centrale de Lyon, LIRIS, UMR5205
36, avenue Guy de Collongue
F-69134, Ecully Cedex, France
{huiliang.jin,bertrand.david,rene.chalon}@ec-lyon.fr

Abstract. Public screens are common in modern society, and provide information services to audiences. However, as more and more screens are installed, it becomes a burden for users to find information concerning themselves quickly. This is because screens cannot understand what users really need, they only display pre-designed information related to a certain location. To ensure better cohabitation between people and screens, one solution is to make screens understand users rather than make users understand screens. Given that it is difficult, even for humans, to interpret other people's intentions, it is far harder for screens to understand users. We need first to decide which kinds of information about users could be helpful for a screen to estimate to users' needs. In this paper, we study a public interactive screen, which can speculate as to users' intentions by interpreting their proxemic attributes (such as distance, movement, etc.) and context information (identity, locations, etc.). Based on proxemic interaction semantics, we built an interactive public screen, which: 1) could interpret users' needs in advance and display relevant information; 2) be available for multi-users and display distinct information to them; 3) be open for data exchanges with users' mobile devices. Through a lab study, we demonstrate that the screen presented in this paper is more attractive to users and could provide users with useful information more rapidly and precisely than traditional screens.

Keywords: Proxemic Interaction, Proxemic Screen, Public Screen, Initiative Interactions.

1 Introduction

The vision of ubiquitous computing is gradually turning into reality. A variety of screens are installed around us, providing useful information, but meanwhile people are confronted with more and more data flowing in from all sources. Public screens are typical ubiquitous media which work for public services. They always show specific information related to a particular location, for example, screens in airports display flight info, screens in shopping malls display shopping guides, etc. These screens are helpful but their functions are too unitary compared with

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 567–577, 2014.

users' diverse needs. If someone looks at a screen that does not provide him/her with useful information, they will ignore this screen. For example, few people will stop in front of an advertisement screen installed in a railway station, because they want to find out information about trains and not shopping. Screens should offer more diversity, but also more precise information to specific users. We conclude that current screens have three disadvantages; First, public screens are static. At present, if a tourist wants to go to the airport of a city to take a flight, he/she cannot obtain an answer from a screen installed in the bus station concerning the best public means of transport to the airport, but needs to check and analyze the route by him/herself. This is a lengthy process and does not rule out errors. Modern screens in bus stations should be able to detect users' actual intentions, for example to detect a user who wants to go to the airport, and to display instantly the best route from the current location to the airport. Secondly, public screens nowadays can only be used by individual users, which means "first come first served": a user needs to wait for the current users to leave for he/she to use the screen. Even if the screen is large enough to display plenty of information, it is a waste of display capabilities. Thirdly, current public screens are blind to ambient devices: it is impossible for users to download any interesting information from a screen directly. The only ways for users to get information from public screens is to memorize it, or take a snapshot by smartphone: neither method is very efficient.

We live in a ubiquitous society where data is exploding. These old-fashioned screens are not intelligent enough to cope with the development of society because they only show information exhaustively without knowing if someone is interested by it. By contrast, an intelligent screen should understand users' needs, and display dynamic information to specific users in specific contexts, thus ensuring the current user can rapidly obtain the exact information he/she wants. Furthermore, an intelligent screen should be open to other devices, especially personal devices (e.g. smartphone, tablet, etc.). Mobile devices are already universal, and thus typical ubiquitous devices. A connection between public screens and personal mobile devices could create a real ubiquitous device network.

Challenges still need to be faced to achieve these prospects. How can we make a screen understand users' needs? What kinds of contexts should be taken into consideration? How can we design interfaces for multiple users? How can we connect personal mobile devices with public screens? Researchers have studied some parts of these issues from different aspects. Most of them focused on natural interactions with a screen by technical methods (by touching, smartphone or mid-air gestures), others studied design principles of a public screen, while others studied issues such as evaluation, photo sharing among personal devices and public screens, etc.

While individual research has already been conducted on these issues, proxemic interaction theories discussed them more systematically, examining human-screen interaction based on proxemic theories, which study nonverbal communication between people. At the beginning, distance was taken into consideration as references of communication between users and a large screen. For example, D.Vogel et al. [1] explored multi-level interactions from implicit to explicit in front of a screen according to users' distances from the screen. They divided the space in front of a vertical screen into four discrete zones which correspond to four interaction phases: ambient screen, implicit interaction, subtle interaction, and personal interaction. They designed sharable interfaces for users in different phases. Their prototypes were mainly distance-based but not completely proxemic interactions.

S.Greenberg et al. [2] extended the proxemic theory of inter-human nonverbal communication [3] to human computer interaction. They referred to the theory of Vogel and Ravin, before coining the term "proxemic interaction" as a novel kind of spatial related interaction. The advantage of proxemic interaction is that it makes a screen interact with users initiatively in different proxemic areas of the screen. It takes not only distinct distance as references of interaction, but also successive orientation and movement, as well as user's identity. Marquardt et al.[4]developed a proximity toolkit, which could easily integrate proxemic data in real time applications. Furthermore, they studied the location attributes of an intelligent room, including the user's spatial relationship with the fixed (doors, walls) and semi-fixed (sofas, chairs) features. Although they built a proxemic interaction theory systematically, they did not sufficiently examine how these proxemic attributes could improve interactions between users and a public screen. Furthermore, regarding the multi-user scenario, they only studied simple collaboration of two users based on several demo applications. With regard to practicability, users have to wear additional markers to be recognized by their system, thus limiting the practicability of their prototypes.

In this paper, we mainly study the initiative interactions of a screen based on the proxemic interaction semantics coined by S.Greenberg et al. Initiative interaction means that the screen described in this paper can attract users by some active responses rather than wait to be discovered, and provide more personalized information to users by interpreting their behaviors rather than making users find this information themselves. This paper is divided into three parts:

1. we build an intelligent screen, which understands the meanings of users' proxemic attributes, and displays dynamic interfaces based on users' spatial relationship with the screen and contexts information;
2. with regard to multi-user conditions, we design individual interfaces for users in different zones, from public to personal;
3. we develop a tool to connect seamlessly users' mobile devices with the screen, and exchange data between these two media;

To study these issues, we have constructed a large vertical screen with a projection surface, installed in a semi-public area of the laboratory. We have studied initiative interactions with the screen based on interpretation of users' proxemic attributes, as well as studying communication between the screen and users' personal mobile devices.

2 Related Work

2.1 Proxemic Interaction

Distance-based interaction was the early form of proxemic interaction. N. Roussel et al. [5] studied distance-based interaction applied with a video communication system. Ju et al. [6] introduced a distance-related interactive whiteboard deployed in a lab for collaborative work. Both these prototypes are based on the user's distance from a screen, as well as on a lean and zoom interface [7] and the work of Vogel, D et al.[1]. S. Greenberg further studied proxemics as references of interaction with a vertical screen[2]. Marquardt et al.[4]developed a proximity toolkit depending on the marker-based VICON motion tracking system and Kinect. This toolkit supports rapid prototyping of proxemic interaction, providing fine-grained proxemic data between people, portable devices, large interactive surfaces and other non-digital objects in a special test room. The proxemic interaction studied by S.Greenberg et al. recognized users' natural behaviors and analyzed them as implicit inputs for interaction: for example, if a user took out a phone then the movie playing paused to wait for him/her to make the phone call. This is interesting, however, as in a real situation the meaning of users' behaviors might be different: the same behavior might have a different meaning depending on the context. Therefore, it is better to let users make choices by explicit commands as well. Proxemic interaction of mobile devices has also been studied, either for controlling the screen (point a mobile phone to a screen), or for transferring files (ProxemicCanvas in[8]).

2.2 Communication between Devices

Alt. F et al.[9]compared methods for posting contents from a user's smartphone to a public notice screen (e.g. directly touch input, phone/screen bump etc), as well as retrieving methods (QR code, email, print etc.). Their results revealed that screen interaction was favored if users could post contents ad-hoc on the screen. Data communication between devices could use: Bluetooth, matrix or bar codes, NFC/RFID, Wi-Fi, Cloud sync (e.g. Dropbox) and other methods (USB, ZigBee). Cheverst et al. [10] explored pictures exchanging between a mobile phone and a screen over Bluetooth. As the author indicated, the reliability of the Bluetooth discovery process was an obdurate problem, and the pairing process was also time-consuming. Matrix codes (e.g. QR code) and barcodes are widely used as practical means of transferring data from a public screen to a mobile phone, but scanning a small matrix code is not always convenient for users (e.g. in a dim light or in a crowded place). By contrast, Wi-Fi is a more popular way: connecting mobile devices with public screens via Wi-Fi is more attractive to users than mere local connections. Although there are many off-the-shelf applications which support data exchange between devices via Wi-Fi, e.g. [11], they are designed for home use, and their operation is somewhat redundant for public use. A specific data exchanging tool needs to be designed for public use.

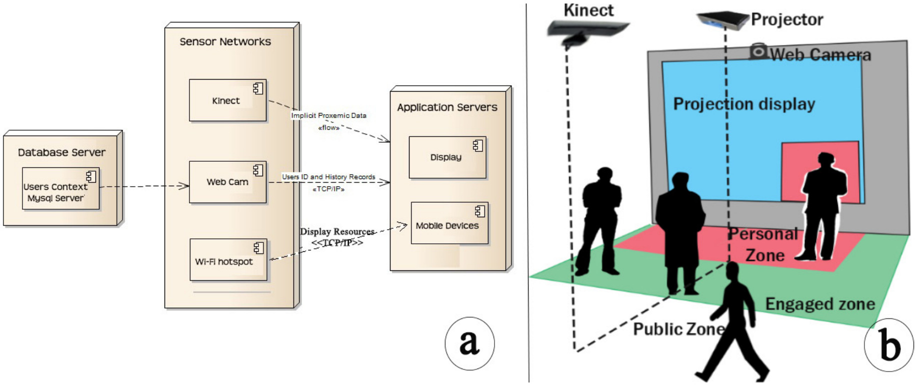


Fig. 1. UML deployment diagram and installation of system

3 System Architecture and Interaction Design

We constructed a large screen with a projection surface and equipped it with Kinect to recognize users' proxemic attributes (e.g. distance, orientation, movement, etc.) and implement mid-air gestures; a web camera is installed above the screen for identity recognition. The sensor data are sent to the screen server, and analyzed. The server then interprets users' potential intentions and renders the relevant contents to audiences. The system UML deployment diagram is shown in Figure 1a, while system installation is shown in Figure 1b.

In like manner to D.Vogel [1], we divided up three discrete zones in front of the screen from far to close (Figure 1b): public zone (PZ), engaged zone (EZ), and personal zone (PeZ). According to the zone, users have different possibilities of interaction, and can read different levels of information (from public to personal). For example, the screen only displays general information to a user passing by PZ quickly, but if he/ she enters EZ, the screen server judges that he/ she wants further info and displays information which could be interesting to him/ her. We consider users' personal mobile devices, mainly smartphones, as another zone: Privacy zone (PrZ), where users can download relevant private information from the public screen to read in the screens of their personal devices. Although we divided up physical zones discretely, contents on screen presenting to users are transiting progressively for better experience.

3.1 Public Zone

Users in a public zone can read general messages, such as advertisements, notices, etc. The screen does not try to work out the intention of users in this zone, because there might be many users passing by. However, it tries to attract the intention of passersby by making them aware that they are detected by the screen. We create a colored circle with no contents for a user in the public zone. This circle progresses according to the user's movements: if he/she moves close

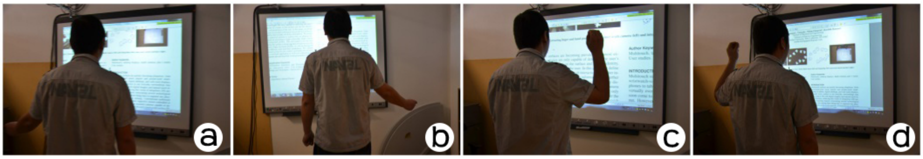


Fig. 2. Gestures Available in EZ (a: scroll up, b: scroll down, c: zoom in, d: zoom out)

to the screen, the circle is enlarged gradually, while if he/she moves away from the screen, the circle shrinks until it disappears.

3.2 Engaged Zone

If a user in a public zone is attracted by the circle's animation and enters the engaged zone, the circle will turn instantly into a small window to show that he/ she has been recognized by the system: this window does not contain contents if there are already users in the personal zone of the screen, but it could remind users that they could explore more interactions. Users in this zone can manipulate the public contents by natural mid-air gestures regardless of users in the personal zone. We support four kinds of gestures in the engaged zone: user could stretch hand to scroll up or down the interface for browsing current contents (Figure 2a, b), or raise hand to zoom the interface to inspect details (Figure 2c, d). The small window belongs to a user pans along with the user's movement in this zone to keep his/ her attention. If there are no users in the personal zone, some interesting information might be displayed in the window, for example a thumbnail of his/ her calendars, or social network notices, etc. To find out more details, the user could step further into the personal zone.

3.3 Personal Zone

If a user enters this personal zone, the window allocated to the user will be enlarged and anchored in front of his/her eyes: this window then becomes a temporarily private display area in the public screen, and more personal details information is displayed to the user in the private display area. More fine- grained gesture interactions are supported for users in this zone: for example, the user can zoom in and out of his/her display area by pinch gestures (in or out) to adjust the window to his/ her favorite size. The user can wave his/her hand to flip over current contents to the last or next page, or glide the current page by swiping his/her hand upwards or downwards. However, as the screen described in this paper is not a touch screen, interactions on the screen are not as precise as with a tactile one. Although tactile screens are more effective, implementation of touch sensitive interactions on a large screen is difficult and expensive, interactions on screen are not the key issue in this paper, we will not compare interaction here with tactile screens. When one user in the personal zone is operating, the other parts of the screen continue to display general public information, as shown in

Figure 3a. In this way, we divide a public screen into a public display area for general users and into several private display areas for particular users. We thus break the rule of "first come first served" by taking full advantage of the display capabilities of a large screen.

Although we try to protect users' privacy by displaying some personal infor mation in a small private window, there is still some private information that users are not willing to display in plain texts. As a result, before displaying infor mation in a user's private window, we let the user decide whether the information should be displayed in texts or in the format of a document. If the user chooses the information to be displayed as a document, then he/she needs to download the document to his/her mobile devices for reading.

3.4 Privacy Zone

As discussed in the above section, users' private windows make sure they can read personal information while still ensuring the security of this information. However, since private windows are part of a large screen, there is still a risk of exposing privacy. Compared with the large screen, personal devices (e.g. smartphone, tablet, etc.) have small screens, which are ideal media for displaying personal information. It is more secure and convenient to migrate personal information on users' private windows to their smartphones. We develop a toolkit known as a direct migrator, able to connect the two types of media seamlessly, thus allowing users to exchange information freely with public display via their mobile devices [12] . This tool is implemented in Java, and is based on Client/Server Wi-Fi Socket protocols. It has two advantages. First, with the toolkit, downloading files from a public screen to mobile devices is fairly simple (Figure 4a, b). We built a Wi-Fi hotspot along with the screen: the user connects his/her smartphone to the hotspot, he/she can then select any item displayed on the screen, and download that item to the smartphone by clicking a download button on the mobile interface. By contrast, if a user wants to post some information on the screen, he/she has only to select the file from the smartphone and click on the post button: the selected file is sent and posted on the screen (Figure 4c). Second, interactions with the toolkit are natural and intuitive. The File Migrator supports not only button-based interactions, but also gesture interactions. For example, swiping your finger from the bottom of the smartphone screen upwards will send a selected file out to the target device (Figure 5a), or swiping your finger from the top of the mobile screen downwards will retrieve a selected file from the screen (Figure 5b).

3.5 Multi Users Situation

A large screen has sufficient display capabilities. However, this is not always taken full advantage of because only one user can interact with a screen at any one time. In this system, several users in the personal zone can read information and interact with the screen simultaneously because each user in the personal zone has a personal display window, as shown in Figure 3a. Moreover, the private

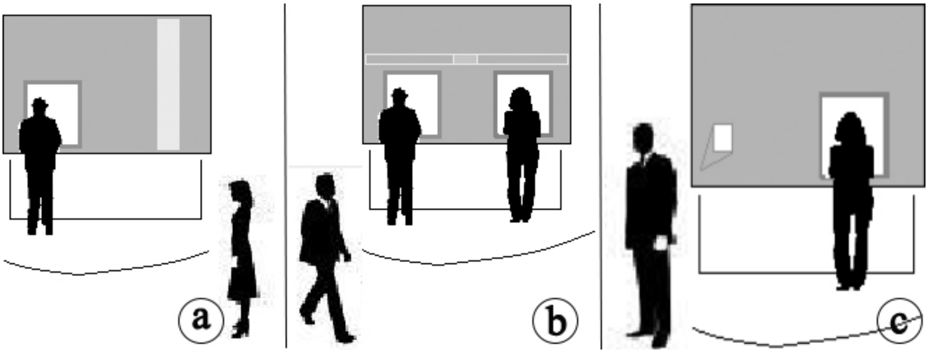


Fig. 3. private display window on screen, a: one user in PeZ, the other in EZ; b: two users in PeZ, one user passes by in EZ; c: one of users walk out and his personal window fade out

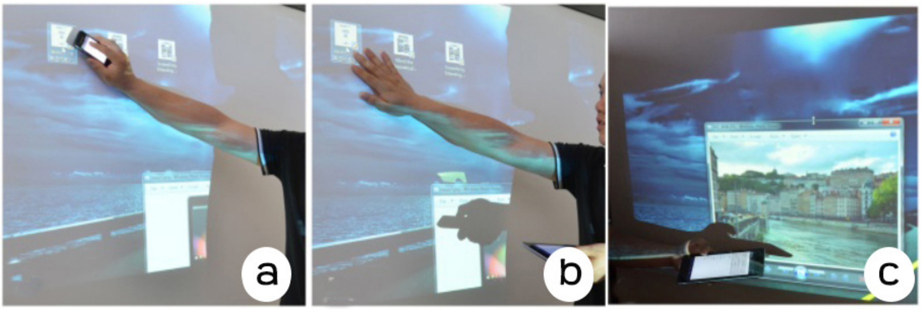


Fig. 4. Interactions in PrZ (a: user selects a file with smartphone and downloads the file, b: user selects a file by hand, and clicks on the tablet to download the file, c: upload an image from the tablet to the screen)

windows only occupy the lower spaces of the screen. If another user approaches the screen at this time, an additional small window is created in the upper space, and the window as well translates along with users' movement to catch his/her attention (Figure 3b), he/she could decide to step further into the personal zone, or just leave away. Private windows make sure users can interact with the screen individually without interference. The private window will be removed if one user walks away. In Figure 3c, one user downloads contents to his smartphone and moves away: his personal window thus fades out.

4 User Study and Discussion

We organized a lab study to evaluate the prototype. We invited 10 volunteers (3 females, 7 males, average age 26.5 years old) to participate in the study.

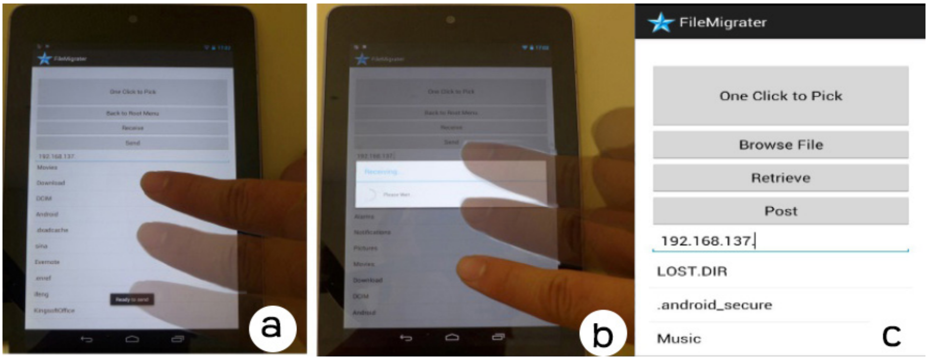


Fig. 5. File Migrator (a:swipe upwards to send file, b:swipe downwards to download file, c: the mobile UI)

4.1 Task and Procedure

The test is divided into two parts: interaction with the proxemic screen, and information exchange between the screen and the user's smart phone. Before the test, we played a short tutorial video for testers, allowing them to quickly understand the functions of the prototype.

The screen displays general messages if there is nobody in front of it. A tester enters the zones in front of the screen from far to close. First he/she passes by the public zone: a colored circle instantly appears and evolves along with the user. The tester is thus attracted to the Engaged zone. At the same time, the circle turns gradually into a window, also evolving along with the user. The window displays some personal messages for the user: as a demo, we display a greeting sentence (e.g. Hello, Mr/Mrs ROBERT). The tester zooms in and browses the current interface by mid-air gestures. Then he/she decides to find more information and steps further into the personal zone. Meanwhile the window has enlarged, and is placed just in the screen in front of him/her, and more personal information is displayed. The tester zooms in and out of his/her private window by pinch gestures, and browses the contents by glide gestures. An icon showing a smartphone is displayed in the corner of the personal window, to remind the tester that he/she can download contents by smartphone. The tester then takes out the smartphone, which we prepared for the test and connected to the local Wi-Fi hotspot. He/she launches the file migrator, selects a document in the personal window and clicks on the download button: the contents in the window are shifted to the smartphone. Finally, the tester moves away from the screen and the personal window is removed. The screen returns to its default status.

4.2 Test Results and Discussion

After the test, all testers are required to fill in a SUS (System Usability Scale) questionnaire to evaluate the usability and learnability of the prototype [13].

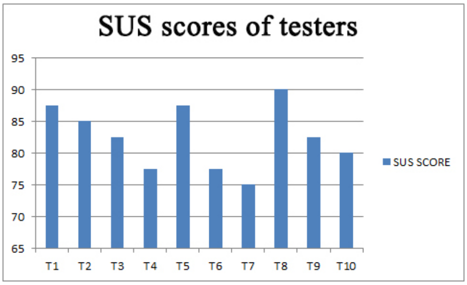


Fig. 6. SUS Scores of the testers

The SUS includes 10 statements (5 are positive and 5 are negative). For example, I think that I would like to use this system frequently, I found the system unnecessarily complex. Each item has 5 response ranges from "strongly agree" to "strongly disagree". The result of the SUS is a score from 0 to 100, able to reveal the objective usability of a product. The average SUS score for the proxemic screen prototype is 82.5, Grade B (individual scores of testers are shown in Figure 6). This score implies that users are willing to recommend this product to friends. All 10 users agreed that the proxemic screen is more attractive than normal screens, and that it is easy to learn even for first-time users. 8 testers out of 10 said they would like to use this system if it is available in real life. However, they also doubted that using a camera to recognize users' identities for public installation might not be practical because it is difficult to collect all information for passersby and to determine rules for deciding what kind of contents could be displayed on a public screen. Testers particularly appreciated the possibility of transferring data from a public screen to personal mobile devices, and thought that the toolkit greatly improves the practicability of a public screen.

5 Conclusion and Future Work

In this paper, we built a public screen which tries to understand users' intentions by their proxemic attributes and context information, thus providing users with personal-related information, instead of making them search for information from mass data. We took full advantage of the display capabilities of a large screen, and designed interfaces that could display distinct information to different users at the same time while still ensuring the security of privacy. Furthermore, we developed a tool for connecting and exchanging data between a public screen and personal mobile devices. Compared with normal public screens, the screen constructed in this paper could change the convention governing people's interaction with public screens. It is more efficient for users to get information they need from public screens because screens understand users, and there is no barrier between personal devices and public media, or between small screens and large screens. The screen is a real ubiquitous media compared with traditional screens. In the future, we will continue to develop and test the prototype, and, during the process, try to discuss more interesting application scenarios of the screen in real life contexts.

References

1. Vogel, D., Balakrishnan, R.: Interactive public ambient displays: Transitioning from implicit to explicit, public to personal, interaction with multiple users. In: Proc.UIST, pp. 137–146. ACM Press (2004)
2. Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R., Wang, M.: Proxemic interactions: The new ubicomp? interactions 18(1), 42–50 (2011)
3. Hall, E.: The Hidden Dimension. Anchor Books (1966)
4. Marquardt, N., Diaz-Marino, R., Boring, S., Greenberg, S.: The proximity toolkit: Prototyping proxemic interactions in ubiquitous computing ecologies. In: Proc. UIST, pp. 315–326. ACM Press (2011)
5. Roussel, N., Evans, H., Hansen, H.: Using distance as an interface in a video communication system. In: Proc. IHM, pp. 268–271. ACM Press (2003)
6. Ju, W., Lee, B.A., Klemmer, S.R.: Range: exploring implicit interaction through electronic whiteboard design. In: Proc. CSCW, pp. 17–26. ACM Press (2008)
7. Harrison, C., Schwarz, J., Hudson, S.E.: Tapsense: enhancing finger interaction on touch surfaces. In: Proc. UIST 2011, pp. 627–636. ACM Press (2011)
8. Marquardt, N., Greenberg, S.: Informing the design of proxemic interactions. IEEE Pervasive Computing 11(2), 14–23 (2012)
9. Alt, F., Shirazi, A.S., Kubitza, T., Schmidt, A.: Interaction techniques for creating and exchanging content with public displays. In: Proc. CHI 2013 (2013)
10. Cheverst, K., Dix, A., Fitton, D., Kray, C., Rouncefield, M., Sas, C., Saslis-Lagoudakis, G., Sheridan, J.G.: Exploring bluetooth based mobile phone interaction with the hermes photo display. In: Proc. Mobile HCI 2005, pp. 47–54. ACM Press (2005)
11. Wi-files, https://itunes.apple.com/us/app/wifi-files/id416409502?mt=8
12. Jin, H., Xu, T., David, B., Chalon, R.: Direct migrator: eliminating borders between personal mobile devices and pervasive displays. In: The 5th IEEE Workshop on Pervasive Collaboration and Social Networking 2014 (PerCol 2014), Budapest, Hungary (March 2014)
13. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. 24(6), 574–594 (2008)

Evaluation of Tactile Drift Displays in Helicopter

Patrik Lif[1], Per-Anders Oskarsson[1], Johan Hedström[1], Peter Andersson[1],
Björn. Lindahl[1], and Christopher Palm[2]

[1] Swedish Defence research Agency, Linkoping Sweden
{patrik.lif,p-a.oskarsson,johan.hedstrom,
peter.andersson,bjorn.lindahl}@foi.se
[2] Linkoping University, Sweden
chrpa087@student.liu.se

Abstract. Brownout during helicopter landing and takeoff is a serious problem and has caused numerous accidents. Development of displays indicating drift is one part of the solution, and since the visual modality is already saturated one possibility is to use a tactile display. The main purpose in this study was to investigate how tactile displays should be coded to maintain or increase the ability to control lateral drift. Two different tactile drift display configurations were compared, each with three different onset rates to indicate the speed of lateral drift. A visual drift display was used as control condition. The results show that best performance is obtained with the basic display with slow onset, and with complex display with constant onset rate. The results also showed that performance with the best tactile drift display configurations was equal to the already validated visual display.

Keywords: Tactile display, helicopter, brownout.

1 Background

Brownout caused by blowing sand is a serious problem for helicopter pilots. The main problem is limitations of visual references during takeoff and landing, which has led to numerous incidents [1]. The three main problems are lateral drift, difficulties of keeping heading to reference cue, and rate of closure. Lateral drift means that the helicopter drifts sideways without the pilot is noticing it. The consequences may be that the helicopter crashes against wires, a mountain side, or flips to the side during landing. Also, blowing sand may lead to the false perception that the helicopter moves in opposite direction. To remedy for this, at least three areas need attention: displays, sensors, and crew-cooperation. However, our focus is on tactile display solutions to compensate for lack of visual references. The basic idea with a tactile display is that the users (i.e. helicopter pilots) can receive tactile information and simultaneously look away from the instruments and instead look out the cockpit window to maintain orientation, look for obstacles on the ground in a landing situation, or keep track of the surrounding. Furthermore, pilots are in many situations already visually saturated.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 578–588, 2014.

Therefore, instead of adding further visual information, a tactile display may improve performance.

Alternative displays, such as tactile displays, can be useful both in military [9], [6] and civilian [2] situations. Furthermore, performance often improves when multimodal displays are used to simultaneously present redundant visual, tactile and auditory information. The idea of bimodal systems for simultaneous presentation of visual and auditory stimuli was proposed by [4]. Later Wickens [13-14] developed a theory called multiple resource theory (MRT). MRT [7], [15] describes our perceptual resources in four dimensions: modalities, stages, access, and responses that are represented in a cube. Risk for mental overload can be predicted by the amount of interference between information processing in the dimensions. A well designed multimodal display, i.e. distribution of the information between visual, tactile, and auditory channel, can improve performance and perhaps reduce mental workload.

In a helicopter study pilots were able to stay closer to a moving target during hover when using a tactile display (than without) and they also rated their situation awareness as higher when they used the tactile display [8]. Curry and Estrada [5] showed that a tactile belt significantly improved drift control during takeoff and hover in helicopters. For fatigued pilots (awake for 31 hours) drift control was significant better with than without a tactile belt. Even though tactile, bi- and multimodal displays are proven to improve performance compared to only visual displays, one question that need further attention is how to optimize the tactile pattern to improve performance further.

In both visual and tactile displays there is a need for good design, and to make sure that the user intuitively understands the information. Design of visual displays has a long tradition, but for tactile displays the design and evaluation of different tactile pattern or tactile messages are rather new, with exception of sensory impaired individuals. In military settings, efforts have been done, in i.e. soldier communication where design of tactile language and evaluation of tactile displays for dismounted soldiers [3],[12] proven to be useful.

Here we focus on tactile displays, and compare different tactile displays with each other and with a visual drift display for helicopter pilots. The main purpose was to investigate how tactile displays should be coded to maintain or increase the ability to control lateral drift. Two different tactile drift display configurations were compared, each with three different onset rates (frequencies) for indication of the speed of lateral drift. A previously validated visual drift display was used as control condition. To make sure that the visual and tactile display could be compared, they were coded in corresponding ways, i.e. make sure that perceptual momentum was achieved [10].

2 Method

2.1 Participants

14 participants without prior experience of flying a real or simulated helicopter participated in the experiment, five females and nine males. Their mean age was 24.1 years, with a standard deviation of 2.7 years.

2.2 Equipment

The simulation was performed with a PC-based simulation of a Bell 206B JetRanger helicopter, with Prepare 3D, an extension of Microsoft® Flight Simulator X. The surrounding graphics were presented on three 46-inch LED screens, with resolution 1920 × 1080 pixels. The three displays were placed in upright position with the left and right display slanted 28 degrees inward, and distance from the participants' eyes to the screen was approximately 110 centimeters. The primary head-down displays was the same as used in the instrumentation in a standard Bell 206B JetRanger and were presented on a 23-inch LCD touch display with resolution 1920 × 1080 pixels (Figure 1).

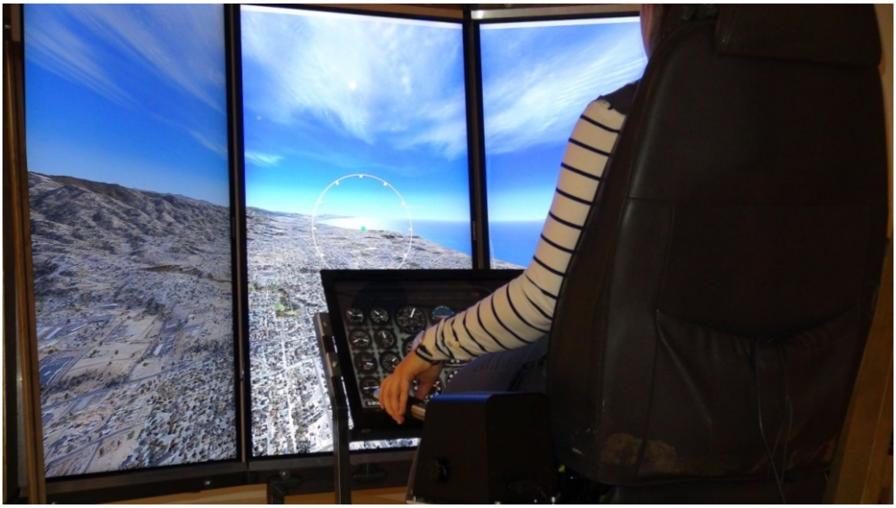


Fig. 1. Experimental setting with visual primary head-down display, and visual head-up drift display on the central screen

The simulator was controlled by joysticks and pedals, steering with a Flight Link G-stick III Plus Cyclic and a Flight Link Anti-Torque pedals, and throttle level by a Flight Link Collective throttle C1. The simulator was motion based with a MOOG Electric Motion Base MB-E-6DOF/12/1000 kg.

Three drift display configurations were used: visual drift display, tactile basic drift display, and tactile complex drift display. Also, for each tactile display configuration three different tactile patterns were used. The visual drift display was based on the hover display in a Seahawk helicopter, but was adapted at the Swedish Defence Research Agency (FOI), i.e. it only presented information about lateral drift, left or right, whereas the original hover display presents drift in all directions. In this experiment the visual drift display was presented head-up on the middle of the three 46-inch displays (Figure 2).

Speed of lateral drift, to the left or right, was indicated by a green vector. The length of the vector increased with the speed of lateral drift. When the drift was zero (or below the threshold of 0.4 m/s) no vector was visible and only the green square in the center was visible. When lateral drift was 5 m/s the vector reached halfway to the white outline of the circle. When lateral drift was 10 m/s, or more, the vector reached the outline of the circle. In Figure 2 the vector indicates lateral drift to the left with 10 m/s or more.

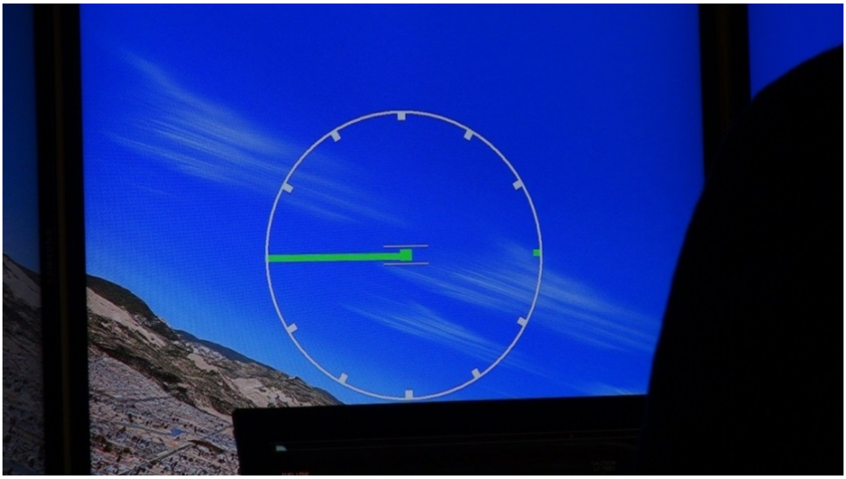


Fig. 2. Visual drift display indicating lateral drift to the left with 10 m/s or more

The tactile drift displays (basic and complex) consisted of a tactile vest developed at the Swedish Defence Research Agency (FOI) with three rings of motor tactors (120 Hz) sewn-in the fabrics. Each ring has 12 motor tactors evenly positioned over the torso and one tactor is always positioned straight ahead. In this experiment only the middle ring was used. (Figure 3). For both the basic and complex displays the length of the pulses was 100 ms.

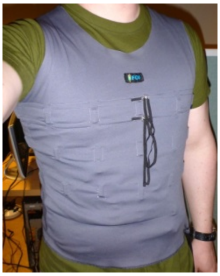


Fig. 3. Tactile vest

The tactile basic drift display only used two tactors, one tactor under each armpit. Drift was indicated by vibrations under the armpit, at the same side as the direction of the lateral drift, left (9 o'clock) or right (3 o'clock) with vibrating pulses. Analogously to the visual drift display, when the drift was zero (or below the threshold of 0.4 m/s) the belt was silent, i.e. no vibrations were given.

The tactile complex lateral display used a more complex motion pattern and used all 12 tactors. Seven tactors were used for indication of drift to left and right respectively. The middle tactors (12 and 6 o'clock) where used both for the left and right drift indication pattern. The pattern indicated drift with pulses that alternated between the tactor under the armpit in the direction of the drift and to two other tactors, with a paus of 200 ms when shifting between tactors. When drift was zero (or below the threshold of 0.4 m/s the belt was silent, i.e. no vibrations were given. When lateral drift to the right exceeded 0.4 m/s the pulses alternated between the tactor under right armpit (3 o'clock) and the central tactors (12 and 6 o'clock simultaneously). When lateral drift to the right reached 1/3 of maximum indicated drift speed (1.67 m/s with fast onset and 3.33 m/s with constant and slow onset rates) the pulses changed to alternate between the tactor under right armpit (3 o'clock) and tactors 1 and 5 o'clock simultaneously. When lateral drift to the right reached 2/3 of maximum indicated drift (3.33 m/s with fast onset and 6.67 m/s with constant and slow onsets) the pulses alternated between the tactor under the right armpit (3 o'clock) and tactors 2 and 4 o'clock simultaneously. Drift to the left was presented analogously.

For both the tactile displays three different onset rates were used. At constant onset rate the interval between the pulses were always 2000 ms. At slow and fast onset rate the interval between the pulses decreases linearly from 2000 – 200 ms when the speed of the drift increased. When lateral drift started the interval between the pulses was 2000 ms, and when lateral drift was 10 m/s or more the interval between the pulses were 200 ms. At fast onset rate the minimal interval of the pulses (200 ms) was reached when lateral drift was 5 m/s or more. In other respect everything was identical to slow onset rate. This means that when the constant onset rate was used the basic display only gave information of direction of the drift, whereas the complex display gave information of both direction and speed of the drift. However, when the slow and fast onset rates were used the basic display gave information about both direction and speed of drift, and the complex display gave information of drift in two ways, both by the pattern and by the interval between the pulses (see Table 1).

Table 1. Types of information of lateral drift given by each tactile display configuration

	Direction of drift	Speed of drift by onset rate	Speed by tactile pattern
Basic constant onset	x		
Basic slow onset	x	x	
Basic fast onset	x	x	
Complex constant onset	x	x	
Complex slow onset	x	x	x
Complex fast onset	x	x	x

2.3 Design and Procedure

The experiment was performed with a within-subjects design with the two tactile lateral drift display conditions by the three onset rates. The visual drift display was used as a control condition. The order of the display configurations was balanced over participants. The participant's main task was to fly straight ahead and avoid lateral drift, and secondary to fly in a specific heading (north, south, west or east), at the altitude of 3000 feet, and with the speed of 30 knots.

Before the experiment started the participant received training, starting with a walkthrough about the functionality of the helicopter and the primary instruments (speed, altitude, compass, and gyro). The next phase was first free flight and then to fly at 3000 feet in 30 knots in a specified direction, analogously to how the task should be performed in the experiment. Then the participant learned the different drift displays, and made a test flight with each drift display configuration. The training continued until the experiment leader assessed that the participant had adequate control over the helicopter and fully understood the seven display configurations. The average training time was approximately one hour.

After the training was completed the experiment started. The participants tested each display configuration in turn (order depending on balancing). The duration of each display configuration was two minutes and was equally performed with all display configurations. After each display configuration the participant answered a questionnaire and after the last display configuration was completed the participant also answered a final questionnaire. Total time for the experiment, including training and instructions, was approximately 2 hours.

During each of the experimental conditions the experimental leader asked the participant to identify speed, altitude, and direction four times. The purpose was to make it necessary for the participant to avert the eyes from the displays and thereby increase the realism. The simulated weather conditions were good. The following performance measures were collected:

- Lateral drift – mean absolute lateral drift (m/s)
- Deviation of speed – mean absolute deviation from intended air speed of 30 knots (m/s).
- Deviation of altitude – mean deviation from intended altitude of 3000 feet (m).
- Deviation of heading – mean absolute deviation from the intended direction (degrees).

The reason for using the mean of the absolute values was to calculate the size of the deviation, i.e. the direction of the deviation was not important. All measures were logged with a sampling frequency of approximately 9 Hz. For each participant, in each display configurations, the mean was calculated from the absolute values of the sampling points.

The questionnaires mainly contained 7-point rating scales. Descriptions of analyzed questions are given in the Results.

3 Results

Data were analyzed with analysis of variance (ANOVA). In the case of violation of the sphericity assumption in the ANOVAs the Greenhouse-Geisser-corrected *p*-value is given. Post hoc testing was performed with Tukey's honestly significant difference (HSD) test.

3.1 Performance Measures

The performance measures of lateral drift, deviation of speed, deviation of altitude, and deviation of heading were each analyzed in two ways. First, a two-way factorial repeated measures ANOVA, 2 types of tactile displays (basic, complex) × 3 onset rates (constant, slow, fast) was used to investigate main and interaction effects of the tactile displays and the onset rates. Secondly, a one way repeated measures ANOVA was used to compare the 6 tactile display configurations with the visual drift display that was used as control condition. The reason for using separate ANOVAs is that there was only one type of coding of the visual drift display, thus main and interactions effects of tactile display configuration by onset rate could not be analyzed with the visual display included.

Lateral Drift. The two-way ANOVA showed a significant interaction effect of display by onset rate (Figure 4), $F(2, 26) = 8.13$, $p = .002$; but no main effects of display

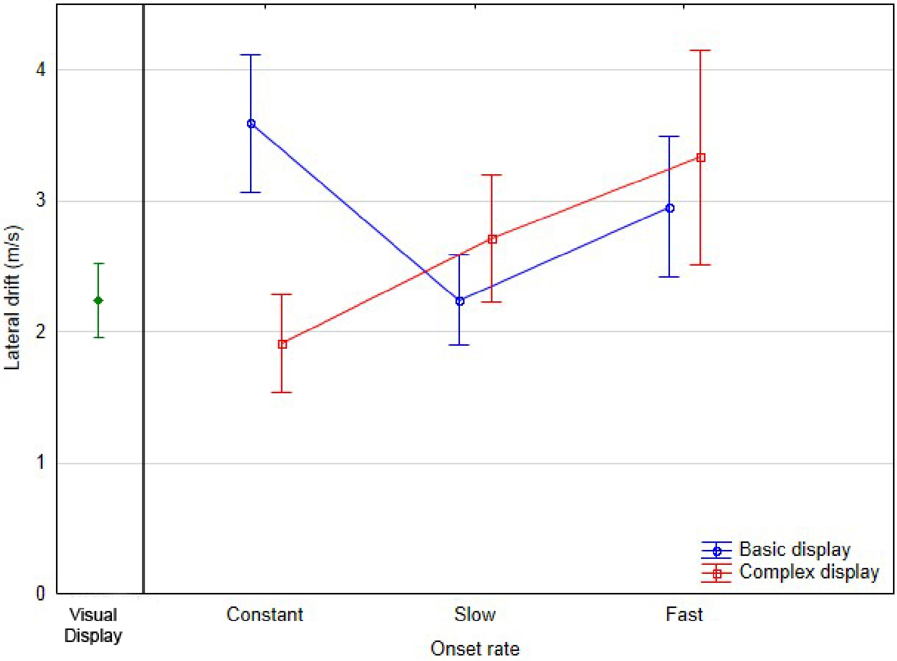


Fig. 4. Mean lateral drift with the visual drift display and the two tactile drift displays with three onset rates

or onset rate. Post hoc testing showed significantly larger mean lateral drift for the basic display with constant onset compared to the complex display with constant onset rate (*p* = .007). Also, for the basic display lateral drift was significantly higher with constant compared to slow onset rate (*p* = .041), whereas for the complex display lateral drift was significantly lower with constant compared to fast onset rate (*p* = .028).

The one-way ANOVA only showed a strong tendency, *F*(6, 78) = 3.00, *p* = .050 of difference between the seven display configurations. Post hoc testing showed no significant differences between the visual display and the tactile display configurations.

Deviation in Speed. The ANOVAs showed no significant differences. The mean deviation from correct speed was 7.1 m/s.

Deviation in Altitude. The ANOVAs showed no significant differences. The mean deviation from correct altitude was 71.2 m.

Deviation in Heading. The ANOVAs showed no significant differences. The mean deviation from correct heading was 15.6 degrees.

3.2 Subjective Measures

The subjective ratings were analyzed in the same way as the performance data, one ANOVA was used to analyze main and interaction effects of two tactile display configuration and the three onset rates, and one ANOVA was used to analyze if there was any differences compared to the visual display configuration. However, since there were three questions about perception of lateral drift and two questions about disturbance of lateral drift, perception and disturbance of lateral drift were analyzed with a three-way repeated measures ANOVA, 2 types of tactile displays (basic, dynamic) × 3 onset rates (constant, slow, fast) × number of questions, to investigate main and interaction effects of the tactile displays and the onset rates, and a two-way ANOVA, 7 display configurations × number of questions, to compare the 6 tactile display configurations with the visual drift display.

Since there was only one question about mental workload, the analysis was equal to that of performance data, one two-way ANOVA to investigate main and interactions effects between the tactile displays and the onset rates, and one one-way ANOVA to compare the 6 tactile display configurations with the visual drift display.

Perception of Lateral Drift Presentation. The participants answered three rating questions about how they experienced the display information of lateral drift: perception, accuracy, and understanding of the drift information (1 = Very difficult – 7 = Very easy).

The three-way ANOVA only showed a significant main effect of onset rate, *F*(2, 26) = 4.55, *p* = .020. Post hoc testing showed that this was due to significantly higher ratings of how lateral drift was experienced at high onset rate (M = 4.6, SE = 0.3) compared to constant onset rate (M = 3.9, SE = 0.2) (*p* = .016).

The two-way ANOVA showed a significant main effect of display, *F*(6, 78) = 5.28, *p* = .003 (see Figure 5). Post hoc testing showed that the main effect of display was due to significantly higher ratings of the visual display (M = 5.3, SE = 0.3) compared to the basic tactile with both constant (M = 3.4, SE = 0.3) and slow onset rates (M = 4.1, SE = 0.4) (ps < .001). And also significantly lower ratings of the basic tactile display with constant onset rate compared to the complex tactile display with high onset rate (M = 4.6, SE 0.3) (*p* = .047).

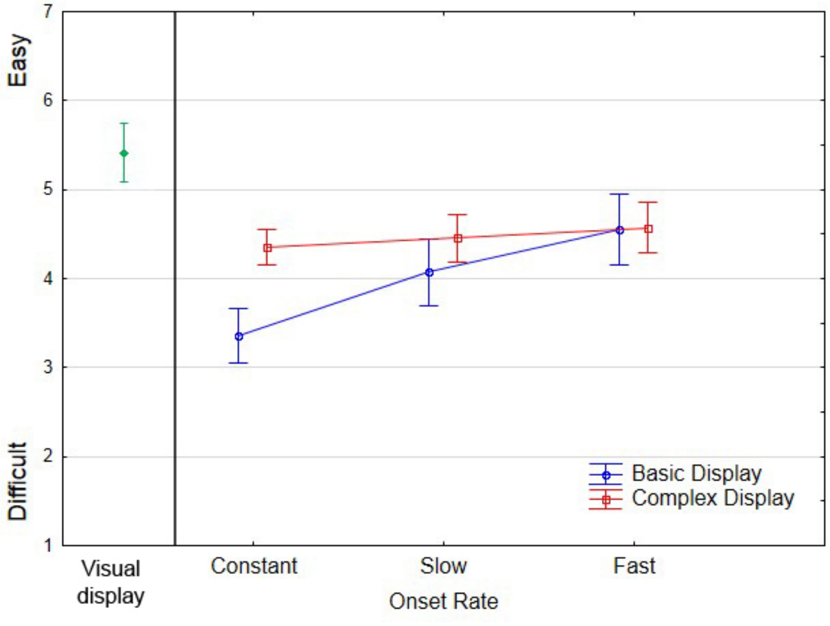


Fig. 5. Subjective ratings for the visual display and the two tactile drift displays with three onset rates

Discomfort. The participants answered two rating questions about if the display presentation was experienced as disturbing: if it caused discomfort and if it was annoying (1 = Not at all – 7 = Very much). The ANOVAs showed no significant differences. The ratings of discomfort were very low (M = 2.1) per display configuration.

Mental Workload. The participants answered one rating question about how they experienced their mental workload (1 = Very low – 7 = Very high). The ANOVAs showed no significant differences. The ratings of mental workload were moderately high (M = 4.8).

4 Discussion

With the tactile basic display lateral drift was significantly better controlled with the slow onset rate compared to the constant onset rate. The reason that performance was worst with the constant onset rate is most likely that this display configuration gave no information of the speed of the drift. That performance was best with the slow onset rate (fast onset was not significantly better than constant) was most likely that the resolution of the drift information was higher with a slower onset rate.

With the tactile complex display lateral drift was significantly better controlled with the constant onset rate compared to the fast onset rate. The reason for the better performance with the constant onset rate is most likely that the dynamic pattern itself gave information of drift speed. Whereas, slow and fast onset rates that gave dual coding of lateral drift made the information more difficult to interpret.

There were no significant differences in control of drift between the tactile displays and the already validated visual drift displays. Although the tendency of the one-way ANOVA indicates differences between the displays, lateral drift with the two tactile displays which provided best control of drift were more or less equal to the visual display (see Figure 4).

Controlling altitude, speed, and heading were in a sense secondary tasks, and performance on these tasks, can thus also be regarded as secondary measures of mental workload. That these tasks were controlled equally well with the tactile drift displays as with the visual drift displays gives further support to the possibilities of using a tactile drift display. Furthermore, the ratings of workload, although relatively high were at the same level with the tactile and visual display configurations. Also, the participants did not experience the tactile displays as discomforting.

The subjective ratings of how the drift information was experienced are somewhat counterintuitive, since these ratings do not reflect the performance measures of lateral drift. For example the basic drift display with slow onset rate was rated lower than the visual display, but performance was comparable.

The results indicate that for the basic tactile display onset rate can successfully be used as a secondary information channel. Whereas, for a complex tactile display configuration that uses the tactile pattern to present speed information, the best alternative was to use the constant onset rate. It is interesting that performance with the complex tactile display, which only used three levels to indicate speed, when used with constant onset rate was equal to the tactile basic display with slow onset rate that presented speed information on a continuous scale. This indicates that if a tactile display is properly coded, speed information can be successfully coded with only a few levels.

That performance with the best tactile drift display configurations (basic with slow onset rate and complex with static onset rate) was equal to the already validated visual display supports the possibilities of using a tactile display. This also confirms results from our previous research that has shown that lateral drift, in principal, can be controlled equally well with both visual and tactile drift displays [11].

A fundamental idea with tactile displays is to make them intuitive, and the results here are important since they provide valuable information about how tactile patterns can be used to increase display interpretability and thereby decrease lateral drift. However, most likely a tactile display will be used together with a visual drift display, in a bimodal drift display configuration. Therefore, further research is needed to investigate if a high level of perceptual momentum can be maintained when these suggested lateral tactile display configurations are used together with present visual display configurations.

References

1. Albery, W.B.E.: Rotary-Wing Brownout Mitigation: Technologies and Training: RTO Technical Report (TR-HFM-162): NATO (2012)
2. Castle, H., Dobbins, T.D.: Tactile displays for enhanced performance and safety Procedings at SPIE 5443, 269 (2004)
3. Chapman, R.J., Nemec, M.L.: The Evaluation of a Tactile Display for Dismounted Soldiers in a Virtusphere Environment. Paper presented at the Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting (2013)
4. Clark, J.M., Paivio, A.: Dual coding theory and education. Educational Psychology Review 3(3), 149–170 (1991)
5. Curry, I.P., Estrada, A.: Drift Cues from a Tactile Belt to Augment Standard Helicopter Instruments. International Journal of Applied Aviation Studies 8(1), 74–87 (2008)
6. Eriksson, L., van Erp, J., Carlander, O., Levin, B., van Veen, H., Veltman, H.: Vibrotactile and visual threat cueing with high G threat intercept in dynamic flight simulation. In: Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, Santa Monica, CA, USA, pp. 1547–1551 (2006)
7. Greenwood, T., Tsang, C.C.: Wickens' Multiple Resource Model, `https://wiki.ucl.ac.uk/display/UCLICACS/Multitasking` (Retrieved May 15, 2012)
8. Kelley, A.M., Cheung, B., Lawson, B.D., Rath, E., Chisson, J., Ramiccio, J.G., Rupert, A.H.: Tactile Cues for Orienting Pilots During Hover Over Moving Targets. Aviation, Space, and Environmental Medicine 84(12), 1255–1261 (2013)
9. Krausman, A.S., White, T.L.: Tactile Displays and Detectibility of Vibrotactile Patterns as Combat Assault Maneuvers are Being Performed. Army Research Laboratory, Aberdeen (2006)
10. Lif, P., Svenmarck, P., Oskarsson, P.-A.: Perceptual momentum for design of multimodal displays. In: De Waard, D., Brookhuis, K., Weikert, C., Röttger, S., Manzey, D., Biede, S., Reuzeau&, F., Terrier, P. (eds.) Proceedings HFES Europe Chapter Conference. Toulouse (2012), `http://hfes-europe.org`
11. Oskarsson, P.-A., Lif, P., Hedström, J., Andersson, P., Lindahl, B., Tullberg, A.: Visual, tactile, and Bimodal Presentation of lateral Drift in Simulated Helicopter. Paper presented at the Human factors and Ergonomics Society 57th Annual Meeting, San Diego, USA (2013)
12. Riddle, D.L., Chapman, R.J.: Tactile Language design. Paper presented at the Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting Boston (2012)
13. Wickens, C.D.: Multiple resources and performance prediction. Theoretical issues in ergonomic science 3(2), 159–177 (2002)
14. Wickens, C.: Multiple Resources and Mental Workload. Human Factors 50(3), 449–455 (2008)
15. Wickens, C., Hollands, J.G., Banbury, S., Parasuraman, R.: Multitasking corrected Engineering psychology and human performance, 4th edn. Pearson, New York (2013)

Development of Interaction Concepts for Touchless Human-Computer Interaction with Geographic Information Systems

Ronald Meyer[1], Jennifer Bützler[1],
Jeronimo Dzaack[2], and Christopher M. Schlick[1]

[1] Institute of Industrial Engineering and Ergonomics, RWTH Aachen University
{r.meyer,j.buetzler,c.schlick}@iaw.rwth-aachen.de
http://www.iaw.rwth-aachen.de

[2] ATLAS ELEKTRONIK GmbH, Sebaldsbrücker Heerstr. 235,
D-28309 Bremen, Germany
jeronimo.dzaack@atlas-elektronik.com
http://www.atlas-elektronik.com

Abstract. Interaction concepts in 3D GIS are yet limited to 2D input methods like mouse and keyboard. This work describes elaboration of a concept of touchless interaction for a prototype that aims to be used in maritime GIS applications. Experts from the maritime field have been interviewed to construct a rigorous scenario settled in the maritime field. Besides the planning and conversion of a stereoscopic GIS prototype a touchless interaction concept for stereoscopic environments under consideration of three different hand models is developed and presented. Implementation of these different hand models is planned for future evaluation.

1 Introduction

Geographic information systems (GIS) provide access to geospatial and context situated data for a wide area of applications and target a broad audience of users from novices to professionals in different areas. Traditional GIS applications are visualized 2D with planar top-view. Advances in computer graphics technology allow visualization of 3D scenes with increasing complexity in real-time. The visualization of GIS benefits from this development since spatial data can now be visualized in 3D virtual space by using modern graphics hardware. Virtual globes like Google Earth[1] or NASA World Wind[2] are versatile tools to exemplify the power of representing spatial relationships in a three-dimensionally rendered environment but both lack of an interaction method that naturally maps on their spatial dimensions. Interaction in three dimensions may be more direct and immersive since there is a shorter *cognitive distance* between a user's action and the system's feedback through *immersion* and *presence* in the virtual world. The feeling of presence, where the physical environment of the user is mentally

[1] http://earth.google.de

[2] http://worldwind.arc.nasa.gov/

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 589–599, 2014.

replaced with the virtual one, induces immersion into the virtual environment which facilitates interaction using natural tasks. This allows users to mentally build up complex mental models of e.g. how an interaction concept works [2].

Since publication of gesture-based game controllers like the Nintendo Wii[1] or the Microsoft Kinect[2] which intend users to interact via body motion or gestures a beginning dissolution of the WIMP (Windows, Icons, Mouse, Pointers) paradigm can be observed also in the non-entertainment area when profiling relevant contributions in the field of human-computer interaction of the recent years. Adaption of technology originating from the entertainment industry has become a famous phenomenon as convenient availability of efficient hardware with stable system abilities is available at low prices through mass market conditions. These new interaction technologies capturing hand and body motion in three dimensions facilitate a reassessment of spatial interaction in virtual globes and 3D GIS.

New devices allow new interaction offering six degrees of freedom (DOF) with millimeter or sub-millimeter accuracy in visual tracking of human body up to finger limbs [10] [17], e.g. the Microsoft Kinect or the Leap Motion controller[3]. These devices' abilities are worth to be investigated in respect to their input precision and reliability in context of a multimodal input for maritime GIS applications since previous research of 3D GIS in maritime context are more focussed on visualization and less on input modalities. Yu et al. see no availability of a true 3D GIS on the current market [18] which is due to several impediments, including deficiencies in data structuring and manipulation of various types of objects, 3D data analysis and large volumes of data.

Information processing in maritime navigation is generally defined through interaction with digital sea maps and virtual object manipulation for navigational purposes. Maritime vehicles and onshore maritime control rooms are likewise equipped with corresponding systems. Maritime safety greatly relies on these systems which must provide availability and accessibility of real-time information through human-computer interactivity. Best possible representation and interaction with data and information allows a high contextual awareness for system users and enhance decision-making processes in time-critical situations.

In this research a concept for touchless human-computer interaction in 3D GIS is developed putting it's focus on applications in maritime context, i.e. use case scenarios for maritime applications are investigated through interviews with experts on the maritime field. On that basis an interaction concept is developed including a hardware component-off-the-shelf prototype which consists of depth tracking devices like the Microsoft Kinect and the Leap Motion controller and a stereoscopic display which are operated through the open source GIS NASA World Wind in a stereoscopic environment. The prototype forms the basis for future evaluation of different touchless input methods for a maritime 3D GIS.

[1] http://www.nintendo.com/wii

[2] http://www.microsoft.com/en-us/kinectforwindows/

[3] https://www.leapmotion.com/

2 State of the Art

Two-dimensionally visualized GIS are prevalent on the commercial market. Mouse and keyboard are adequate input devices as the interaction on 2D maps can easily be achieved through the two degrees of freedom offered by mouse input. With visualizing spatial relations in three dimensions these input devices do not further comply the requirements of interaction and therefore become obsolete.

Modern multimodal input devices as the aforementioned Microsoft Kinect and Leap Motion controller allow a more direct input on spatial data since they provide 3D output data.

Recent research in the field of multimodal input in 2D and 3D GIS documents the identification of the existing problem, anyway solutions are still in a state of research [18]. Rauschert et al. see a necessity of a paradigm shift in the usage of GIS from classical WIMP interaction to multimodal input. Most functionality in GIS is subjected to expert users since special functions can only be accessed via repetitive menu and wizard tool usage [14]. Their research reports on a 2D GIS in a large screen environment that can be operated by spoken commands or free-hand gestural input such as pointing or outlining areas of interest. Fuhrman et al. conducted a user study where a 2D GIS was equipped with multimodal input modalities to rate basic user performance and to generally evaluate user acceptance [5]. The study was conducted under participation of ten voluntary geography graduate students having no or a maximum of three years experience with GIS. The prototype provided functions like data querying, navigating and drawing 2D primitives like points, circles, lines and free-hand drawing. During the study participants were asked to perform tasks like finding a certain map extract, load a data set via speech command or place a marker on the 2D map. The results show that participants with GIS experience had no noticeable advantage in completing tasks over participants with no GIS experience. The participants stated that the speech-based dialogue and gestural input method allowed them to interact with the system easily, without knowing about GIS concepts and database queries. The results indicate a higher learnability in the usage of multimodal systems concerning GIS while the system was generally accepted by the participants.

Gold et al. expand the proprietary 3D GIS *GeoScene3D* with functionality of a collision detection algorithm to support marine traffic monitoring. In addition to marine features, marine objects were integrated into the system, e.g. depth contours in coastal areas, navigational lights, anchorage areas, precautionary areas and even ship wrecks [6]. The interaction with the 3D GIS was constrained through the use of a wheel mouse. Gold et al. describe the interaction as quickly learnable by young and old users likewise but also mention constrains that occurred through the restrictions of mouse input: To deal with different affordances in spatial input Gold et al. implemented a manipulator which enabled the user between different modes to map the same mouse gesture to different types of input depending on the intention *GeoScene3D*. With *Kinoogle* Boulos et al. created a gestural interface to operate Google Earth via the Microsoft Kinect [1]. *Kinoogle* is a composition of middle-ware which maps gestural input captured by the Kinect and processed by OpenNI and NITE[4] to the standard interaction set of Google Earth. OpenNI[5] features processing of the image and depth stream of the Kinect while NITE provides extraction of human body motion and allows the definition of full body gestures. Boulos et al. describe full-body gestural input as to be quite exertive but as a great potential to improve user experience e.g., when interacting with large and stereoscopic displays.

Bruder et al. [3] conducted an evaluation on mid-air selection performance of virtual objects in a stereoscopic virtual reality table environment using three different selection techniques for stereoscopic environments with non-GI systems.

1. **Direct input**, where the user uses the index finger's tip to interact with the virtual object, bears accommodation problems for the human eye as the focus distance among the screen's surface and the finger tip are divergent. The accommodation of the index finger's tip makes the virtual object appear blurred to the user which breaks the stereoscopic effect as accommodation is an oculomotoric depth criterion for visual depth perception [15].
2. **Distant input**, where the user's index finger tip's position is mapped to a virtual cursor which is set off 10 centimeters to the index finger tip. While the index finger is still resided in the stereoscopic environment and blurred to the users eye a focus loss is avoided by mapping the interaction space into the virtual environment.
3. **Distant input with a virtual hand**, which is operated by the user in a distance to the user's real hand of 10 centimeters, similar to the second method. The user's real hand remains blurred in the stereo-scopic environment but is not focused during interaction tasks.

The results by Bruder et al. show a better performance in interaction tasks on using distant input and distant input with a virtual hand on doing a selection test based on Fitt's Law [4] with a slightly increased performance on distance pointing using a cursor.

3 Development of Interaction Concept

The development of an interaction concept for a GIS in maritime context requires expert knowledge in terms of scenario planning, task descriptions for future planning of use cases as well as special cases concerning the conception of a man-machine interface in on- or offshore use. For this purpose the ATLAS ELEKTRONIK GmbH provided experts from the maritime field that currently are involved into user interface development for onboard systems, with some of them having a seafaring background. The ATLAS ELEKTRONIK GmbH is a naval and marine electronics business company domiciled in Bremen, Germany, specialized on integrated maritime electronic systems and sonar sensors.

[4] `http://www.openni.org/files/nite/`

[5] `http://www.openni.org/`

3.1 Guided Expert Interview

A group of five maritime experts of the ATLAS ELEKTRONIK GmbH were surveyed via guided expert interviews [11], [12] to further investigate the field. All participating experts had basic experience with touch-based interaction in 2D GIS by the time of the guided interview which was divided into three categories: *Stereoscopic Visualization*, *Interaction* and *Scenario-based Application.*

Categories one and two analyzed the experts state of knowledge concerning each topic while category three reconsidered the first two categories in the context of possible scenario-based applications. Each category was arranged as a set of standardized questions concerning to each topic. The interview guideline's consistency was analyzed during a pretest with an extra individual participant. Each interview was scheduled to 45 minutes and the interviewer annotated the statements of the interviewees during the interview since none of the interviewees agreed on live audio recordings.

3.2 Interview Analysis

The analysis of results of the guided interviews illustrate possible fields of application for virtual reality GIS software with gesture-based input. The combination of a 3D visualization with a natural user interface was rated as beneficial by the experts for three maritime scenarios:

1. **Harbor Maintenance and Surveillance**
   Data for harbor maintenance and surveillance usually converges in onshore control rooms where data of security and safety systems is reviewed and analyzed in real-time. Besides the planning and coordination of arriving ships and harbor personnel data of harbor facilities like surveillance cameras and underwater inspection data is reviewed. A 3D virtual reality visualization having interactive access on real-time data of all installed sensors in a harbor offers possibilities of a central data reviewing element and mastering high information density.
2. **Offshore Wind Park Facility Surveillance and Maintenance**
   Offshore wind park facilities deserve special protection due to their high infrastructural relevance. Besides surveillance technology like cameras or radar, underwater sensors come into operation for maintenance tasks like echo sounders or under water cameras. An interactive 3D visualization can be a supportive tool for maintenance tasks but also implies challenges concerning reliability issues in terms of gesture-based interaction since an interactive system must reliably work aboard even under swelling sea states concerning to the experts which requires extra research.
3. **Offshore Bathymetry and Economic Geology**
   Bathymetric maps and sonographic recordings of ocean beds and sea grounds are 3D recordings generated by echo sounders. A 3D representation of recorded data in a virtual environment is beneficial for planning and evaluating the allotment of offshore resources concerning expert opinion. A combination of

echo sounder data sets with sea map data also allows safer maritime navigation in shallow waters. The use of a gesture-based interaction in swelling sea states also applies for this scenario which requires extra research.

The aforementioned scenarios form the basis for deriving use case descriptions for a planned prototype. The scenarios can be arranged into two general subjects. Scenario 1 and 2 both describe general interaction using a maritime GIS on- and offshore while scenario 3 is more confined by the idea of working on raw data sets for their analysis. For further classification of use case descriptions one of the three elaborated scenarios is selected for future planning and conversion: Harbor Maintenance and Surveillance. Harbors are central elements in maritime infrastructure with a multitude of safety-critical concerns. A quick access to requested data must be guaranteed with a reliable and robust human-computer interaction. With its multitude on actors and sensors the harbor security and maintenance scenario offers diverse potential for the definition of use-cases for gesture-based interaction in GIS. Safety critical issues in this scenario facilitate increased requirements on reliability. The interviews' results and analysis form the basis of the creation of a scenario-based interaction concept for touchless interaction with GIS in maritime context.

3.3 Basic Operations in Maritime GIS

Further development of the maritime GIS touchless interaction concept requires a look at GIS functionality in general. Goodchild summarized use case scenarios to possible use case groups concerning the functionality in GIS, each based on its conceptual sophistication [7]. Table 1 shows four of the originally six groups, extended with exemplary tasks including an interaction target which counts as manipulation object in the virtual system. The classification of the functionality into basic and complex indicates the number of repetitive or non-repetitive steps to achieve completion of the task. *Basic* indicates one step where *complex* indicates more than one step.

Table 1. Functionality for the maritime GIS application

Functionality	Exemplary Task	Target	Classification
query and reasoning	change map or data view	virtual camera	basic
measurement	measure area or track	virtual object	complex
transformation	manipulate or create data	virtual object	complex
descriptive summaries	tag and group objects or data	virtual object	complex

3.4 Hardware Setup

The implementation of the planned prototype is conceptualized as a component-off-the-shelf system (COTS) consisting of a Leap Motion controller for hand-gesture tracking, the Microsoft Kinect sensor for upper-body and head-tracking, and a 27 inch passive stereoscopic display, operated with an off-the-shelf laptop with dedicated graphics hardware. On the software side the open-source GIS NASA World Wind Java is used which provides all basic functionality of a GIS software and permits efficient expansion of the available source code with a gestural interaction. The system will be provided with an interface that allows the linking of external real-time sensor data for scenario-related experimental purposes.

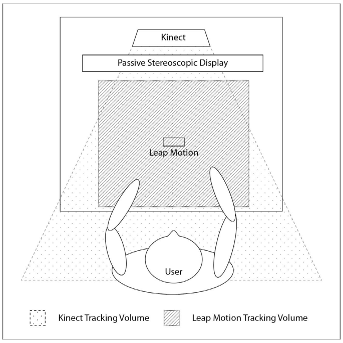


Fig. 1. Workplace as head-tracked stereoscopic environment with gestural interaction

Using both a stereoscopic visualization and coupling of the viewer's head to the virtual camera, enables an immersive virtual reality setting where the user's eyes are not covered as in a setting where virtual reality glasses as e.g. the Oculus Rift[6] are used. This setting allows easy integration of a virtual reality workspace into everyday workspaces [16] as the planned prototype is designated for use in harbor control rooms. Also hand-eye coordination remains unaffected in a stereoscopic head-tracked environment, which additionally allows usage of manual pointing tools for a possible tangible [8] interaction setup for future ideas. Therefore, the leap motion device provides high accuracy on two-hand input and manual tool recognition which can be held in the users tracked hand e.g. as pointing devices.

3.5 Gestural Interaction Model

The division of interaction tasks into basic and complex operations suggests an assignment of complex tasks to the user's preferred hand and execution of basic tasks to the user's non-preferred hand[9]. On the basis of the findings of Bruder et al. concerning interaction models and mid-air selection performance in stereoscopic environments [3] the choice on using the Leap Motion controller as gestural input device is consolidated for use in stereoscopic GIS. This affirms

[6] `http://www.oculusvr.com/`

the idea of a gestural input method with two-hand motion in front of a stereoscopic display. Hence, the concept for a distant input method with a virtual visualization of the user's preferred hand forms the general guideline for further elaboration of interaction concepts for the 3D maritime GIS application. While the preferred hand is visualized in the stereoscopic environment for complex task interaction the non-preferred hand remains unvisualized but should be used as input hand for basic tasks simultaneously.

The human hand consists of 16 rigid elements: Three phalanges of each finger plus the palm. The Leap Motion sensor's algorithmic processing in its current state is capable of tracking each finger by providing position and velocity of each finger tip and the position of the palm. These basic capabilities are sufficient to create concepts of differently visualized hand interaction models for later implementation into the maritime GIS as an interactive prototype and later evaluation. Three different visualizations of virtual hand models are presented in the following in general terms.

1. **Convex Hand Model.** In the convex hand model each finger with their finger tips and the palm of the preferred hand are represented by convex primitives 2. This approach involves potential occlusion of interaction objects which can potentially dissolved by adding different levels of translucency to the convex Hand Model.

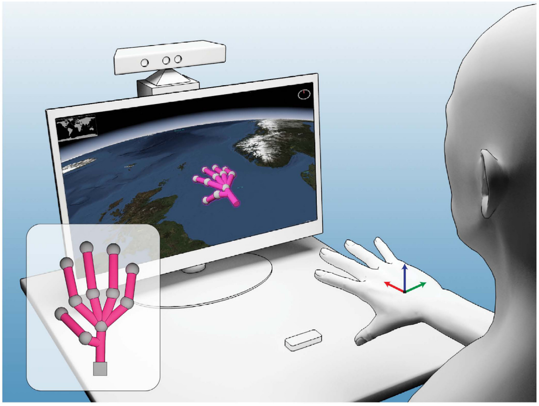


Fig. 2. Interaction model of preferred hand using a convex hand model

2. **Kinematic Hand Model.** The kinematic hand model is represented by a non-volumetric virtual model of the user's hands functional structure visualized through lines 3. The lines thickness is planned as to be adjustable (here shown using a thicker line visualization).

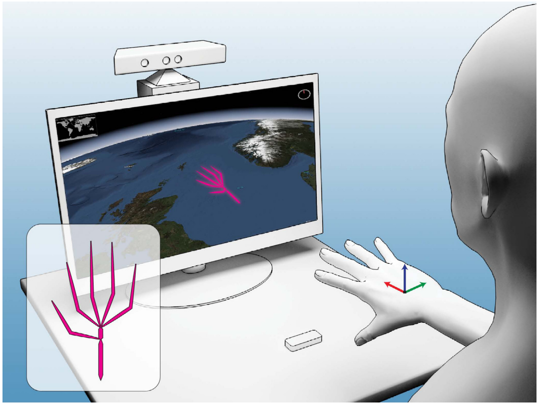


Fig. 3. Interaction model of preferred hand using a kinematic hand model

3. **Point Cloud Model.** The point cloud model represents the key features of the user's hand by visualizing finger tips and ends, and the palm's mass centre by points 4. The points' size is planned as to be adjustable.

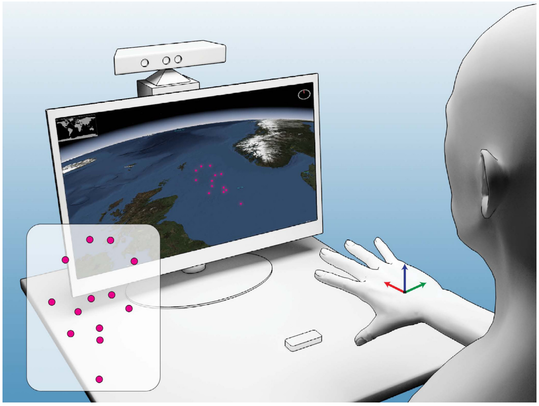


Fig. 4. Interaction model of preferred hand using a point cloud model

4 Summary and Outlook

This work elaborated a basic interaction concept for touchless human-computer interaction for geographic information systems in maritime context. Experts from the maritime field were interviewed for the development of possible maritime scenarios and as a basis for the future planning of use cases. The scenario "Harbor Security" has been identified as having a high potential for future conversion since this scenario implies a variety of actors, potential interaction objects and different sensors. Additionally, the scenario features particular requirements in terms of reliability as harbors are highly frequented areas where clear view on available system data with rapid cognitive information processing and reaction times are indispensable. Basic GIS operations have been tabled and been purposed as basic and complex interaction tasks for the two-hand touchless interaction concept, in which the user's non-preferred hand operates basic tasks while the users preferred hand operates complex tasks. Different hand models for virtual object interaction have been elaborated and presented. All necessary hardware and software components were arranged for a first stereoscopic GIS prototype. Future steps are the implementation of the elaborated scenario into the stereoscopic GIS environment and the integration of the COTS sensor hardware into the GIS software. The implementation of the presented touchless interaction model will be followed by an evaluation in which the three virtual hand models, presented in this work, will be further investigated.

References

1. Boulos, M.K., Blanchard, B.J., Walker, C., Montero, J., Tripathy, A., Gutierrez-Osuna, R., et al.: Web GIS in practice x: A microsoft kinect natural user interface for google earth navigation. International Journal of Health Geographics 10(1), 45 (2011)
2. Bowman, D.A.: 3D user interfaces: Theory and practice. Addison-Wesley, Boston (2005)
3. Bruder, G., Steinicke, F., Strzlinger, W.: Effects of visual conflicts on 3D selection task performance in stereoscopic display environments. In: Proceedings of IEEE Symposium on 3D User Interfaces 3DUI, pp. 115–118. IEEE Press (2013)
4. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47(6), 381 (1954)
5. Fuhrmann, S., MacEachren, A., Dou, J., Wang, K., Cox, A.: Gesture and speech-based maps to support use of GIS for crisis management: A user study. AutoCarto 2005 (2005)
6. Gold, C., Chau, M., Dzieszko, M., Goralski, R.: 3D geographic visualization: The marine GIS. In: Developments in Spatial Data Handling, pp. 17–28. Springer, Heidelberg (2005)
7. Michael, F.: Goodchild. The use cases of digital earth. International Journal of Digital Earth 1(1), 31–42 (2008)
8. Ishii, H.: The tangible user interface and its evolution. Commun. ACM 51(6), 32–36 (2008)
9. Kabbash, P., MacKenzie, I.S., Buxton, W.: Human performance using computer input devices in the preferred and non-preferred hands. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, pp. 474–481. ACM (1993)
10. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 12(2), 1437–1454 (2012)
11. Küsters, I.: Narrative Interviews: Grundlagen und Anwendungen. Springer, Heidelberg (2009)
12. H.O. Mayer.: Interview und schriftliche Befragung: Grundlagen und Methoden empirischer Sozialforschung. Oldenbourg Verlag (2012)

13. Petit, M., Ray, C., Claramunt, C.: A contextual approach for the development of GIS: application to maritime navigation. In: Carswell, J.D., Tezuka, T. (eds.) W2GIS 2006. LNCS, vol. 4295, pp. 158–169. Springer, Heidelberg (2006)
14. Rauschert, I., Sharma, R., Fuhrmann, S., Brewer, I., MacEachren, A.: Approaching a new multimodal gis-interface. In: Proceeding of the 2nd International Conference on GIS GIScience, CO, USA (2002)
15. Schlick, C., Luczak, H., Bruder, R.: Arbeitswissenschaft. Springer, Heidelberg (2010)
16. Ware, C., Arthur, K., Booth, K.S.: Fish tank virtual reality. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, CHI 1993, New York, NY, USA, pp. 37–42. ACM (1993)
17. Weichert, F., Bachmann, D., Rudak, B., Fisseler, D.: Analysis of the accuracy and robustness of the leap motion controller. Sensors 13(5), 6380–6393 (2013)
18. Yu, L.-J., Sun, D.-F., Peng, Z.-R., Zhang, J.: A hybrid system of expanding 2D GIS into 3D space. Cartography and Geographic Information Science 39(3), 140–153 (2012)

Spyractable: A Tangible User Interface Modular Synthesizer

Spyridon Potidis and Thomas Spyrou

University of the Aegean, Dept. of Product and Systems Design Eng. Hermoupolis, Syros, Greece
spotidis@aegean.gr, tsp@aegean.gr.

Abstract. The purpose of this paper is the exploration of the possibilities that Tangible User Interface (TUI) may offer in the area of sound synthesis, by reconfiguring the functionality of the existing TUI tabletop musical instrument called "Reactable" and redesigning most features, adjusting it to a synthesizers needs. For this research we analyzed sound properties, physics and formation, as well as how human used these features to synthesize sound. Afterwards we present the properties and advantages of TUI technology, and its use in sound and music, distinguishing Reactable, for being the most even musical instrument using TUI. As an outcome we develop and present an initial prototype modular synthesizer, called Spyractable. Finally, we subject Spyractable to users' evaluation tests and we present the outcomes, making suggestions for further investigation and design guidelines.

Keywords: Sound wave, harmonics, modular synthesis, modules, tangible user interface (TUI), tangibles, patches, graphical controllers.

1 Introduction

Reactable is a tabletop TUI musical instrument, uses specially developed fiducial markers and gestures to produce intimate music [1]. Design was involving the development of a multiuser organ, with simple and intuitive use, without any guidance or manual needed. It produces music from the first interaction and supplies with optical feedback the users, so to make known of its state and use [2]. Users manipulate finger gestures and special fiducial markers (the tangible tokens), that are connected dynamically in a modular analogue synthesis way, on the reactable's surface, to change sound structure, control its parameters and create music [3].

Even though Reactable has sound synthesis capabilities, it concentrates more on music production and less in sophisticated sound synthesis, introducing just the basic modular synthesis principles [2].

In this paper, sound and sound synthesis, TUI's technology and Reactable's main features and software have been explored and an appropriate interface has been designed to adjust Reactable's features to a sound composers' needs. Finally, a small, but adequate, version of a tangible user interface tabletop modular synthesizer has been designed and implemented, so to evaluate if such a venture can contribute to a

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 600–611, 2014.

simpler and faster way of synthesizing sounds, with a clearer view of the user's actions that enhances experimentation and creativity. The synthesizer is called **Spyractable** and is able to implement classic and more complex, methods of synthesizing, addressed to users with a long range of knowledge on sound synthesis, starting from knowing the basics to more sophisticated skills.

1.1 Sound

When a sound source is being vibrated, it causes a propagation of periodic changes of the atmospheric pressure. This propagated disturbance is called wave and as travels through air (or any other compressible media), it carries energy. The wave's crests correspond to the compressions, troughs to the rarefactions of the atmospheric pressure and zero prices to atmospheric equilibrium. When a wave reaches an ear, and fulfills some conditions as described later, it stimulates the acoustic nerves and the brain interprets these waves as sounds. The simplest wave that can be discrete, is described the sinusoidal wave. Most sounds in nature are not simple but complex, formed by many different sinusoidal waves. The human ear analyzes sound in its consisted sinusoidal waves and sends different signals to the brain [4, 5].

Wave has the following objective and corresponding subjective characteristics that correspond to its nature and to human interpretation.

Frequency and Pitch: "Frequency" is the characteristic that describes how often the periodical disturbance is being repeated and is measured in Hertz. Frequency is perceived as "pitch". Human ear corresponds to sounds from 20 Hz to 20 kHz [7].

Volume and Loudness: The wave's crests and troughs give the amplitude of the wave that represents the amount of energy carried within the wave (the volume) and is being measured in watts/m2. Human interprets volume as "loudness". Different frequencies with same volume are not perceived at same loudness [4], [5].

Phase: The moment of the time circles of a period when the wave started is called "Phase" and being measured in degrees [5], [6].

Time Envelope: Every sound has a start and an end. The curve describing how the amplitude is being developed in that period of time is called Time Envelope. Time envelope is usually divided into four segments: Attack, is the time from the triggering of sound, till it reaches the sound's crest amplitude. Decay is the time sound is taking from crest amplitude to normal level of sound, as long as it endures. Sustain indicates the amplitude the sound has during its duration and is a percentage of maximum amplitude achieved at Attack time and last segment is Release time. The time the sound endures till it stops, after the sound source stops to vibrate [4], [6].

Harmonic contain – Timbre: Most sounds are sums of many different sinusoidal waves, with each one to have its own frequency, amplitude, phase and envelope. This sum is called harmonic contain and it is this characteristic that make people distinguishing the many different sounds, characterizing as "Timbre". In every harmonic contain there is one louder wave (fundamental) that determines the pitch of the sound. In order a sound to sound fine, all the other waves have to be integer multiples and submultiples of fundamental's frequency, obeying the formula $f_n=f_f*n^{\pm 1}$ The multiple frequencies called overtones and the submultiples sub - harmonics. Every other frequency causes the sound to sound bad / inharmonic [4], [6].

1.2 Synthesizers.

A sound synthesizer is a device that has the ability to produce a wide range of sounds, either imitating existing organs, or producing new sounds that don't exist in nature. Synthesizers use various methods and circuits to handle electrical and digital signal, as waveform analogue, and turn it into sound. It has three parts: **Controller** that sets the pitch and other factors of sound. **Speakers** that turn electrical signal into sound and the **generator** that carries all the appropriate equipment to produce sound. There are three main kinds of synthesizers: Analogue that use electrical signal, digital that uses digital signal and VST's that are computer software [6], [7].

Synthesizers, in order to handle sound, use discrete components to do separated elaborations and formations to the signal, till it becomes sound. The basic units are:

- **Oscillator:** Circuit that produces alternating signal in a circular periodical change, just as a wave. It's the mother of sound [4], [6].
- **Low Frequency Oscillator:** This unit is also an oscillator that produces a non hearable signal between 0.1 Hz and 20 Hz. Its purpose is to slightly modulate other module's factors. If it modulates oscillator then we have "vibrato", if it modulates the amplifier's gain, we get the "tremolo" effect [4], [6].
- **Amplifier:** It multiplies signals amplitude, in order the sound to be hearable. It is also used to "shape" a sound by using a time envelope to its signal exit [4], [6].
- **Envelope Generator:** Unit that produces electric control voltage or digital command. As it name reveals, it is the unit that generates time envelopes [4], [6].
- **Filter:** This is a unit that clips or weakens frequencies within a range, defined by its kind and attributes [4], [6].
- **Effect Processors:** Special circuits that modify the acoustic signal in a way that has to do with its environmental behavior [4], [6].
- **Mixer:** Circuit that adds signals, at specified volume levels, into one unique signal.
- **Modulator:** Modulation is a procedure where a signal modifies some characteristics, as frequency, phase, amplitude and harmonic contain, of another signal. This method is used to shape signals and create sounds or within effects [4], [6].
- **Sequencer:** It's a district, additional unit that produces notes [4], [6].

There are nine basic methods of sound synthesis, but VSTs make it possible to produce some more hybrids. The most important Synthesis techniques:

- **Modular Synthesis:** The base of all analogue methods, the first method used to form different sounds. The composer connects modules using simple cables in order to make the desirable sound, forming many different synthesis types [6], [8].
- **Additive:** It does the opposite procedure of ear, meaning it uses many oscillators, producing frequencies with its own phase, amplifier and time envelope [6], [7].
- **Subtractive:** With two or more oscillators and a mixer we produce a complex waveform that is being filtered with one or more filters, so to obtain the harmonic contain the composer wants [6], [7], [9].

- **Physical Modeling:** It is a method, which uses mathematical models to simulate the cause of the sound, as it happens in nature. A computer calculates the sound wave that will be produced from the knocking of a string with a hammer in a wooden box for example, and oscillates the oscillator in that manner [6], [9].
- **Sample Based:** This kind of synthesis uses recorded samples of organs in all the frequencies. The controller triggers the oscillator to play the corresponding frequency. Usually two or more oscillators are used, then mixed and follow the common procedure, filter, amplifier and effects with LFOs and envelopes. [6], [9].
- **Frequency Modulation:** As described before, frequency modulation uses a hearable signal – modulator- to modulate the frequency of another signal-carrier. Carrier sets the fundamental frequency and modulator the overtones. Overtones volume depends on modulators and carriers amplitude ratio [6], [7], [9].
- **Phase Modulation:** Modulator is a special oscillator that is able to change phase circle velocity. If, for example, phase gets from 0 to 360 in half time, this means that the period of the wave lasts have the time so the frequency is doubled [6], [10].
- **Linear Arithmetic:** Digital kind of synthesis that combines subtractive and sample based methods. It uses two mixed tones. Each tone is made by two sound partials. Partial P is a sample based sound made by a sample oscillator and a time variant amplifier. Partial S is a subtractive sound made by an oscillator, time variable filter and amplifier. The partial P sets the Attack of the sound and partial S the decay, sustain and release of the sound [6], [7].
- **Wavetable:** Digital method that uses a matrix of samples. Some samples will constitute attack and decay and another sustain and release. [6], [7], [11].
- **Granular:** Another computer based synthesis. It is similar to linear arithmetic with many parts of samples lasting less than 50msec (grains). These parts form sound shadows that can be treated like simple waves later on [6], [12].

1.3 Tangible User Interface (TUI)

Tangible User Interface, are graspable, physical or embodied user interface with least differences, and are a physical handle to a virtual function that is being used for one and only dedicated manipulation [13]. In TUIs, physical objects (tokens) are both controllers of digital information and physical representation of it [14].

TUIs have the following attributes regarding representation and control [14]:

- Physical representations, computationally coupled to underlying digital info.
- Physical representations embody mechanisms for interactive control, using movement, rotation, placing and other manipulations to control the system.
- Physical representations are perceptually coupled to actively mediated digital representations. Both physical and digital representations play the same important role in representation and are co-depended.
- The physical state of interface artifacts partially embodies the digital state of the system. Even with a switched-off system, tokens may represent, with their state, the implied functionality of the system.

In order an interface to be Tangible, It has to embody the following properties [13]:

- Space-multiplex both input and output: This means that each controllable function has a dedicated controller, occupying its own space.
- Allow for a high degree of inter-device concurrency both for input and output.
- Increase the use of strong specialized input devices. Physical artifacts that control the interface must have exclusive, dedicated control area.
- Have spatial-aware computational devices.
- Have high spatial reconfigurability of devices and device context. Physical controllers may not be used at a specific moment during a handle, but their presence in space, keeps reminding their functionality.

Sheiderman identifies three basic properties of direct manipulation interfaces [13]:

- Continuous representation of the object of interest.
- Physical actions or labeled button presses instead of complex syntax.
- Rapid incremental reversible operations whose impact are immediately visible.

Advantages of TUIs [14].

- It encourages two handed interactions.
- Shifts to more specialized, context sensitive input devices;
- Allows for more parallel input specification by the user, thereby improving the expressiveness or the communication capacity with the computer;
- Leverages off of our well developed, everyday skills of prehensile behaviors for physical object manipulations.
- Externalizes traditionally internal computer representations.
- Facilitates interactions by making interface elements more "direct" and more "manipulable" by using physical artifacts.
- Takes advantage of our keen spatial reasoning skills.
- Offers a space multiplex design with a one to one mapping (control – controller).
- Affords multi-person, collaborative use.

TUI systems have already been used successfully in learning processes, concerning narrative or rhetoric programming, molecular biology or chemistry, physics and dynamic systems [16]. There have been an enormous number of applications that take advantages of TUI in order to produce sound or music [17]. They could be separated in two big categories: Table tops and appliances. Tabletops usually use cameras and embodied sensors to input the handles information and screens or projectors to output the digital representation. The interaction takes place at a specific space. Such systems are Audio D-Touch [18], Audiopad [19], Smallfish [20], Jam o-drum [21] and Reactable [1]. Appliances use electronic tokens carrying the digital representation on them and spatial standalone interacting within their parts (e.g. blocklam [22] and audiocubes [23]). Various other application domains [15]:

- Information storage, retrieval, and manipulation.
- Information visualization.
- Modeling and simulation.
- Systems management, configuration, and control.
- Education, entertainment, and programming systems.

1.4 The Reactable

The Reactable is an instrument, which seeks to be collaborative, multiuser, intuitive, giving no manual or instructions, sonically challenging, non-intimidating instrument, learnable and masterable, suitable both for novices and advanced electronic musicians, for home use, studio or live performance of electronic music [24], [25], [26]. Regarding its functionality, Reactable is based on a translucent round table on which, users interact by moving tokens, changing their position and controlling with these actions the topological structure and the parameters of a sound synthesizer.

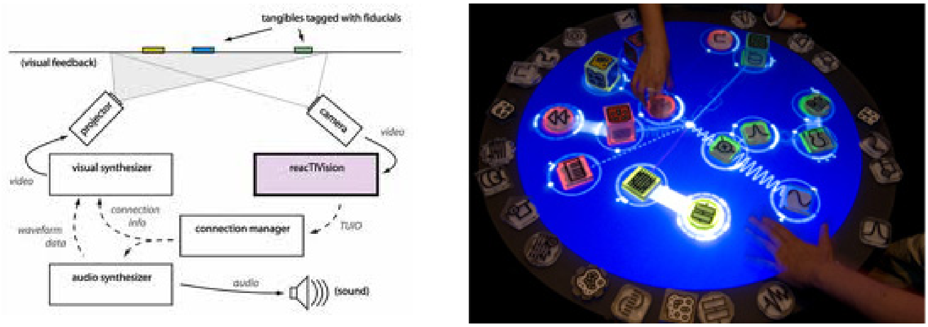


Fig. 1. The Reactable architecture and live action snapshot

Moreover beneath the table, there is a projector dynamically drawing animations on its surface, providing a visual feedback of the state of the synthesizer [24], [26]. Every token brinks a special fiducial marker that is been read by a camera, placed beneath the surface. Software, specially developed for Reactable called “reacTIVision”, reads tokens’ id, orientation, as well as finger placing and gestures, producing information about each token’s position, rotation angle, fingers’ positions and time related sizes, as speed, acceleration etc. [25], [27], [28]. This data is send to connection manager software that will make the appropriate calculations about tokens’ state, based on orientation and proximity, producing control data for the sound and visual feedback [24], [26]. Sound synthesizer is based on modular synthesis principals whereas every token represents a module and their orientation will dynamically set up the desirable connections of these modules [24], [25], [26]. Due to its goals, Reactable is equipped with various sample players, melodic and rhythmic, effects, filters, oscillators and LFOs, sequencers and stuff that will help user to immediately produce music, either his knowing how Reactable works or not [24], [25]. The visual feedback, produced by a visual synthesizer, provides user information about token’s state, sound’s state and modular connections, drawing the formed waveform that "exits" every module. Moreover, it draws graphical controllers, such as sliders, pop-up menus and secondary modules, controlled with fingers [24].

2 Introducing the Spyractable

2.1 Concept and Goals

Getting knowing the Reactable, a simple idea was born. What if all this technology didn't serve the purposes of reactable, an intuitive, ready-to-play organ for electronic music based on modular synthesis, but used to synthesize sounds, an ability that sure Reactable has, but not in a highly sophisticated manner. Would a combination, of computer based modular synthesizer and tangible user interface, offer new possibilities and facilitations to a sound composer? Spyractable was developed to research for this hypothesis. Our goal was to make a computer – based tangible user interface modular synthesizer that would offer to users the opportunity to facilitate sound synthesis in:

- Achieving a target – sound.
- Encourage experimentation.
- Save time.
- Cover user's needs for synthesizing sounds.
- Give clear image about how the sound is being synthesized and what the user did with no complex handles that disorientate composer from synthesis and engross him with button handles and way-finding through confusing menus.

Users and Usage Scenarios: As a user, we define anyone who wants to synthesize sound and is aware of the basic modular synthesis knowledge. We want user to be able to create a sound with various ways, modify a sound as desired in a live time expression situation and correct wrong options as soon as possible, with always be aware of what's happening.

Design and Implementation: In order to have an adequate synthesizer to complete our research it was decided to develop one that will surely implement a modular synthesizer and moreover give a little taste of other synthesis methods, in short mode. It will have various oscillators, filters, effects, LFOs, envelope generators and amplifiers, and could be able to use various synthesis methods. The technology was known, but since we are developing a completely new task, we have to adjust most of the features, keeping the dynamic patching in a modular metaphor by using reactivision's fiducial engine and Reactables architecture. Since this project is done in academic environment and purposes, it was decided to use open source software.

The main program runs in processing, a java based program for graphics that runs in its own compiler [29]. Sound synthesizer was made with pure data, a visual programming language, member of the patcher family [30]. Processing runs the Spyractable. It accommodates the connection manager, receiving messages from reactivision via TUIO library, generates the graphics (visual synthesizer) and reads the pure data file via libpd library.

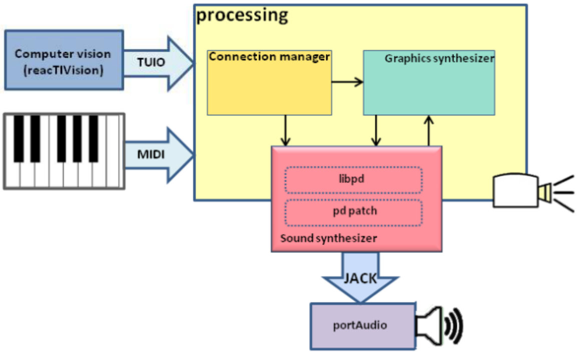


Fig. 2. The Spyractable's software architecture

Spyractable Interface: After draft development and evaluations, we came up with a horizontal interface, where the user stands in front of the appliance surface and puts the tokens in a readable way, developing the modular chain from left to the right. This way he gets the most of the given space and develops his thoughts the way he has learn to present them (according to the west civilianization, but this is easily change).

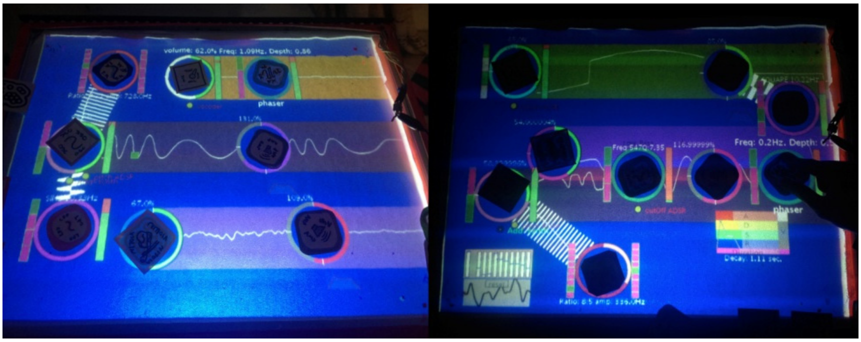


Fig. 3. Spyractable's aspects in action

We have developed 13 tokens including two amplifiers with envelope control, two filters with cut-off frequency envelope control and ability to change filter kind, one time delay effect, one chorus / phaser / chorus+phaser effect, one mixer. Three sample oscillators (violin, trombone and trumpet), a noise oscillator, a microphone input with pitch bend and vocoder, a multi-oscillator that user can choose between sine, saw, triangle, square and their sums wave, pitch envelope and velocity zone control and a sine oscillator with pitch envelope control and an additive synthesizer for forming the waveform by setting fundamental's and overtones' volume and phase (0° or 180°), and velocity zone control. Additional to these, we also made an LFO that can be connected either to wave oscillators or the amplifiers that includes sine, triangle, square and saw waveforms, a master volume and finally a modulator that modulates the sine oscillator's frequency (FM synthesis) with a sine oscillator. The user can choose between the modulation index and the modulator's / carrier's frequency ratio. The pitch is input via midi keyboard.

Hardware: For the hardware we used a 32" Plexiglas surface (4:3) , 45 infrared led lights SFH 485 Osram 880nm , a sony ps3 eye camera without infrared filter, an infrared light pass filter 850nm and a Toshiba TDP45 projector. The software was hosted by a Mac book snow leopard.

Interacting with Spyractable: In order to get a sound, user has to put an oscillator on the surface. Immediately it is connected to the main amplifier, controlled by the master volume. The sound, after is triggered by the midi keyboard controller, is continuous since no volume envelope is applied to shape it. In same case, sample oscillators just play the whole triggered sample till is finished. Next step for the user is to connect an amplifier into the chain. Amplifiers hold the volume ADSR time envelope. In order to make a modular chain, every oscillator draws a color zone, whatever patch (the token accompanied by graphical data forming a module) is in that zone, is dynamically connected in a sound chain, starting with the oscillator at the left and ending at the auxiliary exit at the right side of the screen, connected in turn. The connection between two patches is visualized with the waveform that travels from one to the next module, just like Reactable. LFO and modulator are dynamically connected through proximity rules, to the nearest available patch. All the parameters can be changed either with rotation of the token or with finger handling the side graphical controllers (radio buttons, sliders, switches etc). In order to change one parameter the biggest number of movements, a user may do, in the worst case scenario, is three finger actions, all clearly attached to the controllable patch, saving time and easily clarifying the sound modification. The sound zones may be overlapped. In this case whatever module is in the common zone area accommodates both sound signals derived from its left side.

3 Testing and Evaluation

For the final evaluation of Spyractable, we proceed to usability testing by using two methods. Firstly candidates performed a "Think Aloud" Usability Test [31]. During the test they were given 5 tasks to do. First task was to find how Spyractable works and what each token does. After this we explained to them whatever they didn't find out and proceed to the next tasks. The next three tasks included synthesizing a given sound, with scalable difficulty from task to task. The last one was to make a sound of their own taste. The purpose of these tests was to find how intuitive use Spyractable may have, what usability problems it has, as far as it concerns connectivity, logic and control (both token handling and GUI), how fast, or slow, a sound – goal can be achieved and how it does with free experimentation.

Usability test was supplemented with a semi-structure interview, willing to found out more about what candidates liked or didn't like, what incommoded them, what more did they expected and what surprised them.

Five candidates participated the test, with different level of knowledge in synthesizers (novice to experts), different cultures (analogue devices to complex computer programs) and different focus in synthesis (hobby sound creation to professional sound designers).

Various conclusions were given about the way of use and the behavior Spyractable faced, depending on the categories of users, but as far as it concerns usability problems, most people agreed by finding the same ones.

As far as it concerns logical mistakes and disadvantages, most people didn't like the view of GUI elements or didn't notice some of them at all. All the candidates looked up for the volume envelope to be controlled by the oscillators' patch, even though this is incorrect for the modular logic. Expert users though modulator was for Ring modulation effect, even though they use fm synthesis on VSTs, and many problems were faced with the rotation of patches since when it reached its maximum price it hopped back to the minimum and vice versa.

As far as it concerns usability problems, most people couldn't handle patches on the surface and either they were hiding the GUI elements or had to move them to other positions, sometimes whole of the sound chain. Another similar problem that contributed the first one was the ergonomics of the set-up. Candidates had to either rise from their positions to put a token on the upper space of the screen, or put them at the down side till they couldn't fit. Another serious notice was that most of the candidates were using only one hand to handle them and the other to trigger notes from the keyboard, even though they had alternatives for note triggering. This is probably a matter of habit, because since they were helped with keyboard, they started using both their hands. The third problem was in some cases a though, and in others an incident. What happens if you accidentally step on the surface, displace the tokens or change a GUI given option, moreover during live performance? This mostly had to do with Spyractable's early view that didn't inspire any confidence but on the other hand looked very fragile. Last but not least was the question how it will save a sound, how it will reload it and how the reloaded sound will be controlled? But since we don't come up in design with that matter, we don't really know whether there will be a problem or not.

To answer our assumptions and thoughts, despite the upper problems, Spyractable impressed most of the candidates, willing to fix the problems and expand it with more effects and features. It does shorten synthesis procedure time with fewer steps till the final goal, compared with what candidates used to use. Candidates also claimed that they understand some things they hadn't clear in their mind, and it was much more playful and mind absorbing than the synthesizers they had with much easier manipulation. To end up with we believe Spyractable to enhance creativity, since candidates were truly happy to make their own sounds and took much time for this, used as much patches as possible, even asked for more and impressed by using techniques they didn't have in mind (like the two amplifiers or parallel use of sample oscillator and simple oscillators) and proposed creative ideas to elevate the fan.

4 Future Development and Research

All the conclusions came up with some standards that are:

- It was a brand new experience for all the candidates and were impressed.
- It was a small implementation with the least synthesizer features. This fact made it easier to build an easy to use interface.
- The set up was truly very dreadful to fall apart.

So far TUI looks really helpful with sound synthesis, and most of Spyractable's problem seem to be easily, or not so easily, solved, but if we really want to see if a TUI synthesizer can stand as a commercial product, we have to test it with better hardware, and much more tokens and complicated modules, tested by candidates who are familiar with TUI technology.

References

1. Jordà, S., Geiger, G., Alonso, M., Kaltenbrunner, M., Rouge, B.: The reacTable: Exploring the Synergy between Live Music Performance and Tabletop Tangible Interfaces, TEI 2007, LA, USA. ACM (2007)
2. Geiger, G., Alber, N., Jordà, S., Alonso, M.: The Reactable: A Collaborative Musical Instrument for Playing and Understanding Music. Her&Mus (Heritage & Musicography) N.4, 36–43 (2010)
3. Jordà, S.: The Reactable: Tangible and Tabletop Music Performance. In: ACM Conference on Human Factors in Computing Systems, Atlanta, GA, USA (2010)
4. Loukas, X.: Sound – Music & Technology (Echos – Mousiki & technologia), vol. 1. Sygchrony Mousiki Publishing (1992)
5. Dodis, D.: Sound sampling: Creation with contemporary technology (Echolipsia: E dimiourgia me ti sygchroni technologia). Ion Publishing (1995)
6. Symvoulopoulos, A.: Vocabulary of Contemporary music technology (Lexico oron sygchronis mousikis technologias). Filippos Nakas Publishing (1994)
7. Arnaoutoglou, D.: Technology of music (Mousiki Technologia). Filippos Nakas Publishing (1993), ISBN: 978-0-00290-116-1
8. William, H.: In depth Feture: Inside Synthesis 003 – Modular Basics / Introduction to Modular Synthesis – Sonicstate.com, http://www.sonicstate.com/articles/articlecfm?id=147
9. Chiotis, M.: One glance at VST Synthesizers (Mia matia sta VST synthesizers). Sound Market Magazine (2010)
10. Wiltshire, T.: Phase Distortion Synthesis – Electric druid / Synth DIY pages, http://www.electricdruid.net/index.php?page=info.pdsynthesis
11. Bristow-Johnson, R.: Wavetable Synthesis 101, A Fundamental Perspective. Journal of Audio Engineering Society (1996)
12. Kuehnl, E.: Granular Synthesis, http://adagio.calarts.edu/~eric/gs.html
13. Fitzmaurice, G.W.: Graspable User Interface. Ph.D. Thesis. s.l. University of Toronto, Dept. of Computer Science (1996)
14. Ullmer, B., Ishii, H.: Emerging Frameworks for Tangible User Interfaces. IBM Systems Journal 39(3.4) (2000)
15. Fitzmaurice, G.W., Ishii, H., Buxton, W.: Bricks: Laying the Foundations for Graspable User Interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI 1995, Denver Colorado, USA. ACM (1995)

16. Marshall, P.: Baton Rouge: Do tangible interfaces enhance learning? In: 1st International Conference on Tangible & Embedded Interaction (TEI 2007), LA, USA. The Association for Computing Machinery Inc., ACM (2007) ISBN:977778-1-59593-619-6.
17. M. Kaltenbrunner.: Tangible music, `http://modin.yuri.at/tangibles/`
18. Costanza, E., Shelley, S.B., Robinson, J.: D-touch: A consumer-grade tangible interface module and musical applications. In: Proceedings of Conference on Human Computer Interaction, HCI 2003 (2003)
19. Patten, J., Recht, B., Ishii, H.: Audiopad: A Tag-based Interface for Musical Performance. In: Proceedings of the 2002 Conference on New Interfaces for Musical Expression, Dublin, Ireland (2002) ISBN: 1-87465365-8
20. Fujihata, M.: Furukawa. K., Munch. W.: Notes on Small Fish. ARS ELECTRONICA. [Η*λεκτρονικ*ό], `http://90.146.8.18/en/archiv_files/20001/E2000_306.pdf`
21. Blaine, T., Perkis, T.: The Jam-O-Drum Interactive Music System: A Study in Interaction Design. In: Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, Brooklyn, New York. ACM (2000) 1-58113-219-0/00/0008
22. Newton-Dunn, H., Nakano, H., Gibson, J.: Tangible Interface for Interactive Music. In: Proceedings of the 2003 Conference on New Interfaces for Musical Expression (NIME 2003), Montreal, Canada (2003)
23. Schiettecatte, B., Vanderdonckt, J.: AudioCubes: A distributed cube tangible interface based on interaction range for sound design. In: Proceedings of the Second International Conference on Tangible and Embedded Interaction (TEI 2008), Bonn, Germany. ACM Press (2008) 978-1-60558-004-3
24. Jorda, S., Kaltenbrunner, M., Geiger, G., Bencina, R.: The reactable*. In: Procs. of the International Computer Music Conference (ICMC 2005), Spain, pp. 579–582 (2005)
25. Geiger, G., Alber, N., Jordà, S., Alonso, M.: The Reactable: A Collaborative Musical Instrument for Playing and Understanding Music. Her&Mus (Heritage & Musicography) (4), 36–43 (2010)
26. Jordà, S., Geiger, G., Alonso, M., Kaltenbrunner, M.: The reacTable: Exploring the Synergy between Live Music Performance and Tabletop Tangible Interfaces. In: TEI 2007, Baton Rouge, LA, USA. ACM (2007) ISBN 978-1-59593-619-6
27. Bencina, R., Kaltenbrunner, M., Jorda, S.: Improved Topological Fiducial Tracking in the reacTIVision System, San Diego, CA, USA. IEEE Computer Society, Washington, DC (2005) ISBN:0-7695-2372-2-3
28. Bencina, R., Kaltenbrunner, M.: The Design and Evolution of Fiducials for the reacTIVision System. In: 3rd International Conference on Generative Systems in the Electronic Arts, Melbourne, Australia (2005)
29. Reas, C., Fry, B.: Processing: A programming handbook for visual designers and artists. The MIT Press Cambidge, Massachusetts (2007) ISBN:978-0-262-18262-1
30. IOhannes m zmolnig.: Reflections in Pure Data. In: Linux Audio Conference, Parma, Italy (2009)
31. Ericsson, K.A., Simon, H.A.: Protocol analysis. MIT Press (1985)

Neural Interface Emotiv EPOC and Arduino: Brain-Computer Interaction in a Proof of Concept

Eduardo Emilio Reder, Amilton Rodrigo de Quadros Martins,
Vinícius Renato Thomé Ferreira, and Fahad Kalil

IMED – Faculdade Meridional, Senador Pinheiro. 304,
99070-220 Passo Fundo, Brazil
eduardo.reder@gmail.com,
{amilton,vinicius,fahad.kalil}@imed.edu.br

Abstract. This study aims to demonstrate the interaction between the human being and the machine through a neural pattern recognizing interface, namely Emotiv EPOC, and a robotic device made by Arduino. The union of these technologies is assessed in specific tests, seeking a usable and stable binding with the smallest possible rate of error, based on a study of how the human electrical synapses are produced and captured by the electroencephalogram device, through examples of projects that achieved success using these technologies. In this study, the whole configuration of the software used to bind these technologies, as well as how they work, is explained, and the result of the experiments through an analysis of the tests performed is addressed. The difference in the results between genders and the influence of user feedback, as well as the accuracy of the technologies, are explained during the analysis of the data captured.

Keywords: Emotiv EPOC. Arduino. Brain-Computer Interface. Interaction. Electroencephalogram.

1 Introduction

According to the French philosopher Pierre Lévy, social and economic changes of the 21st century came to establish a new paradigm of social life in the contemporary society. Such abrupt change of paradigm is what sociologists of current technology come to call “The Technological Revolution”. In his book “O que é virtual?” (Becoming Virtual: Reality in the Digital Age) (1996), Lévy sets and demonstrates how the virtual abstraction of information has an impact both on everyday life and on more complex business processes, science and even contemporary human thought.

As well pointed out by Levy, the development of the very scientific method is already conditioned to the use of digital technology and information technology. Medicine and neuroscience, through digital image modeling for diagnostics, establish a pattern of work, making use of resources which, for simple clinical psychology, seemed unreachable.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 612–623, 2014.

It is with this intention that Emotiv EPOC, a helmet which consists of a complex portable noninvasive electroencephalogram apparatus produced by Emotiv, is presented in this study.

In addition to the Emotiv EPOC, other technologies, such as robotics (e.g. through Arduino), are easily falling into the hands of ordinary people. One can buy the Arduino Starter Kit for an affordable price and this is a very easy platform to work on.

The Emotiv EPOC, as well as his brother, the Electroencephalogram (EGG), use a set of fourteen sensors and two references to tune the electrical signals produced by the brain in order to detect patterns of thought, feelings and expressions in real time (EMOTIV, 2013), and based on the assumption that the device can receive such information, it is possible to use them for any purpose; from playing a specific game for this interface or controlling common actions in any application on your computer without using a mouse or a keyboard, to performing physical actions through the use of Robotics in conjunction with the EPOC, everything is defined by how the device is used and how the information captured will be treated and used.

An idea to make use of these data is, precisely, its combination with robotic interfaces, such as the Arduino, to create prototypes and test the concept of functionality, thus expanding the horizon of utilization for larger applications and having a greater influence on the scientific and commercial environment.

According to McRoberts (2011, p. 22), the Arduino is a micro single-board controller and a set of software to program it. In practical terms, it is a small computer that you can program to process inputs and outputs from the unit and the external components connected to it.

Based on the assumption of using the EPOC as the data input device, and the Arduino as the output device, this study intends to unite the two technologies and develop an interface (software), with the purpose of demonstrating the functional visibility of neural reading for robotic control.

Through this project, the idea that the link between the EPOC and the Arduino is really possible will be proven, thus generating new possibilities far greater than four simple lamps lighting up controlled by the power of the brain.

2 Electrical synapses and the EEG study

The nervous system "is the basis of our capacity for perception, adaptation and interaction with the world in which we live" (GAZZANIGA, 2000 Apud STERNBERG, 2008, p. 43) and "through this system we receive, process and then respond to the information that come from the environment" (RUGG, 1997 Apud STERNBERG, 2008, p. 43).

Through devices such as the EEG and study areas such as the neuroscience an important objective of the work and study of the brain can be attained, locating the areas responsible for a given function or behavior (STERNBERG, 2008, p. 42).

Everything that happens in the brain and, consequently, in other parts of the body occurs due to the synapses. Lent (2005, p. 99) refers to the synapse as "a biological chip, because the computations that the neural circuits are capable happen in them“.

Among all the 1,000 synapses formed and the 10,000 received by a neuron, the majority consists of chemical synapses, conducted through the neurotransmitters. The rest occurs by electrical impulses through the presynaptic membrane and through the channels that bind presynaptic and postsynaptic cells (KANDEL & SIEGELBAUM, 1995 Apud MONTENEGRO, 2012, p. 01), and these impulses are called electrical synapses, responsible for the transmission of information between neurons, which are picked up by EEG devices.

The whole process of capturing the electrical signals from the brain by the EEG device depends on the existing electric current of the brain, but this in itself has no power. For the electrical current to flow a conductive bridge connected to the brain is necessary – considered, in this case, as the source of energy – thus creating a circuit. This “brain circuit” is formed as any other electric circuit, and the science of electricity works the same way.

Simply put, the electroencephalographic record consists of capturing the cerebral electrical using electrodes, transmitting to the electrode box and then to the amplifiers of the EEG device. Then, the record is made. The breadth and width of each wave are based on the voltage and the frequency of the electrical current captured by the electrodes.

According to Montenegro (2012, p. 8), the electrodes are the metallic medium where the signals will be received. Directly applied to the patient's scalp according to the international 10-20 system, which consists of an internationally recognized method to describe and apply the location of electrodes on the scalp, the conductors register the electrical currents that will be forwarded to the amplifiers. The current passes from scalp to the electrode through the current created by the ions present in the solution, which is a conductive gel, a paste or even a liquid. An electrochemical phenomenon occurs on the link between the scalp and the electrode, turning the ionic current in an electron flow, which generates an electrical current capable of being transmitted to the EEG amplifiers which increase 1500 times the voltage picked up in order to be able to record the electrical activity of the brain.

The technique of electrodes on the scalp is widely used, since it is not intrusive, however, it has many limitations, mainly by suffering influence of the skull itself.

3 Studied Technologies and Its Limitations

The interaction of human thought and a given software or hardware goes into test when tools that allow it to happen reach the market. The studied union is made possible through the interaction between the Emotiv EPOC neural interface and the Arduino.

The Emotiv EPOC, despite being relatively new in the market, makes use of EEG technology to perform actions on digital media through a headset, but it's not a novelty that brain patterns are used to control digital actions in the field of robotics.

Thus, some projects using portable EEG headsets are emerging. Improving accessibility has been a much aimed target in current studies in different areas, which leads most of these studies to be directed to the field of disabled people in order to facilitate their lives.

For example, the University of Zaragoza, Spain, has been studying, in recent years, the possibility of creating a wheelchair commanded by thought captured by an EEG machine.

Through sensors installed in the chair, a scenario is assembled by a software and options are given to the user, who makes the choice of where to go only by thinking. After the path is designed, the user can rest and the chair calmly goes through the path created earlier, decreasing the mental exhaustion present in interactions that require concentration to achieve the desired action.

Although the Spanish project is revolutionary, it was not the only one to be created in order to improve the lives of wheelchair users. Shortly after the creation of the Spanish project, a team of researchers at the Federal Polytechnic School of Lousana (EPFL), Switzerland, also created a project for wheelchair users driven by electrical signals detected by EEG, also adding augmented reality to the set of technologies involved.

In this project, two cameras are positioned in front of the chair to detect and recognize close objects on real-time, avoiding, in a relatively simply fashion, possible collisions and accidents that may occur involving the chair and the rest of the environment in which it lies. Using software developed for this project, the user controls the wheelchair through defined movements, for example, the simulation of movements of the right hand to move the chair to the right and of the left hand to move it to the left.

Besides the two aforementioned projects, the Emotiv EPOC is being used at the University Center of Pará (CESUPA), in Brazil, to control a simple wheelchair, built on Arduino. Being simpler than those cited above, this project makes use of a regular wheel chair, fitted with motors and a laptop. The signals are sent to the computer wirelessly and it issues specific commands to the chair, making it to move.

Projects in this area do not cease to be created, which clearly demands further study of these technologies.

3.1 Emotiv EPOC Neuroheadset

According to the institutional site of Emotiv (2013), the neuroheadset EPOC is a personal interface for human interaction with the computer through the acquisition of electrical signals produced by the brain, through techniques of electroencephalography (EEG), in order to identify thoughts, feelings and expressions in real time.

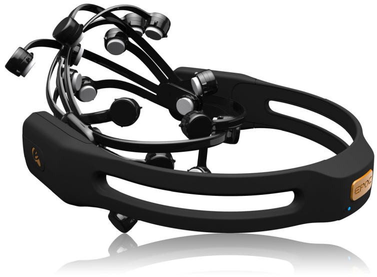


Fig. 1. Emotiv EPOC Neuroheadset

Through 14 sensors and 2 references that offer optimal positioning for accurate spatial resolution, brain patterns are obtained for certain functions, in addition to reactions such as eye blinking and smile, and movements such as lifting or lowering the head and turn it to the right or left.

Along with the neuroheadset, accompanies an interface for training where 13 actions can be trained and patterns of each user for these actions are recorded so that when played, are identified correctly. These actions are: move right, move left, up, down, push, pull, rotate clockwise, counterclockwise, rotate left, right, front and back.

3.2 The Choice of Arduino as Platform for Application of Concept

According to McRoberts (2011, p. 22), "Arduino is a small computer that you can program to process inputs and outputs between the device and external components connected to it".

The Arduino has numerous components that can be connected to it as LEDs, dot matrix displays, buttons, motors, switches, temperature sensors, pressure, distance, GPS receivers, Ethernet modules or any other device that transmits data (MCROBERTS, 2011, p. 22), and, depending on the Arduino, USB ports that enable connection to PC or Mac to exchange information.

A free development platform allows software, or sketches, to be created in a C-based language, which Arduino understands, and as these are open source as well as the hardware, have full compatibility with each other. These sketches are loaded into the Arduino board allowing it to interact with the components connected to it and to the environment.

The ease of creating and by its recognition for exercising its role, the Arduino was considered one of the best choices for use in proving the concept of usability of Emotiv EPOC connected to an independent hardware platform.

4 Project, Development and Research Application

The interaction between these technologies is presented in this project. The idea of uniting the brain activities of a subject to a robotics interface is the main focus of this scientific work.

The information used in this paper comes through observation of LEDs on an Arduino board. This board is connected to the computer via USB cable and receives specific data sent to it by software that was developed on Windows platform, which in turn receives data from EmoKey software and the Control Panel EPOC, as can be seen in the representation follow.

Solely the union of the technologies presented in this project does not prove its usability. It can run as expected and technologies communicate through the software mentioned, as shown in Figure 2, however, if the error rate and / or difficulty of controlling these technologies are very large, the way which is being made this combination is no longer useful.

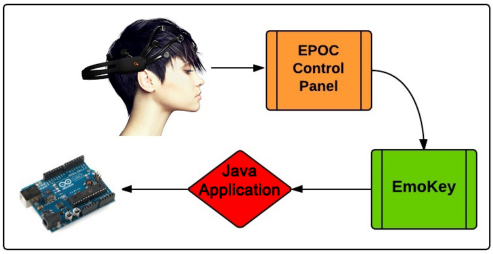


Fig. 2. Representation of the interaction between technologies

Given this thought a series of tests for a total of 10 people is proposed, which are divided between males and females in equal percentage, in order to assess the usability and the percentage of hits and misses on a range of small assisted tests.

The first and most important step of the whole process of testing is the calibration and training of movements and thoughts in Emotiv EPOC Control Panel. In this step the user wears the EPOC, each sensor is properly positioned at the correct point of the head based on the International 10-20 system and the signal of each sensor is checked through the existing representation on the display.

After all sensors are sending signal, ranging medium to good, the next step is the training of thoughts or movements in "Cognitiv Suite" tab of the Control Panel where the five thought patterns are recorded individually, "Neutral" , "Left" , "Right", "Lift" and "Down". Each of these can be trained as often as necessary according the ability of each user to focus.

The first pattern to be recorded is obviously the "Neutral" because this is the basis of comparison for all other movements that are trained later. For this pattern the user is oriented to relax and try to keep the mind free of thoughts, or simply do not focus on any particular thought, so that electrical signals are captured from a peaceful mind without defined patterns of thought. For 8 seconds this pattern is recorded, process that can be repeated until the user considered that was, in fact, relaxed and was not focused on anything.

The order of recording after the "Neutral" standard be recorded is irrelevant, because the training is done individually for each motion, and the order is not important. The only point that one should take care on this step is that the user is calm and aware of what he/she is doing. They should know that their goal is to focus their thinking on the proposed motion and taking care not to record the same patterns in different movements, so they are oriented thinking is quite distinct from one another and that during training the user has the greatest focus on activity.

After The patterns are all recorded the next step is the configuration of the second software used, EmoKey, whose function is to interpret the signal recorded by the Control Panel and play a pre-programmed action.

The EmoKey works from rules and for each of these is identified and enabled each one receives one or more Trigger Conditions which when executed triggers the rule.

For execution of the proposed battery of tests, four (4) specific, simple rules are created, each having a unique action and well-defined conditions trigger. The target

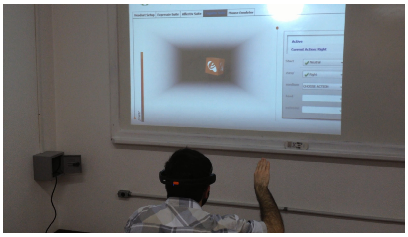


Fig. 3. User being training in EPOC Control Panel

application of all rules is in focus on the screen and is the third software that should be working to achieve these tests, as shown in Figure 2, the Java application.

The rule for "Right" receives a key that consists of writing the number one (1) and simulate an "Enter" on the keyboard. The condition for this rule to be performed is that thought "Right" is played and is consistent.

Each of the other three rules, "Left", "Above" and "Below", follows the same pattern of the "Right" rule, only changing its action for the respective thought and the value that is written and sent to the Arduino.

The third software used, the Java application was developed with the Swing graphics library used for visual programming. Through this library a simple window with a field for entering text and two buttons "OK" and "Cancel" is created, precisely to receive the value passed by EmoKey.

The rule in EmoKey is set to write the command for the thought reproduced by the user and simulate a click on the "Enter" key on the keyboard, i.e. confirm the submission of this information. When confirmed, the value is captured by the Java application and sent to the Arduino which should light up the corresponding LED.

The sketch written in Arduino is used to interpret the signal sent by the Java application through the use of the RXTX API, which connects via serial port the computer and Arduino. In this sketch are configured, very simply, the rules to light up each LED when it receives the expected value, keeping it lit for one (1) second and then deletes them.

The Java application receives the value written by EmoKey previously configured, and sends it to the Arduino, which turns on a light for each value (Figure 4) as follows:

- Thought "Right": send the value 1 and turns on the green LED;
- Thought "Left": send the value 2 and turns on the red LED;
- Thought "Up": send the value 3 and turns on the white LED;
- Thought "Down": sends the value 4 and turns on the yellow LED.

The Java application is terminated when the user cancels, using the existing button in the window, the application without sending any data to the Arduino.

After all patterns are trained in the Control Panel, the Java application is properly working, the sketch is recording in Arduino and the rules are set in EmoKey, testing can begin effectively. Each user goes through four (4) test sessions, on the first session only a standard is tested and on the fourth session all standards and rules are in test.

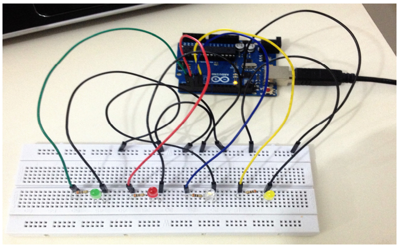


Fig. 4. LEDs on Arduino

Individually, each session is divided into two parts. The first consists of fifteen (15) replicates composed of green arrows pointing in a certain position among the four existing. The user should reproduce the thought relative to the direction shown by the arrow at each repetition and the results are captured by Arduino are registered in a specific table. This part of the session can be seen in Figure 5, where the user has no access to the Arduino.

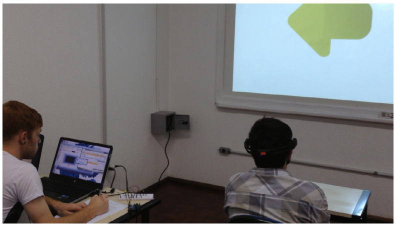


Fig. 5. Test without user feedback and results being captured

The second part is nothing more than a repetition of the first part, that is, the user is challenged to play a specific oriented thinking shown by the arrows, but the Arduino is in field of vision allowing the user to monitor in real time if their thoughts are executing the proposed action. With that he/she can focus more attention on the proposed standard aiming at turning on the LED.

Each repetition takes six (6) seconds to happen, and during half the time (3 seconds) the arrow appears and during the other half of the time the screen turns white, giving the user time to relax and clear his mind to run the next thought when the next arrow is displayed.

After all testing sessions were conducted and all data is properly captured the users were able to describe how they felt during the testings and how they evaluate it.

Records containing tables of hits, wrongs and feedbacks of users were analyzed, and some results were obtained from them in order to improve the view on the technologies presented.

5 Data Analysis

The analysis of the data captured during the tests intended to check the accuracy of the union of the technologies presented, and it was confirmed that the stimulus produced by the user's brain reaches the Arduino and turns on the LED as proposed.

To analyze this accuracy, ten (10) people aged between twenty (20) to twenty nine (29) years, invited by the author were used in the tests previously discussed. Without following any particular order, five (5) users of male gender and sex, four (4) users of female sex and gender and one (1) user of male sex and female gender were tested individually.

One of the issues discussed was to seek to know the relevance of the user to be aware whether or not he was lighting the LED as proposed. Each session had two tests, one without the user seeing the Arduino board and the other user could check in real time if the LED which should light up or was not responding. It was expected that when receiving feedback via the Arduino, the user could increase his/her level of hits, assuming that one could focus more thought in the proposed action, but as we can see in the table below the difference between the first test and the second one not was very large, and we can assume that the way the feedback was passed was not as efficient as expected.

Table 1. Difference of feedback to the user

Result	Without feedback	%	With feedback	%	Total	%
Hit	302	51%	314	52%	616	51%
Wrong	50	8%	46	8%	96	8%
Opposite	61	10%	66	11%	127	11%
Without answer	187	31%	174	29%	361	30%
Total	600	100%	600	100%	1200	100%

The entire analysis was based on four (4) possibilities to be able to get the most accurate results. Each stimulus passed to users could result in:

- Hit: when the respective LED lit according to the passed stimulus;
- Wrong: when the result had no relationship with the stimulus, for example, we obtained the result “up” with the “left” stimulus;
- Opposite: when the result is the opposite of stimulus, for example, we obtained the result “left” with “right” stimulus;
- Without Answer: When no LED lit during the stimulus.

As proposed, and judging from the way of selecting users to be tested, a relationship between males and females users was created. From the users it was possible to see a greater focus and attention of male users when compared to females. As can be seen in the table below, there is no significant difference in the results regarding the gender of users.

During the tests, and even through the opinions of the users tested, there is a clear relationship regarding the difficulty when a new thought pattern was added. The first test was with only one thought pattern and at each new test a pattern was added.

Table 2. Differentiation between Genres

Result	Male	%	Female	%	Total	%
Hit	318	53%	298	50%	616	51%
Wrong	50	8%	46	8%	96	8%
Opposite	58	10%	69	11%	127	11%
Without answer	174	29%	187	31%	361	30%
Total	600	100%	600	100%	1200	100%

At each test performed hits decreases and the amount of different than correct results increases, as it requires greater concentration of the user and thoughts that were not well recorded end up being forgotten and reproduction of these is almost impossible.

After all tests and some relevant information have been drawn from these data, there is still a relationship that becomes necessary. The overall ratio of successes, failures, opposite results and lack of physical response through the Arduino not caring about which test resulted that information. The table below shows these results.

Table 3. Relationship between Test Analysis and General Analysis

	Test 1	%	Test 2	%	Test 3	%	Test 4	%	Total	%
Hit	244	81%	182	61%	116	39%	74	25%	616	51%
Wrong	56	19%	51	17%	105	35%	149	49%	361	30%
Opposite			67	22%	27	9%	33	11%	127	11%
Without answer					52	17%	44	15%	96	8%
Total	300	100%	300	100%	300	100%	300	100%	1200	100%

It is noteworthy that the second larger part of the data, which represent the tests “Without Answer” does not indicate that the user was not able to perform the movement correctly, but the thought pattern was not concise enough to trigger the rule set in EmoKey to light the LED on the Arduino, because during the testing sessions it was observed in the EPOC Control Panel where the cube used for training suffered an effect according the user's thinking.

Based on this information, it is worth restating that the percentage of correct responses and accuracy of movements may prove to be augmented with a greater amount of training and with individual settings for each user at EmoKey in the same way it was done in the Control Panel during the training of thought standards.

6 Final Remarks

This work demonstrates and reaffirms that the technology of reading patterns of neural waves is existing and can be viable and affordable. The devices used in this study are publicly available outside the academic community. Once object of scientific projects feasible only for medical systems or high-precision and criticality, technology presented here is available to commercial developer and student, so it's only a matter of time and market so they can reach the software industry and information technology solutions.

The tests performed and the results achieved in this work are from people without any disabilities, which may have uninterested them, making it difficult to perform the tests. Bringing these technologies for people with special needs, who see this as a change in their lives, totally change the paradigm and likely achievements.

After all tests, even requiring more precise technologies, it has been proven that there is no randomness in the project and that at least one pattern of thought possessed great hit percentage, which indicates a valid union and that with well-defined training and development of higher level software and hardware the paradigm we live in today may be changed.

The LED lamps the Arduino board turning on as each thought pattern recorded by EPOC and projects already undertaken, such as wheelchair controlled by thought and the tractor moved by brain signals, show that simple designs can evolve and become major innovations, as we see in studies in the area of augmented reality, or even the Google Glass project, for example, that have evolved in recent years and may even be used in sets with techniques using brain-machine interfaces.

The results of data analysis indicate that a lot of training and skill to achieve a positive result is required. This suggests that there is need for research in psychology and sensory medicine that will contribute to the development of technology.

From a systems analysis approach, the default robotic control to light small LED lamps may not mean a great advance itself beyond the technology and the physical principles applied in electronic devices offer. However, the integration of these two technologies is what we hope to be inserted at the forefront of technological trends in contemporary society. The virtual world and human communication, through the branch of affective computing hitherto presupposed communication interfaces with visual and tactile-motor sensory interactions. The paradigm of capturing user intentions directly from its source, for us, includes a bright future for the development of information technology.

References

1. McRoberts, M.: Arduino Básico. Novatec Editora LTDA, São Paulo (2011)
2. Lent, R.: Cem bilhões de neurônios: Conceitos fundamentais de neurociência. Editora Atheneu, São Paulo (2005)
3. Montenegro, M.A., Guerreiro, M.M., Cendes, F., Guerreiro, C.A.M.: EEG na prática clínica. Revinter, Rio de Janeiro (2012)
4. Sternberg, R.J.: Psicologia Cognitiva. Artmed, Porto Alegre (2008)
5. Nascimento, F.S., Santos, F.C.: Controle de uma cadeira de rodas servo motorizada a partir de Brain Computer Interface não invasivo. Centro Universitário do Pará – CESUPA, Belém (2011)
6. Levy, P. O.: que é virtual?. Editora 34, São Paulo (1996)

7. Gomez-Gil, J., San-Jose-Gonzalez, I., Nicolas-Alonso, L.F., Alonso-Garcia, S.: Steering a Tractor by Means of an EMG-Based Human-Machine Interface. Department of Signal Theory, Communications and Telematics Engineering. University of Valladolid, Valladolid (2011)
8. Emotiv. EPOC Feactures, `http://www.emotiv.com/epoc/` (May 14, 2013)
9. Emotiv. EEG Feactures, `http://www.emotiv.com/eeg/` (May 14, 2013)
10. Meio Bit. Cadeira de Rodas High-Tech, `http://meiobit.com/73247/cadeira-de-rodas-high-tech/` (June 5, 2013)
11. Engadget. Mind-controlled wheelchair prototype is truly, insane awesome, `http://www.engadget.com/2009/05/04/mind-controlled-wheelchair-prototype-is-truly-insanely-awesome/` (June 5, 2013)
12. EPFL. Brain-Machine Interface @EPFL: Wheelchair, `http://www.youtube.com/watch?v=0-1sdtnuqcE` (June 5, 2013)
13. BCIZARAGOZA. Brain-Controlled Wheelchair, `http://www.youtube.com/watch?v=77KsE` (June 5, 2013)
14. TVSERPRO. Cadeira de rodas acionada com a força do pensamento, `http://www.youtube.com/watch?v=XSt4YIMpE4g` (June 5, 2013)

A Heuristic Model of Vibrotactile Haptic Feedbacks Elicitation Based on Empirical Review

Anak Agung Gede Dharma[1] and Kiyoshi Tomimatsu[2]

[1] Kyushu University, Graduate School of Design, Fukuoka, Japan
[2] Kyushu University, Faculty of Design, Fukuoka, Japan
dharma.satya.utama@gmail.com,
tomimatu@design.kyushu-u.ac.jp

Abstract. We propose a novel heuristic model of vibrotactile feedbacks elicitation. The model is based on two known tactile elicitation principles, i.e. perceived tactile sensation and apparent haptic motion. Our previous studies, along with empirical reviews were used to provide an insight of how these two principles work individually. Our preceding works on the mapping of texture phase diagram of artificial vibrotactile stimuli reveals 3 main perceived vibrotactile sensation, i.e. dampness, friction, and hardness. Furthermore, we have conducted a preliminary research to observe apparent haptic motion in our proposed haptic vest interface. Our findings and the empirical reviews imply that these two haptic principles can be used concurrently to create a novel user experience.

Keywords: vibrotactile haptic feedback, heuristic model, tactile perception.

1 Introduction

In recent years, haptic feedback has been intensively researched as one of the important elements of multimodal human computer interaction. Haptic feedback plays indispensable roles in some user scenarios due to its unique characteristics. They include the capability of delivering information in a non-intrusive way while simultaneously exciting the cutaneous sense with rich tactile sensations. Hence, it can be effectively used to shift user's attention from the periphery to the center of attention and vice versa. Furthermore, haptic feedback can be applied for numerous purposes such as wayfinding [1], tactile mapping in augmented reality [2], or therapy [3].

Vibrotactile stimulus is one of the types of haptic stimuli, which utilizes various kinds of waveform to convey haptic stimuli to cutaneous sense [4]. The applications of vibrotactile stimuli varies from the vibration alarm in cellular phone, game controller, touchscreen feedback, etc.

In this paper, we propose a heuristic model of vibrotactile feedbacks elicitation. Two known principles in haptic feedback, i.e. perceived haptic sensation and apparent haptic motion are discussed in section 2. In addition to empirical reviews in section 2, we have conducted researches on perceived haptic sensation (section 3) and apparent haptic motion (section 4). Furthermore, our proposed model is discussed thoroughly in section 5. Section 6 discusses about conclusion and future works.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 624–632, 2014.

2 Related Works

2.1 Perceived Haptic Sensations

The mechanoreceptors beneath the skin cause a psychophysiological haptic effect, i.e. the perception of different kinds of vibrotactile sensations such as roughness, hardness, and friction [5].

These effects are caused by different amplitude thresholds of three main mechanoreceptors (i.e. SA I, FA I, and FA II). We have proposed several models to explain three major sensations that are elicited by vibrotactile stimuli [5, 6]. In our proposed model [5], we have developed vibrotactile waveforms by the superposition of three different kinds of frequency range. Low frequency stimuli were set at 0.4 – 7 Hz, medium frequency stimuli at 25-40 Hz, and high frequency stimuli at 200 – 250 Hz.

Our findings are similar to other studies about psychological mapping of perceived haptic sensations on authentic materials [7, 8]. Roughness, hardness, and friction can be identified as three main factors to describe perceived vibrotactile haptic sensations.

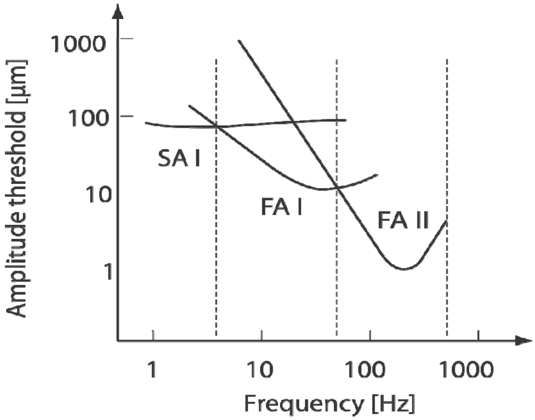


Fig. 1. Detection thresholds of vibratory stimuli (based on Bolanowski et al. [9])

2.2 Apparent Haptic Motion

Another aspect that needs to be taken into account when designing vibrotactile stimuli is apparent haptic motion. Apparent haptic motion is a haptic motion illusion that occurs when a series of haptic stimuli move within specific Stimuli Onset Asynchrony (SOA) [10].

Preceding works by Vaucelle et al. [2], Ertan et al. [11], and Tan et al. [12] suggest that an n x n array of actuators can be used to convey haptic information to the users, by utilizing this apparent motion. In our previous study, we have proposed a haptic vest that utilizes 5x12 arrays of vibrotactile actuators [13]. This haptic vest is designed to be worn on the torso. When a user wears the vest, it can elicit various apparent haptic motion as described by Israr et. al. Furthermore, we have observed that users react differently according to the type of haptic pattern. Their behavior can at least be measured by 3 variables, i.e. reaction time, apprehensibility rating, and comfort rating.

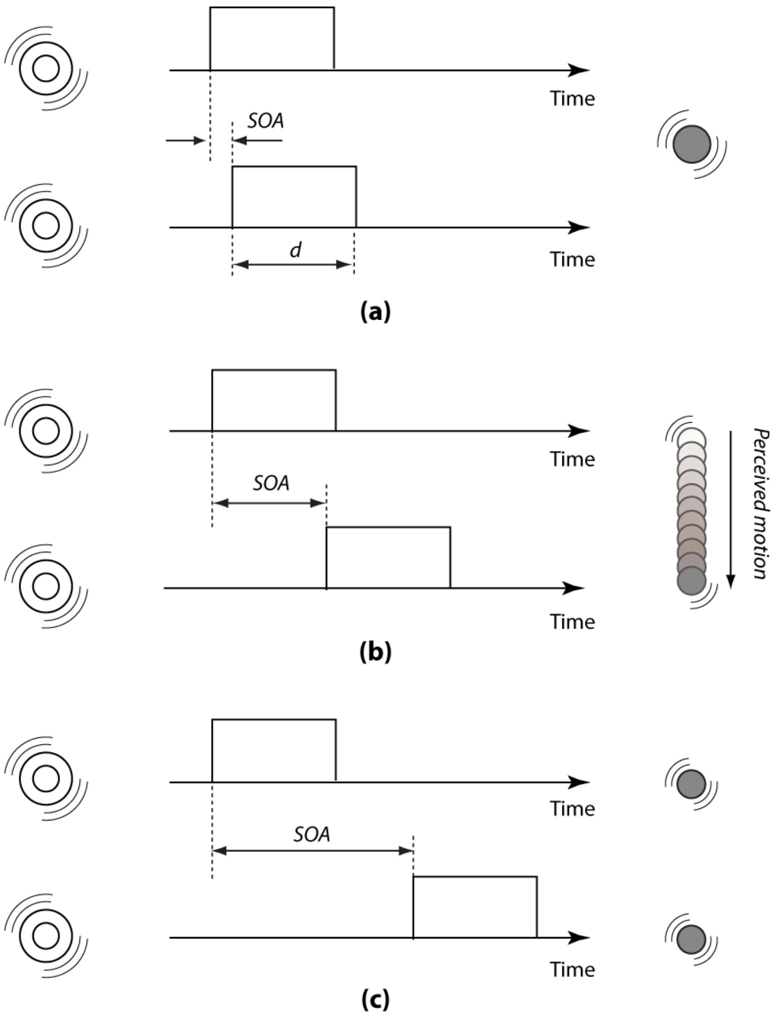
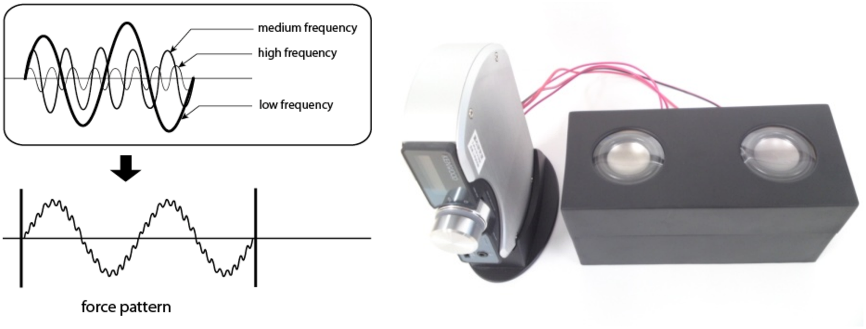


Fig. 2. Apparent tactile motion as described by Israr et al. [11]; (a) simultaneous stimulation, (b) apparent tactile motion, (c) successive stimulation

3 Preceding Works on Perceived Haptic Sensation

3.1 Research Method

Stimuli design and playback. Our vibrotactile stimuli design concept and the correlation between its variables are illustrated in Figure 3-a. Each stimulus was designed



(a) (b)

Fig. 3. (a) Three different types of frequencies to selectively stimulate SA I (low frequency), FA I (medium frequency), and FA II (high frequency); (b) A prototype to display vibrotactile stimuli that consist of a pair of vibrotactile actuators and digital amplifier

by the superposition of three haptic vibrations of different frequency ranges, i.e. the constructive interference of three different frequency ranges.

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 3-b).

Table 1. Amplitude and frequency variables for a given force pattern

Amplitude Variables	**Receptor Target**	**Amplitude Range (micron)**	**Frequency Variables**	**Receptor Target**	**Frequency Range (Hz)**
Amplitude_FA1	FA1 (Meissner)	0 – 450	Frequency _FA1	FA1 (Meissner)	25– 40
Amplitude_FA2	FA2 (Pacinian)	0 – 120	Frequency _FA2	FA2 (Pacinian)	200 – 250
Amplitude_SA1	SA1 (Merkel)	0 – 600	Frequency _SA1	SA1 (Merkel)	0.4– 7

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 3-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 [5].

3.2 Results and Discussion

We have developed a new classification method of tactile sensations to explain perceived artificial vibrotactile sensation. A texture phase diagram has been developed that can be used to explain the correlation between artificial vibrotactile stimuli and tactile perception. This experiment extracted 3 principal components of vibrotactile stimuli, i.e. dampness, friction, and hardness that account for 56.52% of overall tactile perception. This study reveals similar results with Hayakawa et al. (friction, hardness, and moisture) [8] and Hollins et al. who propose softness-hardness and roughness-smoothness as two of the most important element in tactile sensation.

Furthermore, the explained cumulative variance is relatively low (56.52%). Hence, 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.

In addition to this experiment, another experiment to measure the correlation between tactile sensations generated by physical materials and their physical properties has also been discussed [6]. We have found that roughness tactile sensation strongly correlates to surface geometrical roughness (μm), softness to compliance (μm/gf), and stickiness to coefficient of friction (μ).

4 Preceding Works on Apparent Haptic Motion

4.1 Research Method

Participants. Sixteen subjects (7 males and 9 females) participated in the experiment. All subjects are undergraduate students from School of Design, Kyushu University.

Experiment Setting. Our experiment subject wears the haptic vest as shown in Figure 4-b. The experiment sequence is as follows: (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; (3) the stimulus is being played in a continuous loop until the subject gives a response; (4) the subject was asked to choose between four directions (back, front, left, and right) based on his/her perceived direction of the haptic stimulus; (5) The subject were asked to rate the stimulus' comfort and apprehensibility on a five step Likert scale. These sequences are repeated until the last stimulus has been selected.

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 [13].

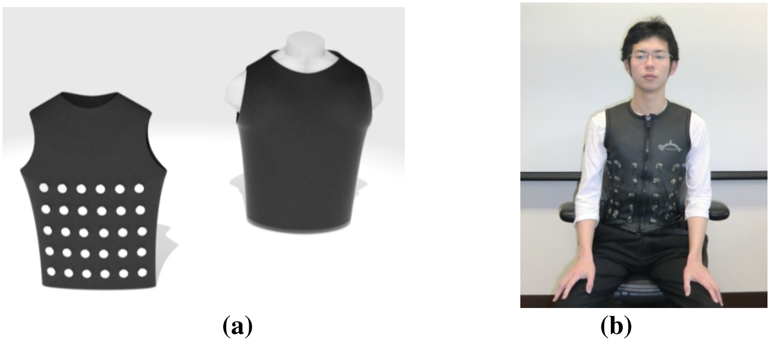


Fig. 4. (a) The design concept of our proposed wearable haptic vest, the left figure and right figure show its outer and inner view, respectively; (b) Experiment setting

4.2 Results and Discussion

We have developed the haptic vest as an attempt to observe to effect of apparent haptic motion. Furthermore, according to user testing, basic characteristics of perceived haptic stimulation to the users, such as comfort and apprehensibility, have been explored.

In terms of comfort and apprehensibility, 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.

In addition to apprehensibility and comfort ratings, users also reported apparent haptic motions during the experiment. Users reported apparent haptic motions for some patterns that have transitions every 0.5 seconds. However, we did not measure the effect of SOA and minimum space that is required to stimulate apparent haptic motion.

5 Discussion: Combined Effects of Perceived Tactile Sensation and Apparent Tactile Motion

In this paper, we propose two indispensable factors for vibrotactile haptic feedbacks design, i.e. perceived haptic sensations and apparent haptic motion. Perceived haptic sensations are cognitively evaluated by users that can be described by adjectives. They have been identified in our preceding researches [5, 6] and Hayakawa et al. [9]. These sensations are elicited by direct contact between the actuators and the skin. We have confirmed that stimulations to three different types of mechanoreceptor (SA I, FA I, and FA II) can affect perceived haptic sensations. These stimulations are caused by different ranges of mechanoreceptor frequency thresholds, as described by Bolanowski et al. [9].

Spatial and temporal settings of vibrotactile actuators affect apparent haptic motions that are perceived by users. Apparent haptic motion is elicited when vibrotactile stimuli moves spatially on the surface of the skin [10] or the actuators are positioned in parallel on the back and front (i.e., to generate thrusting apparent motions) [14]. Apparent haptic motion requires two or more haptic stimuli to vibrate within a specific SOA and within a specific distance.

The key characteristics, psychophysiological factors, and elicited sensations have been identified and concluded in our proposed model (Figure 5). Those characteristics and their corresponding sensations need to be taken into account for effectively designing vibrotactile haptic feedbacks in various user scenarios to create immersive user experience.

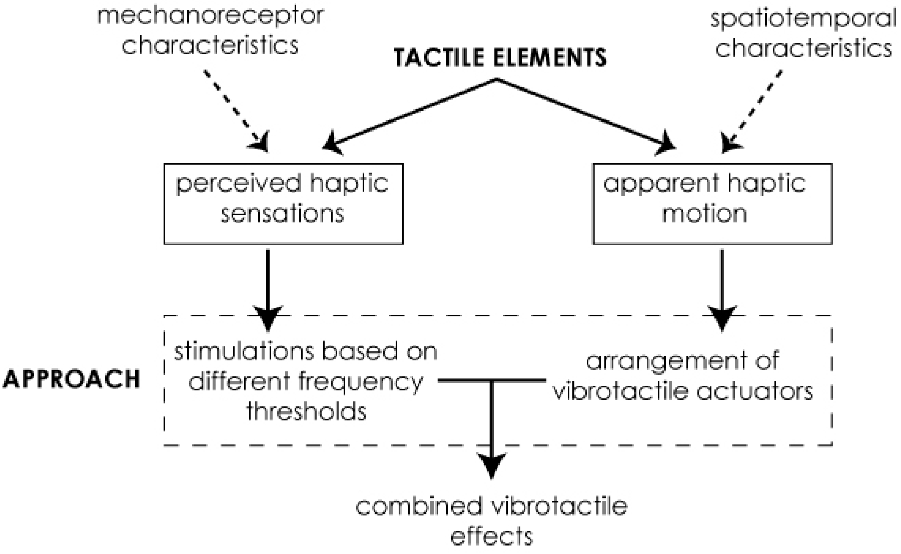

mechanoreceptor characteristics
TACTILE ELEMENTS
spatiotemporal characteristics
perceived haptic sensations
apparent haptic motion
APPROACH
stimulations based on different frequency thresholds
arrangement of vibrotactile actuators
combined vibrotactile effects

Fig. 5. Our proposed concept of the elicitation of vibrotactile effects by the combination of perceived haptic sensation and apparent haptic motion

One of prominent examples of applying these vibrotactile principles has been described in Surround Haptics by Disney Research [16]. Disney Research developed a two-dimensional haptics display mounted on the back of a chair. They fully utilize apparent haptic motion, in addition to integrating stereoscopic visuals and surround sound to simulate an immersive user experience. However, adding perceived haptic sensations into the system would create rich tactile sensation for the users, for example: the possibility to simulate rich tactile sensations by alternating the level of roughness, softness, or friction.

Moreover, we have identified other unknown factors that still need to be explored. As pointed by our proposed model in the perceived tactile sensation experiment, only 56.52% of vibrotactile stimulations could explain overall tactile sensation. Furthermore, in our haptic vest study, users reported that transitions of different haptic patterns affect their way of perceiving directions, which imply that apparent haptic motion have a role in determining user experience.

6 Conclusion and Future Works

This study provides additional insights on user interactions with vibrotactile haptic feedbacks. Known physiological traits of tactile elicitations have been summarized. By combining both of perceived haptic sensations and apparent haptic motion, we suggest that rich tactile effects can be created.

As the future works, we intend to explore more vibrotactile haptic parameters to discover other unknown factors in the interaction with vibrotactile stimuli. More in-depth usability tests will be performed to discover those factors and to test the validity of the proposed concept.

References

1. Heuten, X., Henze, N., Boll, S., Pielot, M.: Tactile wayfinder: a non-visual support system for wayfinding. In: Proc. 5th Nordic Conference on Human-Computer Interaction: Building Bridges, Lund, Sweden, pp. 172–181 (2008)
2. Bau, O., Poupyrev, I.: REVEL: Tactile feedback technology for augmented reality. ACM Transactions on Graphics (TOG) 31(4), 89 (2012)
3. 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, pp. 467–470 (2009)
4. Okamura, A.M., Dennerlein, J.T., Howe, R.D.: Vibration feedback models for virtual environments. In: Proc. IEEE International Conference on Robotics and Automation, Leuven, Belgium, pp. 674–679 (1998)
5. Dharma, A.A.G., Tomimatsu, K.: Mapping texture phase diagram of artificial haptic stimuli generated by vibrotactile actuators. In: Kurosu, M. (ed.) Human-Computer Interaction, Part IV, HCII 2013. LNCS, vol. 8007, pp. 578–586. Springer, Heidelberg (2013)
6. Dharma, A.A.G., Matsumura, Y., Tomimatsu, K.: Design of a tangible prototype for displaying hapticons. International Journal of Asia Digital Art and Design 13, 5–12 (2010)
7. 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)
8. 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)
9. 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 (1998)
10. Israr, A., Poupyrev, I.: Control space of apparent haptic motion. In: Proc. IEEE World Haptics Conference 2011, Istanbul, Turkey, pp. 457–462 (2011)
11. 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 (1998)
12. Tan, H.Z., Gray, R., Young, J.J., Traylor, R.: A haptic display for attentional and directional cueing. Hapticse: The Electronic Journal of Haptic Research 1(3) (2003)

13. Dharma, A.A.G., Oami, T., Obata, Y., Yan, L., Tomimatsu, K.: Design of a wearable haptic vest as a supportive tool for navigation. In: Kurosu, M. (ed.) Human-Computer Interaction, Part IV, HCII 2013. LNCS, vol. 8007, pp. 568–577. Springer, Heidelberg (2013)
14. Ooshima, S., Fukuzawa, Y., Hashimoto, Y., Ando, H., Watanabe, J., Kajimoto, H.: Gut feelings when being cut and pierced. In: ACM SIGGRAPH 2008 New Tech Demo, p. 14 (2008)
15. Hollins, M., Faldowski, R., Rao, S., Young, F.: Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis. Perception & Psychophysics 54, 697–705 (1993)
16. Israr, A., Popyrev, I.: Exploring surround haptics displays. In: CHI 2010 Extended Abstracts on Human Factors in Computing Systems, Atlanta, USA, pp. 4171–4176 (2010)

Auditory Emoticons: Iterative Design and Acoustic Characteristics of Emotional Auditory Icons and Earcons

Jason Sterkenburg[1], Myounghoon Jeon[1], and Christopher Plummer[2]

[1] Cognitive & Learning Sciences,
[2] Visual and Performing Arts
Michigan Technological University
Houghton, MI, USA
{jtsterke,mjeon,cplummer}@mtu.edu

Abstract. In recent decades there has been an increased interest in sonification research. Two commonly used sonification techniques, auditory icons and earcons, have been the subject of a lot of study. However, despite this there has been relatively little research investigating the relationship between these sonification techniques and emotions and affect. Additionally, despite their popularity, auditory icons and earcons are often treated separately and are rarely compared directly in studies. The current paper shows iterative design procedures to create emotional auditory icons and earcons. The ultimate goal of the study is to compare auditory icons and earcons in their ability to represent emotional states. The results show that there are some strong user preferences both within sonification categories and between sonfication categories. The implications and extensions of this work are discussed.

Keywords: auditory icons, earcons, auditory emoticons, non-speech sounds, sonification.

1 Introduction

Since the first International Conference on Auditory Display (ICAD) in 1992, research on sonification, the use non-speech sounds [1], has proliferated. As one of the simplest sonification techniques, auditory icons [2] (representative part of sounds of objects) and earcons [3] (ear + icons, short musical motives as symbolic representations of objects) have been successfully applied to electronic devices as auditory feedback for user activity [e.g., 4, 5]. Following those precursors, spearcons [6] (compressed speech) and spindex [7] (speech + index) have also shown improved performance and reduced workload with menu navigation tasks in diverse contexts. Fairly recently, musicons [16] (music + earcons) and lyricons [17] (lyrics + earcons) have also been introduced to enhance aesthetic aspects as well as functional mappings of the non-speech sound cues. However, despite successful improvement in performance measures, relatively little research has focused on emotional or affective aspects of those auditory cues. If any, research treated with either auditory icons [8] or earcons [9] only, but few studies compared affective effects of both auditory cues in a

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 633–640, 2014.

single study [exception, 10]. The other research gap includes that affect research has depended merely on the simple valence dimension [positive – negative, e.g., 11]. Moreover, there was little research to identify the relationship between acoustic parameters of the sounds and diverse affective dimensions for a design guideline. To take a more systematic approach to affect-related auditory cue design research, the present paper describes iterative design processes of auditory emoticons (i.e., emotional auditory icons and earcons) and evaluation results of both auditory cues. Additionally, we provide an analysis of their acoustical characteristics for future design guidelines.

2 Iterative Design Processes

Sixteen college students, who major (or minor) in sound design or audio technology at Michigan Tech, created in total 640 auditory icons and earcons for 30 affective adjectives (calm, cold, comfortable, delicate, depressed, dreamy, surprising, fancy, free, fresh/cool, impressive, intimate, magnificent, modern, plain, pleasant, simple, soft, strong, warm, harsh, boring, confused, dark, dynamic, scared, uneasy, angry, disgusting, lively) based on multi-phase design panel discussions [12] under the two sound design experts' supervision. Affective adjectives were selected from previous research using the statistical reduction processes (factor analysis and multidimensional scaling) [13, 14] and a couple of adjectives were added to include basic six emotions [15]. After completing iterative design panel sessions (3 times) and removing acoustically similar sounds, we selected (112 auditory icons and 115 earcons) for further evaluations.

3 User Evaluation

3.1 Method

Thirty three undergraduate students were recruited using the online recruitment system (SONA) at Michigan Technological University. Auditory stimuli were presented via computer and headphones (Sennheiser HD 380 Pro headset). The auditory stimuli used fell into two categories: 1) auditory icons and 2) earcons. Each participant listened to several (2 – 7: *M* = 3.73 for auditory icons; *M* = 3.83 for earcons) sound clips from one of the categories. They could listen to the same sound repeatedly as much as they wanted. After listening, participants were asked to record which of the sound clips best conveyed a specific affective adjective (e.g., angry, fearful, etc.). In total, thirty adjectives were used. Upon completion of the task for one category (e.g., auditory icons), participants did the same for the other category (e.g., earcons). The order of affective adjectives, the order of category (auditory icons and earcons), and the order of sound clip presentation were randomized. Finally, participants were asked to decide between their favorite for each category which better conveyed the specific emotion.

3.2 Results

Table 1. Each row shows an affective adjective, a description of each auditory cue type and the percentage of participants who preferred the sound. * indicates p-values < 0.05.

Affective Adjective	**Description of Preferred Auditory Icon**	**Percentage Preferred**	**Description of Preferred Earcon**	**Percentage Preferred**
Angry	Traffic Jam	52%	Distorted percussive guitar chords	48%
Boring	Sigh	55%	Descending base (plucked)	45%
Calm	Breeze through trees and birds chirping	52%	Dreamy pad	48%
Cold	Wind and shivering	67%	Wind and descending piano notes	33%
Comfortable	Sigh of relief and creaking of chair as sinking in	61%	Woodwind chords	39%
Confused	Quizzical grunt	55%	Pitch bent tuning fork	45%
Dark	1) Thunder clap, 2) Distant ominous sound, 3) Owl hooting	58%	Ominous descending strings	42%
Delicate	Glass breaking	45%	High-pitched Oscillating piano notes	55%
Depressed (sad)	Dog whimpering	39%	Sad piano song	61%
Disgusting	Man Vomiting	64%	Descending deep synthesized tones	36%
Dreamy	Synthetic Pulsing	6% *	Whole tone scale	**94% ***
Dynamic	Crowd Cheering	39%	2 high pitched trumpet sounds	61%
Fancy	Spoon tapping Champaign glass	30% *	Baroque style harpsichord	70% *
Free	Wings flapping and bird chirping	64%	Synthesized choir and chime	36%
Fresh/cool	Water pouring into an ice-filled glass	**70% ***	Funk music baseline	30% *
Harsh	Grating metal	42%	Combination of high pitched keyboard notes	58%
Impressive	Amazed “woah”	55%	Trumpet fanfare	45%
Intimate	Girl pleased “ooh”	18% *	Aura (pad) and bass plus snare	**82% ***
Lively	Cheering and applauding crowd	**70% ***	Ascending synthetic violin with percussion	30% *
Magnificent	1)Trumpet fanfare, 2) Thunder clap	45%	Synthesized choir	55%
Modern	Typing, and cacophony of beeping	24% *	Fuzzy pad and staccato melody	**76% ***
Plain	Typing on keyboard	36%	Single flute note	64%
Pleasant (happy)	Child laughing	70%	3 ascending piano notes	30%

Table 1. (*Continued*)

Scared (fearful)	Woman blood curdling screaming	30% *	Tremolo string sound	**70% ***
Simple	Single tick of clock	48%	Xylophone	52%
Soft	Wobbly bell	42%	Descending piano (with reverb)	58%
Strong	Loud bang	42%	Synthetic bass drum	58%
Surprising	Man short gasp	52%	Ascending Fuzzy Keyboard	48%
Uneasy	Scraping fingernails on chalkboard	36%	Tremolo Keyboard	64%
Warm	Fire crackling	67%	Acoustic guitar chords	33%

Clear trends appeared in preference within sound categories. There were preferences shown for many affective adjectives, as determined by chi-square goodness of fit tests. Further, clear trends in categorical preference arose. To illustrate, there was a strong preference for auditory icon representation of words, such as cool (water poured into ice-filled glass) (p = .024), and happy (laughing child) (p = .024). Meanwhile, earcons were preferred to represent words, such as dreamy (whole tone scale) (p < .001), fancy (Baroque style harpsichord sound) (p = .024), intimate (pad, bass, & snare) (p < .001), and scared (tremolo string sound) (p = .024).

3.3 Discussion

Why are auditory icons preferred sometimes while earcons are preferred other times? To answer these questions it is necessary to review the history of auditory icons [2] and earcons [3]. The seminal works of Gaver and Blattner et al. established the defining characteristics of auditory icons and earcons, respectively. These characteristic features can be used to differentiate the two categories. A closer examination of these may be useful in identifying and understanding user preferences in various contexts.

Auditory Icons. Auditory Icons are often considered analogous to visual icons. Of course, the major difference is the sensory modality, with visual icons utilizing the visual system and auditory icons dependent upon auditory channels. But the analogy is important for understanding the key characteristics of auditory icons.

Similar to visual icons which contain features that human visual systems can detect in parallel (i.e., size, contour, color, etc.) auditory icons can be said to possess analogous features (pitch, tone, volume). Visual icons contain information which represents "real world" actions or events (e.g., an image of a camera represents a camera function in a software program). Auditory icons are the sounds which are paired with those same events (e.g., camera shutter sound on a phone indicating a picture was taken). Auditory icons then, are simply the "naturally-occurring" sounds coinciding with actions and events. This is in contrast to other forms of auditory stimulation like alarms, music, and earcons (to be introduced in the next section). In this view auditory icons can be thought of as the sounds which result from the interaction of real-world objects.

The nature of auditory icons makes them better suited for conveying different types of information. Auditory icons take advantage of the natural mode of listening which is “to identify the events that caused them” [2, p. 169]. Insofar as this the meaning of an auditory icon is universal so too can the effects of auditory icons be considered universal. That is, auditory icons represent interactions between objects in the environment (e.g., sound of a camera shutter, represents real-world actions occurring inside a phone) and as long as the sound of the camera shutter has meaning to the listener it has the potential to convey the same information to that listener. This fact highlights one of the potential benefits of auditory icons, namely, it can transcend language and cultural barriers.

Often auditory icons can easily convey a lot of information. This is because auditory icons map the sounds to their respective sources in a way that takes advantage of premade knowledge structures. In other words, mappings between auditory icons and their sources have already been learned, unlike synthetic mappings (e.g., earcons) which have not been learned.

Gaver [2] also suggests that the auditory icons should be useful for representing dimensional data. For each change in dimension (size, weight, speed, etc.) which is to be represented there is a corresponding change in the sound in the physical world. For example, if an icon is supposed to represent an increase in size of some value, then it can sound heavier (louder, deeper, longer lasting, etc.) This still takes advantages of the preexisting knowledge structures of the listener. This makes the mapping seamless and easy relative to other synthetic sounds and can reduce or eliminate the effort necessary to learn the mapping of the auditory icon to the function it represents.

Earcons. Earcons, in contrast to auditory icons, are not naturally occurring. Blattner et al. defines earcons as “nonverbal audio messages used in user-computer interface to provide information to the user about some computer object, operation, or interaction.” [3 p. 13]. In Blattner’s work she describes earcons in a more inclusive way than is intended in this paper. Earcons, here, are better described as synthetic nonverbal auditory messages. Earcons are called synthetic to mark a distinction between them and auditory icons which are either naturally occurring sounds or caricatures of naturally occurring sounds.

Earcons have specific advantages over other auditory categories, including auditory icons. First, because they are synthetic, earcons can be organized more easily. For instance, earcons can be simple sounds (motives) or they can be more complex, involving multiple layers of simple sounds. In this case, each layer can represent a different detail about the real world it is intended to represent. Earcons can be grouped together into families based on how many features they share. These relationships can be shown hierarchically.

Therefore, earcons have a generative syntax which allows participants learn the meaning of specific earcons without ever having heard them before. In many ways it can be considered similar to be a language of nonverbal sounds. However, learning the relationships between each of the motives and families and the syntax of the earcons can be arduous.

User Preferences. When participants' indicated the sound which they thought best captured each emotion, what factors influenced their decisions? Mapping is likely one of those factors. Sounds which do a poor job of representing the event or action which they were intended to describe should likely do poorly. Conversely, sounds with good mapping should do relatively well. What constitutes good mapping? What follows is an interpretation of what factors might be influencing user preferences in auditory emoticons.

It is plausible to suggest that auditory icons and earcons differ in their mapping ability. Further, these differences might influence user's perception of auditory emoticon effectiveness in representing an emotion. The ability of a sound to represent an emotion is dependent upon its connection to a mental representation of an emotion in the user's memory. Thus, in part, these user preferences can be considered a reflection of each user's past experience and memory. For example, the results show the word "scared" (violin tremolo) was best represented by an earcon. This is likely because scary things are often paired with a violin tremolo sounds in popular media and entertainment. Additionally, the affective state happiness is much more saliently linked to the sound of laughing (auditory icon) than any earcon, at least in the minds' of a significant majority of the participants in this study. The link between the auditory signal and a mental representation residing in memory determines how well each emoticon represents an emotion, and insomuch it informs their preferences. An auditory emoticon which is strongly linked to a mental representation of an affective state will be preferred to a weakly linked emoticon. So, user preferences are subject to the variation in the structure of the users' knowledge and memory. Of course, these preferences are also subject to influence from other variables, such as personality and external factors like stress, mood, and other forms of affect.

Additionally it appears that there could be a large cultural component, especially in the case of earcons. Many of the preferred earcons are similar to sonifications used in entertainment (i.e., movies, TV shows, news media). For example, the violin tremolo is often used to convey fear in movies, the harpsichord (affective adjective: fancy) is often used to represent aristocracy in movies. In fact, in this experiment, a case could be made that all earcons which were chosen are similar to those commonly used sonifications in entertainment.

A further interesting observation is apparent by looking at Ekman's basic emotion set [15] and close emotions to the set. A cursory analysis shows that preferences are almost evenly split between the auditory icons and the earcons. However, a closer look reveals that the earcons are more often chosen for the emotions which lie on the negative (or avoidance) dimension (e.g., sad, scared). Conversely, the auditory icons are chosen for the positive valence emotions (e.g., happy, lively). Even though auditory icons are more salient, if they remind users of unpleasant events or experiences, they could be avoided by users. In those cases an earcon might be preferable, as in the case of "scared" where participants preferred a tremolo violin sound to a woman screaming sound. This could be because of an avoidance of unpleasant sounds like woman screaming in the electronic products.

Consequently, it is expected that auditory emoticons which are strongly linked to mental representations of affective adjectives will outperform auditory emoticons which are weakly linked or unrelated in user's minds. This expectation does not predict that auditory icons or earcons will be better at representing any specific affect, but it does imply that the preferences (overall winners) will be the auditory emoticons which were most commonly linked to affective mental representations. Further, it has been speculated that there could be a relationship between the preference of earcons used in this study and the popularity of those sonifications in media and entertainment. Additionally, it was postulated that Ekman's basic emotions may be treated differentially, with preferences for positive valence emotions to be represented by auditory icons and negative emotions to be represented by earcons. This information suggests that auditory display designers should take into consideration the culture, experience, and memories of their target audience when creating affective auditory displays. Further research is required to investigate the complex nature of the relationship between auditory emoticons and user preferences in auditory displays.

4 Conclusion and Future Work

Emotional auditory cue sets were created and refined by iterative design processes and validated by user evaluation. We are collaborating with international researchers to replicate and extend this study to generalize its implications across different cultures. Moreover, we will construct affective dimensions of emotional auditory cues and compare them with affective dimensions of emotional visual cues to see the commonalities and differences between modalities. As a practical application of the auditory emoticons, we plan to test those cues in various contexts, ranging from mobile devices, telecommunication applications, to in-vehicle infotainment.

References

1. Kramer, G.: An introduction to auditory display. In: Kramer, G. (ed.) Auditory Display: Sonificaiton, Audification, and Auditory Interfaces. Addison-Wesley, MA (1994)
2. Gaver, W.W.: Auditory icons: Using sound in computer interfaces. Human-Computer Interaction 2, 167–177 (1986)
3. 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)
4. Gaver, W.W.: The SonicFinder, a prototype interface that uses auditory icons. Human-Computer Interaction 4, 67–94 (1989)
5. Brewster, S.A.: The design of sonically-enhanced widgets. Interacting with Computers 11(2), 211–235 (1998)
6. 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, Online First Version (2012)
7. 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), 10:11-26 (2011)
8. Schleicher, R., Sundaram, S., Seebode, J.: Assessing audio clips on affective and semantic level to improve general applicability. In: Proceedings of the DAGA (2010)

9. Lemmens, P.M.C., De Haan, A., van Galen, G.P., Meulenbroek, R.G.J.: Emotionally charged earcons reveal affective congruency effects. Ergonomics 50(12), 2017–2025 (2007)
10. Larsson, P., Opperud, A., Fredriksson, K., Västfjäll, D.: Emotional and behavioural response to auditory icons and earcons in driver-vehicle interfaces. In: Proceedings of the 21st International Technical Conference on the Enhanced Safety of Vehicles, Stuttgart, Germany, June 15-18 (2009)
11. Lemmens, P.M.C.: Using the major and minor mode to create affectively-charged earcons. In: Proceedings of the 7th International Conference on Auditory Display, Limerick, Ireland (2005)
12. Pirhonen, A., Tuuri, K., Mustonen, M.-S., Murphy, E.: Beyond clicks and beeps: In pursuit of an effective sound design methodology. In: Oakley, I., Brewster, S. (eds.) HAID 2007. LNCS, vol. 4813, pp. 133–144. Springer, Heidelberg (2007)
13. Lee, J.-H., Jeon, M., Han, K.H.: The analysis of sound attributes on sensibility dimensions. In: Proceedings of the 18th International Congress on Acoustics (ICA 2004) , vol. II, Kyoto, Japan (April 2004)
14. Jeon, M., Lee, J.-H., Kim, Y.E., Han, K.H.: Analysis of musical features and affective words for affection-based music search system. In: Proceedings of the 2004 Korean Conference on Cognitive Science (KCCS 2004), Seoul, Korea (June 2004)
15. Ekman, P.: An argument for basic emotions. Cognition and Emotion 6, 169–200 (1992)
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 SIGCHI Conference on Human Factors in Computing System, Vancouver, BC, Canada (2011)
17. Jeon, M.: Lyricons (Lyrics + Earcons): Designing a new auditory cue combining speech and sounds. In: Stephanidis, C. (ed.) HCII 2013, Posters, Part I. CCIS, vol. 373, pp. 342–346. Springer, Heidelberg (2013)

Natural Forms of Communication and Adaptive Behaviour in Human-Computer-Interaction

Madlen Wuttke[1,*] and Kai-Uwe Martin[2,*]

[1] Institute for Media Research, Chemnitz University of Technology, Chemnitz, Germany
madlen.wuttke@phil.tu-chemnitz.de
[2] Chair of Computer Engineering, Chemnitz University of Technology, Chemnitz, Germany
kai-uwe.martin@informatik.tu-chemnitz.de

Abstract. Scientific research over the last two decades imputes a beneficial effect on human-computer interaction by depicting a virtual communication partner onscreen due to the persona effect and the media equation theory. On the other hand, looking back at the historic component of human-computer interactions, the burden of adaptation has always been on humans to understand the machine and to communicate in accordance with its standards. This paper describes natural communication and interaction strategies of humans and computers as well as their importance to scientific research.

Keywords: Intelligent and agent systems, Pedagogical Agents, Natural Forms of Communication, Adaptive, Mobile Learning.

1 Human-Computer-Interaction

This paper discusses the possibilities of HCI from two perspectives: the social sciences, which aspire to develop a natural, human-like multi-channel communication, and the computer sciences, which engages with technology's capabilities and limitations. Historically, the burden of accommodation has always been on the side of humans [1]. Since machines are being built by humans and we already possess the evolutionary advantage of being able to adapt our behaviour, it seems logical to assume this division of labour. The problem is the hereby artificially limited pool of humans capable of interacting with computers. A broader distribution and application of computer systems in the general population was therefore bound to a more user accessible form of interaction.

Following in this line of thought, the development of different kinds of interfaces and the intuitive knowledge about their functions and usage is key to any successful human-computer interaction. Examples of such are already visible in the market today.

The Samsung Galaxy S4 [2] mobile phone has a feature installed which enables the phone's camera to accurately track the gaze of a user. Although the built-in front camera is not capable of tracking eye-movements, this quit simple gaze-tracking allows for a wide array of helpful and user friendly features. For example in a biometric

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 641–647, 2014.

way which allows to unlock the phone, otherwise it enables the video player to pause a presentation once the user physically diverts his attention away from the phone. Another example is the scrolling of long text documents or websites. Once the gaze of the user is reaching the bottom of the screen, the gadget automatically begins to scroll down – allowing for a more intuitive way of reading text on a hand-held device like a smartphone.

TV manufacturers are currently implementing cameras into their latest products which, obviously, allow for an implementation of video-chat functionalities. Furthermore, it enables the devices to check for people looking at the screen. Devices capable of displaying 3D videos without the help of wearing goggles, by adjusting the surface of the screen, use the camera to identify a user's head and position within the visible space in front of the TV [3].

And, of course, car manufacturers are installing additional ways of interaction into their cars. Examples are vibrating steering wheels once diverting across lanes accidentally, cameras, checking the eyes for signs of fatigue or simply microphones to be able to react to voiced commands.

But these cases of application are examples of specific devices while the implementation of new interactive methods in the case of everyday computer systems is stagnating. The same is the case for the establishment of additional Human-Computer-Interaction-Methods in software with an educational purpose.

2 Designing, Development and Evaluation

When designing, developing and evaluating human computer interactions, it is helpful to use a multi-disciplinary approach to facilitate a broader insight of its processes and requirements. This holistic approach facilitates a more profound approach to the challenges of future HCI research. As a result, this paper describes natural communication and interaction strategies of humans and computers, inspects the technical requirements as well as developments and deals with future possibilities and improvements.

While the direction of interactional development has not changed much since its early stages, the challenge remains to establish a natural form of HCI, meaning to enable computer systems to both acknowledge and interpret multi-facetted user generated input and to create adequate responses.

Especially during the last decade natural input methods like gesture based controls using range cameras, voice recognition based on cloud computing and face recognition based approaches were developed, advanced and established. All these methods share a common base: Whenever either side is trying to influence the other by an interaction, an adapted response is expected. This results in the simple conclusion that every form of interaction demands at least some form of adaptive capability [4].

Research regarding new forms of human-computer interaction is invested in developing either intuitive ways of operating or about implementing new and yet untouched forms of dialogue based communication channels. For example Krämer [5] refers to an increased demand for user-centered behavior by computer systems, since they have the computing power and the capability for computer scientists to produce such systems. She pertains further to the growing number of users which are less and less educated in a certain way about how to engage with computer systems, therefore the natural and intuitive ways of interaction will have to increase. The most sophisticated way of a human computer interaction is based on the principles of human to human communication, face-to-face, as humans are evolutionized to do.

The challenge for establishing a real-life-communication between these two entities is to implement all the audible, tactile, olfactoric, visible and invisible channels used in an everyday conversation. While the human part of HCI is equipped with, and used to, all the communication possibilities, the expected interpretation and equally capable distribution of these non-verbal forms of communication on the side of the computer system demands a daring effort.

Since the 1990s, scientific research is focused on empirical evidence for beneficial effects regarding human-computer interaction. Most experimental setups rely on depicting an ever so lifelike but still virtual communicational agent onscreen. The idea behind it is the evidence shown by the persona effect and the postulations of media equation theory. Ideally this would be a person-like robot, but the depiction of a person onscreen has been proven appropriate for facilitating a connection between user and computer [6], [7]. But while the rendering of a person's face, its mimic abilities and lifelike animations have progressed quite rapidly over the last decade [8], [9], the development in the field of text to speech software is not progressing quickly enough to keep up with the computers abilities to graphically display itself as a photo-realistic communication partner [10], [11].

This discrepancy hinders any dynamic approach of interactive communication to pre-recorded voice samples of another human, talking on behalf of the computer. Besides developing real-life appearances and realistic voice effects, research regarding the other non-verbal cues of communication is even more prolongated.

The specified approaches, from simple Input/Output to adaptive forms of interaction, new forms of interaction and finally natural ways of communication deserve a clear exploration and evaluation matching the socio-cultural requirements with the potential of the current available technology to provide a steady progress in human computer interaction. Since this is a convergence of interdisciplinary scientific research traditions, social sciences for explanation of human behavior and applied information sciences to develop software in accordance with user expectations, the potential for future research is extensive.

3 Input and Output

In order to establish new communication channels between the user and the computer, adequate and accessible hardware has to be implemented. For instance the early printout feedback of machines have been replaced by an electronic monitor to display information, while the previously used machine assembler code has been replaced by an attempted natural form of language which is still used today in the form of ‘if’, ‘then’ and ‘else’. This usage of a more accessible form of communication led to the formation of a dialogue oriented attempt of interacting with machines [5].

Due to this development computers became more accessible, were easier to use for specific tasks and allowed for their embedded use today. To promote this development further, the goal should be to enable computers to be more refined and possess improved forms of interactive input and output. Current input methods are basically improvised forms of communication, although processor power and software development are conceptually ready for a more natural and capable of real-world representation of a communicational counterpart.

In the case of an educational system, this would enable a pedagogical agent to perceive and respond to individual issues of knowledge acquisition. Drastically changing the way of presenting material to a learner who, for the first time, is enabled to react appropriately and naturally once being distracted or unable to follow the presentation [12].

4 Adaptive Forms of Interaction

Looking at human-computer interactions, it is important to state that the idea behind any form of interaction is a palpable result. Humans want to interact with a computer in order to make it compute something. And whenever a computer is prompting an error-message it attempts to communicate to a user to change something regarding the previously attempted interaction. So whenever either side is trying to influence the other by an interaction, an adapted response is expected.

This results in the simple conclusion that every form of interaction demands at least some form of adaptive capability [4]. Even the push of a radio-button on a website results at least in the visual depiction of a selected item. But the attempted effect of an adaptive interaction is most probably to facilitate an open form of dialogue between the user and a machine where one influences the other due to reciprocal communication – being in the shape of a natural language or by selecting metaphorical icons which represent an intuitive function like the recycle-bin.

5 New Forms of Interaction

Research regarding new forms of human-computer interaction is invested in developing either intuitive ways of operating or about implementing new and yet untouched forms of dialogue based communication channels. And due to the success of social network sites, the development which begun with the already mentioned metaphorical icons via apparent impasses like the data glove, is now picking up speed mostly because of the possibility to finance new and creative ideas via social-online-crowdsourcing platforms [13], due to which the different ways of interaction might increase quite rapidly over the next couple of years.

This is also true for the formation of new communication channels which is dependent on progress and the development of better and faster hardware. But a simple webcam to track gaze and infer attention can be enough to enhance the way we interact with a computer system, since it opens up a visual channel for the system to act on behaviour instead of relying on the keyboard and mouse interface to simply react to the user.

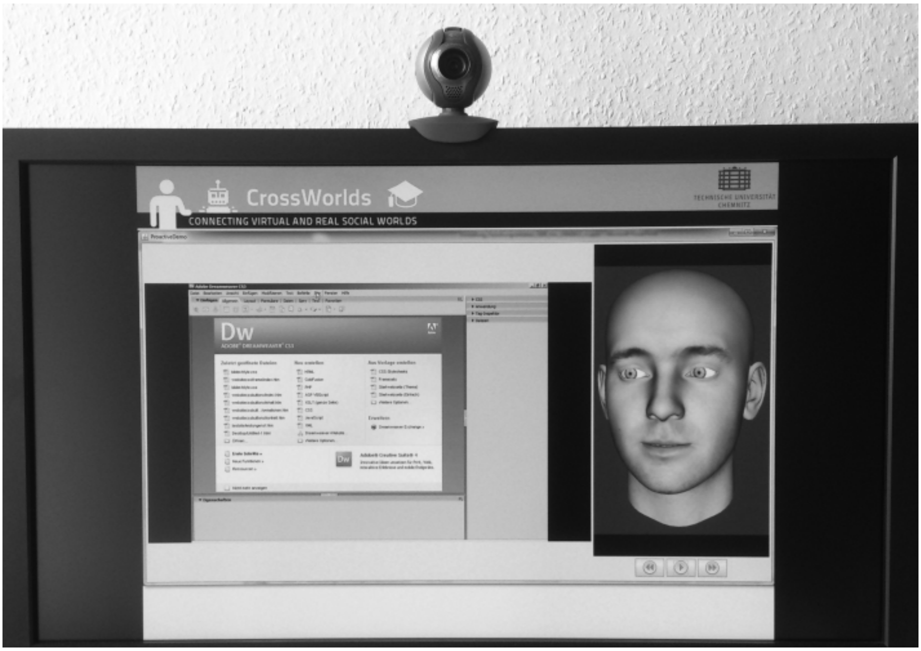

CrossWorlds
CONNECTING VIRTUAL AND REAL SOCIAL WORLDS
TECHNISCHE UNIVERSITÄT CHEMNITZ
Dw

Fig. 1. Educational software with a gaze perceptive pro-active pedagogical agent

6 Educational Settings

An educational setting is per se a program which relies heavily on interactive processes between a learner and some form of teaching software. Since the implementation of pedagogical agents seems to have a beneficial effect on the retention and transfer of knowledge [6] the enhancement of the dialogue between learner and software is paramount. So called ‘conversational agents’, have a high impact on research efforts. And again, a weak point in human-machine interaction is the limitation to the few already implemented channels of communication. Compared to a real-world face-to-face, teacher to student communication setting, additional information is being conveyed than currently possible.

But as in the rest of HCI research, most of empirical studies is focused on checking for beneficial or hindering aspects in appearance or social behaviour. Instead, research should be broadened to include not simply methods of instructional design but also to include real-world expectations of behaviour. For example if the noise level of the environment might be distracting to the learner. Or if a student’s gaze is directed at the correct area of the screen.

As described by Wuttke [12] the pro-active educational aspects might best be dealt with inside a distinct module, called the Electronic Educational Instance (EEI).

The EEI is basically an add-in component capable of being implemented in a variety of information presenting devices like educational software or even recreational devices like Smart-TVs or even cars. Inside the EEI all the sensory information is

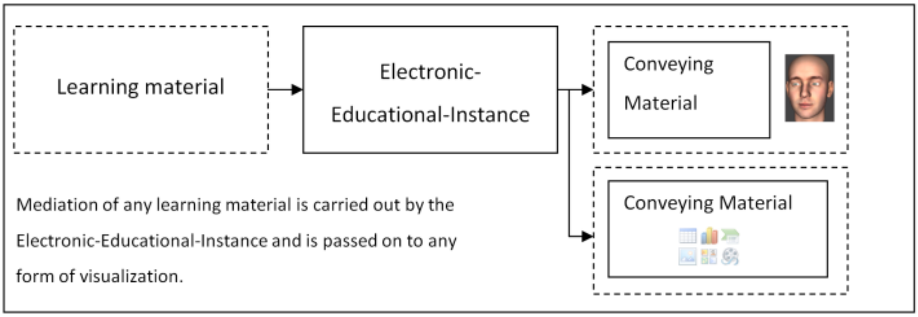

Learning material
Electronic-Educational-Instance
Conveying Material
Conveying Material
Mediation of any learning material is carried out by the Electronic-Educational-Instance and is passed on to any form of visualization.

Fig. 2. Electronic Educational Instance (EEI)

gathered and situational context is created. Allowing for the device to be aware of potential disturbances and enabling it to react adequately.

7 Mobile Systems

The presented forms of interactions and communication have mainly been discussed within the background of traditional computer systems. With the ongoing development and growing distribution of mobile devices like smartphones and tablets, the research is continually focused on developing explicitly intuitive forms of communicating with the device. Traditional forms of providing input like a keyboard and mouse have to be represented onscreen in order of providing these new devices with the advantage of being easy to transport and easy to handle. Touch-based input is as intuitive as a form of input can get – you see, you touch (or push) – whatever reaction one wants to evoke.

Speech-recognition and optical character recognition (OCR) are other forms, which essentially attempt to emulate human perception systems. As an intermediate step, the ubiquitous distribution of Quick-Response-Code (QR-Code) is another example how human communication is adapted to the lower level of computer recognition. OCR still is simply too inaccurate to reliably implement it as a form of input, but due to the QR-Code, letter recognition is not any longer a prerequisite for transferring textual information from the real world into the computer system. While augmented reality applications enable users to perceive the real world and the additional layer of supplementary information, accessible only by using a computer system.

In the case of mobile learning applications, the sensory equipment which is presently installed into a handheld device enables a time- and space-dependent variation of displaying the teaching materials. For example the GPS would allow the application to accurately predict the location of a learner. In the case of the situation 'waiting for the bus' a non-verbal and easy to understand explanation could be offered in the form of a small knowledge-nugget. While in the case of being at the library it would show a textual but more elaborated chunk of knowledge. And once the device's GPS and the wireless network at home is detected, it will produce an auditory chunk of knowledge like teaching vocabularies.

8 Conclusion

A wide array of issues has been discussed regarding the ways and capabilities of human computer interactions. While the direction of interactional development has not changed much since its early stages, the challenge remains to establish a natural form of HCI, meaning to enable computer systems to both acknowledge and interpret multi-facetted user generated input and to adequately respond in kind. Being able to use all the aforementioned channels human would use in a common conversational setting as well – both verbal and non-verbal.

Since this is a convergence of interdisciplinary scientific research traditions, social sciences for explanation of human behaviour and applied information sciences to develop software in accordance with user expectancies, the potential for future research is extensive. To facilitate this research, a theoretical discussion of the essential aspects of expected components and behaviour has been presented.

References

1. Draper, S.W., Norman, D.A.: Introduction. In: Norman, D.A., Draper, S.W. (eds.) User Centered System Design: New Perspectives on Human-Computer Interaction, pp. 1–6. Erlbaum, Hillsdale (1986)
2. Galaxy S4, http://www.samsung.com/us/topic/our-galaxy-smartphones (retrieved on January 31, 2014)
3. Toshiba 55zl2, http://www.toshiba.de/contents/de_DE/PRODUCT_DESC/files/1134731.pdf (retrieved on January 31, 2014)
4. Niegemann, H.M., Domagk, S., Hessel, S., Hein, A., Hupfer, M., Zobel, A.: Kompendium multimediales Lernen. Springer, Heidelberg (2008)
5. Krämer, N.C.: Soziale Wirkungen virtueller Helfer: Gestaltung und Evaluation von Mensch-Computer-Interaktion. Kohlhammer, Stuttgart (2008)
6. Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., Bhogal, R.S.: The persona effect: Affective impact of animated pedagogical agents. In: Pemberton, S. (ed.) Human Factors in Computing Systems: CHI 1997 Conference Proceedings, pp. 359–366. ACM Press, New York (1997)
7. Reeves, B., Nass, C.: The Media Equation: How people treat computers, televisions, and new media like real people and places. Cambridge University Press, New York (1996)
8. Kim, Y.: Learners' expectations of the desirable characteristics of virtual learning companions. International Journal of Artificial Intelligence in Education 17(4), 371–388 (2007)
9. Agada, R., Yan, J.: Research to Improve Communication by Animated Pedagogical Agents. Journal of Next Generation Information Technology 3(1), 58–69 (2012)
10. Atkinson, R.K., Mayer, R.E., Merril, M.M.: Fostering social agency in multimedia learning: Examining the impact of an agent‘s voice. Contemporary Educational Psychology 30, 117–139 (2005)
11. Veletsianos, G.: The impact and implications of virtual character expressiveness on learning and agent–learner interactions. Journal of Computer Assisted Learning 25, 345–357 (2009)
12. Wuttke, M. (2013). Pro-Active Pedagogical Agents. In: Fakultät für Informatik (Ed.). Proceedings of International Summerworkshop Computer Science, pp. 59-62 (July 2013)
13. Marjanovic, S., Fry, C., Chataway, J.: Crowdsourcing based business models. In: Search of Evidence for Innovation 2.0. Science and Public Policy 39, pp. 318-332 (2012)

Human-Robot Interaction

Backchannel Head Nods in Danish First Meeting Encounters with a Humanoid Robot: The Role of Physical Embodiment

Anders Krogsager[1], Nicolaj Segato[1], and Matthias Rehm[2]

[1] School of Information and Communication Technology, Aalborg University, 9000 Aalborg, Denmark
{akrogs09,npeder09}@student.aau.dk

[2] Faculty of Engineering and Science, Aalborg University, 9000 Aalborg, Denmark
matthias@create.aau.dk

Abstract. Head nods have been shown to play an important role for communication management in human communication, e.g. as a non-verbal feedback signal from the listener. Based on a study with virtual agents, which showed that the use of head nods helps eliciting more verbal input from the user, we investigate the use of head nods in communications between a user and a humanoid robot (Nao) that they meet for the first time. Contrary to the virtual agent case, the robot elicited less talking from the user when it was using head nods as a feedback signal. A follow-up experiment revealed that the physical embodiment of the robot had a huge impact on the users' behavior in the first encounters.

Keywords: Culture-aware robots, backchannels, feedback, physical embodiment.

1 Introduction

Robots have begun to move from restricted environments that are specially designed for them into public and semi-public spaces where they are envisioned to interact in a socially acceptable manner with users. Head nods have been shown to play an important role for communication management in human communication, e.g. as a non-verbal feedback signal from the listener. Humans are very good in creating opinions about a communication partner based on first impressions from initial meetings. From cross-cultural studies we know that using the wrong social signals in these first encounters easily lead to severe misunderstandings between the communication partners. One aspect of the many social signals is backchannel feedback, specifically head nods. In a previous Japanese study with virtual agents [11] it was shown that the use of head nods helps eliciting more verbal input from the user when they are congruent with culture-specific head nod patterns, in this case for the Japanese culture in contrast to US American patterns. Based on this results, we present a replication of this study here that changes two parameters:

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 651–662, 2014.

(i) The cultural background of the users: Targeting Danish users, we concentrate on Danish nodding patterns based on the analysis of a multimodal corpus of first meeting encounters (NOMCO).
(ii) The embodiment of the agent: Instead of using a virtual character, we replicate the experiment with humanoid robot (Nao), assuming that the physical embodiment will have an impact on the results.

The paper first presents related work in the area of virtual and physical agents. Then the replicated experiment and results are presented. Results show a strong influence of the physical embodiment leading to a follow up study with a virtually present robot, which is presented next before the paper concludes with a discussion.

2 Related Work

Several studies on virtual agents have shown that the paradigm of a listener agent has a good potential of building rapport and engaging the user in prolonged interactions [5–7]. An important aspect is the production and recognition of appropriate social signals in order to realize affective interactions, which are seen as a prerequisite for successfully establishing rapport with the user and it can be safely assumed that this also holds true for interactions with physically embodied agents, i.e. robots. Research on head nods in robots have so far mainly been concerned with recognizing and interpreting head nods by human users (e.g. [8]) but not so much for employing head nods as a means for the robot to structure and maintain the dialogue with the user. Exceptions are the work by [9] and [10].

As has been acknowledged previously, some parameters of head nods seem to vary across cultures like the frequency of head nods in dyadic conversations. Koda and colleagues [11] present an experimental setup for analyzing this cross-cultural variety for a virtual agent system. They showed that human users speak longer to an agent that takes these cultural differences in the realization of head nods into account. Shortcomings of their approach include the fact that they only tested on Japanese subjects. Thus the reported results might be attributable to the fact that more nodding generally elicits more talking from the speaker.

Here we will use the basic experimental setup to test if Danish participants would also prefer to talk longer if a humanoid robot displays culturally adequate feedback signals in terms of head nods. In order to realize this experiment, more information on head nods is necessary. Head movements in general are a well researched feature of human communication focusing on the physical movement itself, on how to classify different movements as well as on the communicative function of the different movements.

McClave [13] distinguishes between two motions for the American culture, an up/down movement (nod) used to signal affirmation and a side to side movement (shake) to signal negation. Allwood and Cerrato [14] present several relevant head movements and distinguish between nod (forward movement of the head going up and down, which can be multiple), jerk (backward movement of the head which is usually single), shake (left-right or right-left movement of the head which can be multiple), waggle (movement of the head back and forth left to right), and 'swturn' (side-way turn is a single turn of the head left or right). Based on these earlier suggestions, Paggio and Navarretta [15] classify head movements into Nod, Jerk, Head-Forward, HeadBackward, Tilt, Side-Turn, Shake, Waggle and HeadOther.

The physical features of head movements have been the focus of Hadar and colleagues [16], who present a number of different results concerning frequency, amplitude, and cyclicity. They report that subjects exhibited head movements every 7.41 seconds on average (frequency of 8.1 movements/minute) without distinguishing between nods, shakes or other movements. The data was also used to analyze the correlation between the amplitude of a head movement and the conversational function, showing e.g. that a mean amplitude of 13.3 degree can be observed with an affirmation ('Yes') and 11.4 degree with a movement that is synchronous to speech. Results are in so far questionable as the means for measuring are very obtrusive and required the subjects to wear a specific head mounted equipment, making the situation far from natural. Also, the analysis is based on a data corpus of around 16 minutes in total. McClave [13] presents some data from Birdwhistell relating to the velocity that can be observed in head nods. She reports the typical velocity range among Americans of 0.8 degrees to 3 degrees per 1/24 second over a spatial arc of 5 to 15 degrees. Maynard [17] is concerned with the frequency and distribution of head nods and presents an in-depth analysis for Japanese dyadic interactions. His analysis reveals that Japanese do one head movement per 5.75 seconds on average (frequency of 10.4 nods/minute) in contrast to Americans with a movement every 22.5 seconds on average (frequency of 2.7 movements/minute). The distribution of head nods between speaker and listener is almost balanced with listeners being responsible for 44% of head nods while speakers are doing 56% of the nods. Paggio and Navarretta [2] present similar data derived from a Danish corpus, which consists of 12 first meeting encounters with a total duration of around 51 minutes. Based on this data, Danish participants nod on average every 5.82 seconds (frequency of 10.3 nods/minute). The above studies reveal a cultural difference in the frequency of head nods, with US Americans nodding less frequently compared to Danish and Japanese.

Apart from information about the physical qualities of head movements, literature on the function of head movements and specifically head nods is vast. In an early study, Dittmann and Lewellyn [12] focus solely on up/down movements, which are recorded by a tailor-made device that the subject had to wear on his head. They attribute two functions to these head nods, either a signal for the speaker that the listener intends to get the floor or as a feedback signal to the speaker. In both cases vocalizations may accompany the head nod, but the head nod may well precede the vocal channel. Heylen [18] gives a comprehensive overview of functions associated with head nods that draws from multiple sources. He distinguishes between 26 different functions but is a bit fuzzy on the use of head nods, as some functions are more associated with gaze than with head nods or include posture changes. In a similar fashion, McClave [13] presents a range of different functions for head nods from semantic over narrative to interactive. Kogure [19] shows that frequent nodding is a phenomenon observed in Japanese conversations whenever a silence in the conversation occurs (so called loop sequence). Thus, they distinguish nods with and without accompanying speech for their analysis. Maynard [17] specifically analyzes Japanese head movements in contrast to American ones and lists the following interactional functions: (1) affirmation; (2) claim for turn-end and turn-transition; (3) pre-turn and turn claim; (4) turn-transition period filler; (5) back channel, and (6) rhythm taking.

Allwood and Cerrato [14] show head movements to be the most frequent feedback signal in dyadic conversations with nodding either single or multiple being by far the most frequent signal they found. In a follow-up analysis Boholm and Allwood [20] show that a majority of multiple head nods accompany speech that also expresses the feedback information (74%).

To sum up, head nods are an important non-verbal feedback signal in human communications. They are found across cultures with variations in their actual realization, e.g. regarding their timing and frequency in an interaction.

3 Online Survey

In order to establish a baseline for the experiment with the humanoid robot, a repeated-measures online survey was conducted to determine which style of head nodding is preferred by Danish users in the context of a listening robot. Head movements were derived from existing video material of students in dyadic first encounter conversations. The videos were analyzed for nodding patterns in velocity, frequency and angle magnitude. Based on these three variables eight varieties of head nods were defined, programmed into a Nao robot and then video recorded. The eight videos shows the robot passively listening to a voice and nodding, where each video depicts a different value combination for the three variables.

After watching each video, participants were asked to report how well they liked the style of head movement according to an 11 point Likert scale. They were also asked to report what emotions they thought the robot seemed to express from a list of 10 arbitrarily selected emotions, equally distributed between positive and negative affect.

Online Survey Results. 41 participants completed the surveys, 24 men and 17 women, with ages ranging from 20 to 57 years (median = 25). The Likert ratings of the videos were analyzed using the Friedman test. The analysis showed no significant results of the rating between the videos. The head nod that was chosen for further experimentation was the head nod that got the highest average Likert score and that correlated with the highest number of positive emotions.

4 Experiment 1: Co-location

From human interaction it is known how feedback positively influences conversation, and the experiment presented in [11] has shown this relates to interactions with a virtual agent. Based on these insights, it can be assumed that users talk longer with a robot which uses culturally appropriate head nods, compared with head nods form another culture or no movement at all. Moreover, it has been shown that speaking activity is a good predictor of the extraversion trait [21]. In order to test these assumptions, an experiment with a physically embodied agent in the form of the Nao robot has been designed with the following hypotheses:

H1. *A robot that nods elicits longer stories from the user compared to one that does not nod.*

H1a. *A robot that shows culture-specific nodding behavior elicits longer stories form the user compared to a robot that shows unspecific or no nodding behavior.*

H2. *Participants scoring high in extraversion will talk considerably longer independent of the experimental conditions compared to user that score low on extraversion.*

H3. *The user will perceive the robot as more intelligent when it elicits backchannel head nods.*

An independent measures Wizard of Oz experiment is conducted. The independent variable in this experiment is the backchannel feedback of the robot. Before the session, participants were informed that they were going to talk to an intelligent robot, that will listen to them but otherwise remain passive. The automatic head-tracking of the Nao was activated to simulate eye contact. Participants were asked to talk to the listing robot about an open-ended, preselected topic from a list of 15 topics[1]. Participants are randomly assigned to either a control group or one of two groups with backchannel feedback. The dependent variable is the duration of how long the participant speaks. Participants are asked to talk to the robot about the chosen topic as long as they can, but for practical purposes are stopped if they speak for more than five minutes. The test leader observes the conversation and heuristically triggers head nods remotely and without the participant knowing. After the test, participants were required to fill out a short questionnaire regarding personality, impression of the robot and demographic. Figure 1 demonstrates the setup.

Participants. Recruiting was done at a 'university college'. All participants were native Danish speaking students with limited knowledge about robots and had various academic backgrounds. They were debriefed after the experiment.

[1] Fifteen topics: fashion, sports, pets, food, books, movies, music, travel, work, studies, games, cars, vacation, hobbies and ambitions

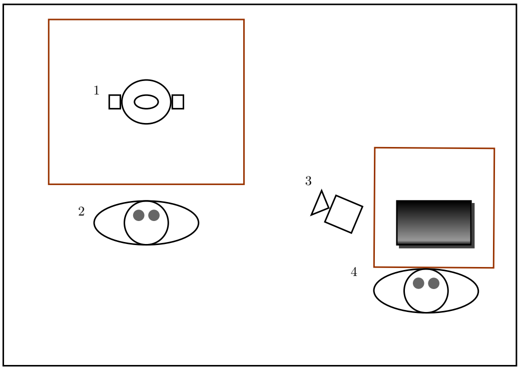


Fig. 1. Top-down sketch of the experiment setup: 1. Nao Robot, 2. Participant, 3. Recording camera, 4. Test facilitator with laptop

Apparatus. The study used a Nao H25 robot by Alderbran robotics. A script was written to trigger the robot to nod upon keyboard input. The remote triggering of the robot was done on a laptop computers with an Intel i7 processor. A video camera was used to record each test session.

Extraversion Measures. Extraversion of each participant was acquired using a shorter version of Eysenck's revisited Eysenck Personality Questionnaire (EPQR-A) [4]. The EPQR-A was administered prior to the each session and only the extraversion dimension was used.

Perceived Intelligence. The participants' perceived intelligence of the robot was obtained using part of Bartneck et al.'s "Godspeed" questionnaire [3]. The questionnaire consists of a series of mutually opposing adjectives, concerning intelligence, working as anchors. The questionnaire consist of five five-point Likert scale questions.

4.1 Results of Co-located Experiment

Twelve female and eight male students participated in the experiment and talked to a physically present robot. Figure 2 shows an example from the test. Their age ranged from 20 to 49, mean = 25, SD = 6.5. Four participants interacted with the robot in the American nodding group and spoke on average 42.8 seconds (SD=9.6). Five participants were in the Danish nodding group and spoke on average 44.6 seconds (SD=9.5). The larger control group (no nodding; NN) had eleven participant who spoke on average 93.8 seconds (SD=28.5). Hypothesis 1 is tested by running two independent measures t-tests between DK-NN and US-NN groups. Bonferroni correction is applied and so $\alpha = (0.05/2) = 0.025$.

The speech duration of participants in the DK group was significantly shorter than the control group t(14) = 3.7, p<0.025, r = 0.7. The speech duration of

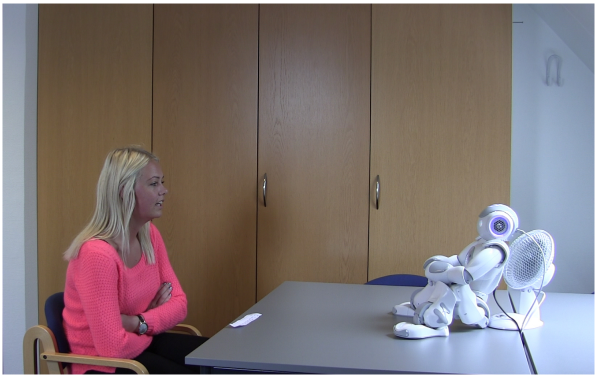


Fig. 2. Image of a participant engaging with the robot. The robot remains in the sitting position throughout the experiment and maintains eye contact.

participants in the US group was significantly shorter than the control group t(13) = 3.4, p<0.025, r = 0.69. Contrary to the virtual agent study by Koda et al. the robot elicited less talking from the user when it produced backchannel head nodding. Thus hypothesis 1 is rejected.

To test hypothesis 1a an independent measures t-test is run between DK-US groups. It shows that there is no significant difference between speech duration: t(7) = 0.289, p<0,05. Thus hypothesis 1a is rejected.

Pearson's correlation coefficient is calculated for the relation between extraversion and duration of speech of a participant in the co-location experiment. There was a positive correlation between the two variables, r = 0.524, n = 20, p = 0.018. A scatter plot of the data is shown if figure 3. Hypothesis 2 is retained.

The results of the Perceived Intelligence are analyzed by comparing the scores of each question between participants of the control group (n=11) and US-DK group combined (n=9). Five independent t-tests, $\alpha = 0.05$, are run. They all showed non-significant difference except for question 5; t(18) = -2.87, p < 0.05, r = 0.56. Thus hypothesis 3 is rejected. While statistically insignificant there was a slight tendency for participants to rate the robot more intelligent on average when they interacted with the robot that elicited feedback. Participants in the nodding-free control condition rated it to be less intelligent.

5 Experiment 2: Virtual Presence

Based on the unexpected outcome of the co-location experiment another experiment is conducted. In this, the independent variable is changed to a virtually present robot that performs Danish head movement to make it more similar to the original study.

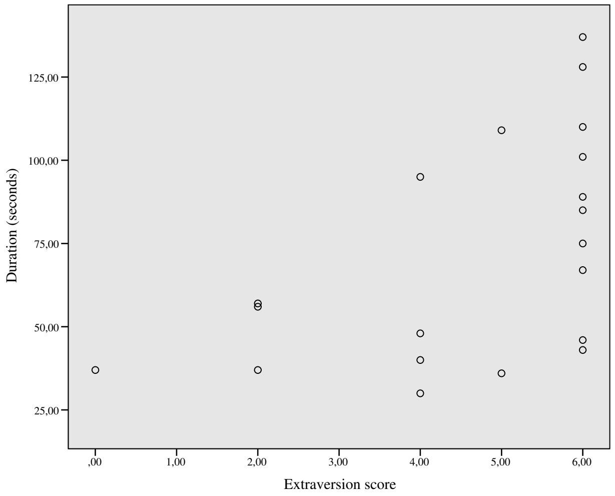


Fig. 3. The measured extraversion of participants in the co-location experiment correlates positively with their duration of speech

The experiment was run with just one condition, the Danish head nodding behavior as a comparison to the previous results. The test was conducted using the same Wizard of Oz method as in the first experiment. The test facilitator is seated with the Nao robot in a separate room and the participants speaks with the robot through a Skype call. The robot performs the same Danish head nod movements as in the first experiment. The same questionnaire data regarding perceived intelligence, extraversion and demography are collected. The following hypothesis is guiding the experimental setup:

H4. *Users will speak considerably longer to a robot that is only virtually present compared to a robot that is physically co-located in the room.*

As in the first experiment participants are asked to speak to a robot about one of the 15 topics. They are given the same instructions as in the first experiment except they are required to answer a Skype call with the robot. The test conductor leaves the room with the explanation that he is monitoring the Skype call remotely.

5.1 Participants

Nine participants could be won, 4 females and 5 males, all native Danish speakers with an age range from 22 to 64 years (median = 25) of which the majority were students.

5.2 Apparatus

The participant laptop had a 15.6 inch screen and was placed on a table in front of the participant. An external microphone is plugged into the laptop and placed

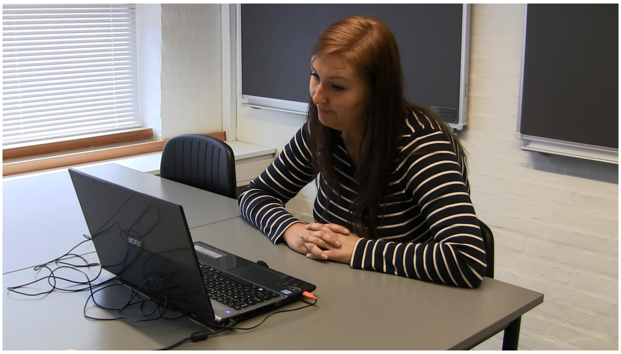


Fig. 4. Image of a participant engaged in a Skype conversation with the robot. The robot remains in the same sitting position as before but only upper body and head is visible on the screen.

in front of the laptop to ensure the facilitator clearly hears the participant during the test. Figure 4 shows the setup of the Skype experiment.

5.3 Results

The 9 participants in the virtual presence experiment spoke on average for 204,7 seconds (SD = 93.2). This is compared in an independent measures t-test, $\alpha = 0.05$, with the results of participants in the co-location experiment in the DK group (n = 5, mean = 44.6, SD = 9.5). Participants in the virtual presence experiment spoke significantly longer: t(8.29) = -5.08, $p < 0.05$. The hypothesis H4 is thus retained. Participants' average extraversion was 5.5.

5.4 Discussion

Participants indeed talked significantly longer when the robot was not physically present in the room, on average more than twice as long as in the control condition. This is also in line with the findings by Koda and colleagues for the virtual agent case [11]. It should be noted that participants in this group had high extraversion which may partially account for the high duration. Thus, we can conclude the physical presence of the robot has a huge impact on the users' behavior, at least in cases where users meet a robot for the first time in their life. It remains to be shown if this effect vanishes, when users get more familiar with the robot, e.g. in subsequent sessions.

6 Discussion and Limitations

The results of the first study contradict the assumptions made from human communication and previous studies with virtual agents. The head movement of the robot negatively influenced the speech duration of participants compared to participants who spoke to the robot that did not produce any head feedback. We speculate that the physical presence of a robot is the cause of this outcome, as a contrast to the virtual agent of the original Japanese study. The difference between co-located and virtual presence, could have been influenced by the presence of the test facilitator during the co-located test and have caused discomfort of the participants. On the other hand, this would not explain why participants spoke longer in the control condition (no nod). The experiment noticeably differs in that it uses a robot, compared to the original experiment which uses a virtual character. We assume that the results might be attributed to either the use of a robot or the presence of the test facilitator during the test, encouraging for future experiments.

7 Conclusion

Danish participants spoke to a robot under different condition of presence and backchannel feedback. Contrary to our hypothesis the duration of speech was significantly shorter when the robot produced head movement compared to a control condition with no head nods. There was no significant difference in speech duration whether the backchannel feedback of the robot was culture specific. As expected a positive correlation was found between speech duration and extraversion. Participants spoke significantly longer when the robot was virtually present. It is not trivial to replicate human communication in interaction with robots. Nor is it a matter of simply reproducing communication signals in a robot to make users interact with it as if it were human.

Acknowledgements. The present work benefitted from the work of Constanza Navaretta and Patrizia Paggio researchers at ”Center for Sprogteknologi” at Copenhagen University who provided the NOMCO data. Moreover the experiments benefits from the help of Vagn E. Pedersen, Sarah Catterine, Birthe Brøndum and Harry Brøndum for their help in running the experiments described in this paper.

References

1. Paggio, P., Navaretta, C.: Feedback and gestural behaviour in a conversational corpus of Danish. In: Proceedings of the 3rd Nordic Symposium on Multimodal Communication NEALT (2011), pp. 33–39 (2011)
2. Paggio, P., Navarretta, C.: Head Movements, Facial Expressions and Feedback in Danish First Encounters Interactions: A Culture-Specific Analysis. In: Stephanidis, C. (ed.) Universal Access in HCI, Part II, HCII 2011. LNCS, vol. 6766, pp. 583–590. Springer, Heidelberg (2011)

3. Bartneck, C., Kuli, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics 1, 71–81 (2009)
4. Francis, L., Brown, J., Philipchalk, L.B., The, R.: The development of an abbreviated form of the revised eysenck personality questionnaire (EPQR-A): Its use among students in England, Canada, the U.S.A. and Australia. Personality and Individual Differences 13(4), 443–449 (1992)
5. Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., Wöllmer, M.: Building Autonomous Sensitive Artificial Listeners. IEEE Transactions on Affecite Computing 3, 165–183 (2012)
6. Yoichi, S., Yuuko, N., Kiyoshi, Y., Yukiko, N.: Listener Agent for Elderly People with Dementia. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 199–200. ACM, New York (2012)
7. Meguro, T., Higashinaka, R., Dohsaka, K., Minami, Y., Isozaki, H.: Analysis of listening-oriented dialogue for building listening agents. In: Proceedings of the SIGDIAL 2009 Conference, pp. 124–127. Association for Computational Linguistics, Stroudsburg (2009)
8. Rich, C., Ponsler, B., Holroyd, A., Sidner, C.: Recognizing engagement in human-robot interaction. In: Proceedings of HRI, pp. 375–382. Institute of Electrical and Electronics Engineers (IEEE) (2010)
9. Riek, L.D., Paul, P.C., Robinson, P.: When my robot smiles at me Enabling human-robot rapport via realtime head gesture mimicry. Journal of Multimodal User Interfaces, vol 3, 99–108 (2010)
10. Liu, C., Ishi, C., Ishiguro, H., Hagita, N.: Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In: Human-Robot Interaction, pp. 285–292. ACM (2012)
11. Koda, T., Kishi, H., Hamamoto, T., Suzuki, Y.: Cultural Study on Speech Duration and Perception of Virtual Agent's Nodding. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS (LNAI), vol. 7502, pp. 404–411. Springer, Heidelberg (2012)
12. Dittmann, A., Llewellyn, L.: Relationship between vocalizations and head nods as listener responses. Journal of Personality and Social Psychology 9, 79–84 (1968)
13. McClave, E.: Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32, 855–878 (2000)
14. Allwood, J., Cerrato, L.: A study of gestural feedback expressions. In: Paggio, P., Jokinen, K., Jönsso, A. (eds.) First Nordic Symposium on Multimodal Communication, Copenhagen, pp. 7–22 (2003)
15. Paggio, P., Navarretta, C.: Head movements, facial expressions and feedback in conversations: Empirical evidence from Danish multimodal data. Journal on Multimodal User Interfaces 7, 29–37 (2013)
16. Hadar, U., Steiner, T., Rose, F.: Head movement during listening turns in conversation. Journal of Nonverbal Behavior 9, 214–228 (1985)
17. Maynard, S.: Interactional functions of a nonverbal sign Head movement in japanese dyadic casual conversation. Journal of Pragmatics 11, 589–606 (1987)
18. Heylen, D.: Challenges ahead: head movements and other social acts during conversations. In: Joint Symposium on Virtual Social Agents. AISB (2005)

19. Kogure, M.: Nodding and smiling in silence during the loop sequence of backchannels in Japanese conversation. Journal of Pragmatics 39, 1275–1289 (2007)
20. Boholm, M., Allwood, J.: Repeated head movements, their function and relation to speech. In: Kipp, M., Martin, J., Paggio, P., Heylen, D. (eds.) Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pp. 6–10. LREC (2010)
21. Lepri, B., Subramanian, R., Kalimeri, K., Staiano, J., Pianesi, F., Sebe, N.: Employing social gaze and speaking activity for automatic determination of the extraversion trait. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, p. 7. ACM (2010)

Recommended Considerations for Human-Robot Interaction Communication Requirements

Stephanie J. Lackey, Daniel J. Barber, and Sushunova G. Martinez

University of Central Florida, Institute for Simulation and Training, Orlando, FL
{slackey,dbarber,smartine}@ist.ucf.edu

Abstract. Emerging robot systems increasingly exhibit greater levels of autonomy, requiring improvements in interaction capabilities to enable robust human-robot communication. This paper summarizes the present level of supervisory control in robots, both fielded and experimental, and the type of communication interfaces needed for successful Human-Robot Interaction (HRI). The focus of this research is to facilitate direct interactions between humans and robot systems within dismounted military operations and similar applications (e.g., law enforcement, homeland security, etc.). Achieving this goal requires advancing audio, visual, and tactile communication capabilities beyond the state-of-the-art. Thus, the requirement for a communication standard supporting supervisory control of robot teammates is recommended.

Keywords: Supervisory control, autonomy, human-robot interaction.

1 Introduction

Combat teams increasingly consist of human ground troops and robot/unmanned assets. Robot assistance comes in the form of weaponized platforms, spy drones, and other Intelligence, Surveillance, and Reconnaissance (ISR) vehicles. The National Defense Authorization Act for Fiscal Year 2001 mandates the Armed Forces dramatically increase the use of unmanned and/or remotely operated systems to one-third of the ground combat vehicles employed in theatre [1]. The Unmanned Systems Roadmap presents strategies for meeting requirement by describing master plans for unmanned air, ground, undersea, and surface systems over the next 25 years [2].

A critical enabling capability for the successful implementation of tactically advantageous, but disruptive, robot systems within dismounted operations is effective and efficient HRI. Transitioning from continuous remote control (i.e., teleoperation) to supervisory control is essential to advancing HRI methods and to optimizing the employment of emerging military robot systems. Sheridan defines supervisory control as "…one or more human operators intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors and sensors to the controlled process or task environment."[3]. In other words, a human operator/controller commands a robot through a computer interface on a discrete rather than continuous basis. The operator programs

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 663–674, 2014.

actions the robot will take and the robot then executes. Throughout the tasking, the robot interacts with the operator and outputs information. This represents a closed-loop command and information exchange. For the purposes of this effort, supervisory control is operationally defined as operators/team members exercising discrete control of a robot, made possible by some level of autonomy the robot possesses.

Enabling the seamless integration of human and robot assets within mixed-initiative teams requires system developers to address four key concepts: (1) the type and level of automation inherent to the robot; (2) how humans communicate with each other; (3) interface design characteristics; and (4) human capabilities and limitations. The purpose for the present effort is to discuss each of these important topics with the aim of facilitating development of a supervisory control standard to drive development of next-generation robotic systems.

2 Level of Interaction

Three levels of HRI comprise supervisory control based on Rasmussen's "skills, rules, knowledge" and include: skill-based, rule-based, and knowledge-based human behavior [3] [4]. Skill-based robots perform a specific skill(s) with a given command. Rule-based robots are programmed to recognize certain stimuli and from those stimuli make pre-specified decisions. Knowledge-based robots essentially learn from their environment and make decisions based on that knowledge. When applying this paradigm to conventional robots and unmanned systems, skill and rule-based behaviors are found in currently fielded systems. Studies in knowledge-based behaviors continue to expand the area; however additional advancements are required prior to military deployment.

The three levels of HRI map directly to a robot's level of automation. Each branch of the military may tailor these definitions, but the core concepts apply to all. For example, leading research conducted by the U.S. Army suggests ten levels of automation known as Parasuraman's Levels of Autonomy [5] [6]. On this scale, one represents a system fully controlled by the human, and ten, represents a fully automated system. This concept parallels the two extremes in Rasmussen's Paradigm ranging from complete manual control to full autonomy. Using Rasmussen's paradigm, the following examples illustrate the levels of autonomy exhibited by various fielded and experimental systems.

2.1 Skill-Based Robots

Skill-based robots typically provide lower levels of autonomy, and require teleoperation to perform a particular skill given a command [3]. An example of this type of robot is a PackBot called the Valkyrie (see Figure 1). This remotely controlled Unmanned Ground Vehicle (UGV) used to extract a fallen Soldier from enemy fire is deployed to the Soldier in need. After the Soldier places him/herself on the Sked (i.e., bed), the robot encapsulates and carries them to safety [7].

2.2 Rule-Based Robots

Rule-based systems use predetermined stimuli to follow the commands of an “if-then” algorithm, [3]. The Global Hawk qualifies as one of these systems (see Figure 1). The main tasking for the Global Hawk is reconnaissance and surveillance achieved by employing an experimental multi-agent system. This system includes multiple UAV’s controlled via a single operator interface used to input mission relevant information. Global Hawk’s Human-Machine Interface (HMI) interacts directly with the operator and simultaneously maintains control of other system elements [8]. For example, when a lead UAV goes missing or becomes dysfunctional the HMI system reassigns an existing UAV on the same mission as the new lead. This occurs without input from the operator [9].

Similar to the Global Hawk, the Predator’s tasking focuses on navigation and surveillance. The Predator is equipped with hellfire missiles, which differentiates this system from the Global Hawk [11]. A need for armed combat systems capable of getting close to targets prompted the U.S. military to equip the Predator with weapons. Although this platform does not require continuous human input for flight and navigation, it relies on operator input for missile deployment (see Figure 1).

The Black Knight Unmanned Ground Combat Vehicle (UGCV) is an example of a research tool exercising ruled-based capabilities (see Figure 1). The Black Knight provides an operational test environment for Soldiers assisting in the development of UGVC tactics, techniques, and procedures. Capabilities such as a rule-based autonomous navigation system evolved from empirical experimentation. The Black Knight generates a path using autonomous capabilities that synchronize perception and path planning subsystems rather than relying on an operator to manually supply a route. The perception system senses obstacles and hazards, and coordinates with the path planning system to reroute and avoid detected obstacles [13]. Even though the Black Knight is not fielded, experimentation with this type of system demonstrates a foothold for future military vehicles.

2.3 Knowledge-Based Robots

Knowledge-based robots, as defined by Rasmussen, possess capabilities that assess a situation, and perform certain actions by considering multiple goals, decision points, and scheduling aspects [3]. Furthermore, robots with such capabilities currently exist in the research and development stages. An example scenario illustrating the goal level functionality of a knowledge-based robot is a commander tasking a robot to, “go to the back of a building and send me a picture of any person that leaves wearing a red shirt.” In this scenario, the system must identify the following subtasks: (1) move towards the back of the building, (2) monitor for someone exiting the building, (3) determine the person is wearing a red shirt, (4) take a picture of the person in a red shirt, and (5) report back to the operator with a notification and image. Accomplishing a task of this complexity requires knowledge-based system to understand the main objective, finding someone in a red shirt, and prioritize subtasks based on the main objective. If the system detects someone in a red shirt leaving the building before reaching the back of the building, it needs to execute the main objective of notifying its commander of the target of interest. These types of systems have yet to progress to field operations. Systems fielded today support skill and rule-based abilities; however a supervisory control protocol should support future capabilities including knowledge-based systems.

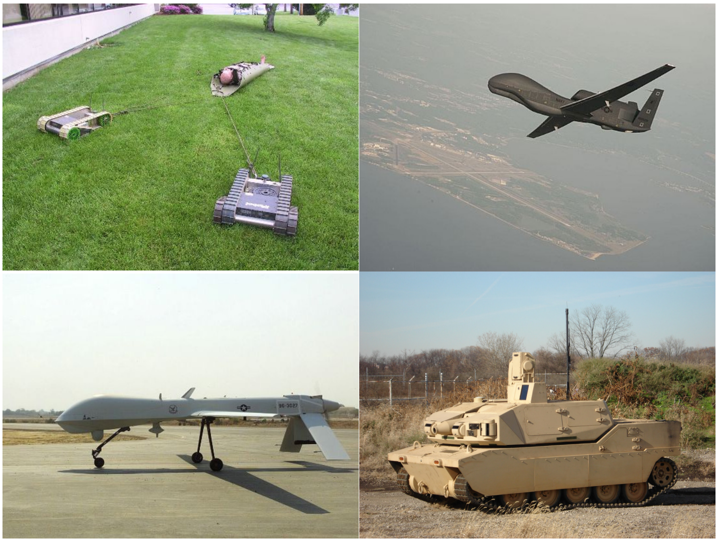


Fig. 1. Top left: iRobot Valkyrie [12], photograph retrieved with permission from http://robotfrontier.com/gallery.html. Top right: RQ-4 Global Hawk [13], photograph retrieved with permission from http://www.navy.mil/view_image.asp?id=125696. Bottom left: MQ-1 Predator [14], photograph retrieved with permission from http://www.navy.mil/view_image.asp?id=883. Bottom right: Black Knight UGCV [15], photo courtesy of the National Robotics Engineering Center. © Copyright 2007-2012, Carnegie Mellon University. All rights reserved.

3 Human-Human Interaction

3.1 Communication Process

Development of supervisory control standards requires evaluation of the human component of mixed-initiative teams in addition to robot skills and behaviors. Understanding the way humans communicate facilitates HRI. In human-human interaction the communication of a message involves three components; a sender, receiver, and a channel used to convey the message (Figure 2).

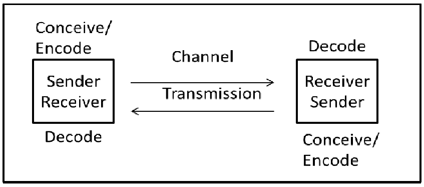


Fig. 2. Communication Process: Method used for human-human communication, the process begins with conception and ends with the receiver decoding the message, adapted from Weinschenk and Barker [16]

Communication begins with conception, when the sender creates a thought to convey to a receiver. The next step of the process, encoding, occurs when the sender considers which method to communicate the message. Transmission follows; in this step the sender selects the channel to convey the thought, ending the process of the sender [16]. The receiver then receives the message and decodes it. The process ends with feedback from the receiver verifying receipt of the message. At this time the receiver may opt to instigate a role reversal (i.e., receiver becomes sender) [16]. This process described by Weinschenk and Barker [16] for verbal communication is similar to the transactional model described by Barnlund [17], Figure 3.

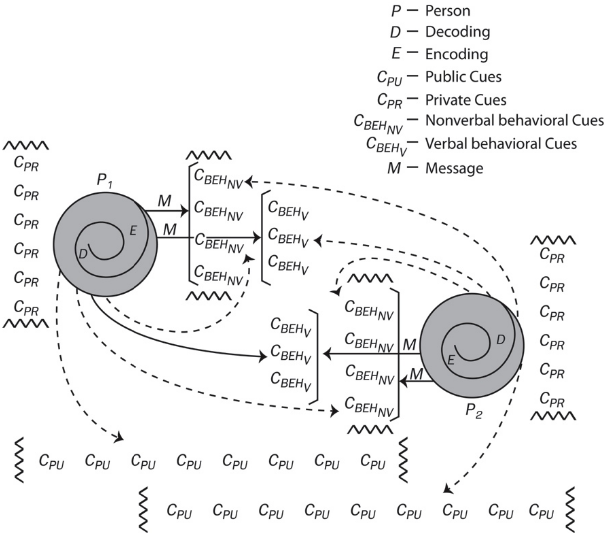


Fig. 3. Transactional Model of Communication, adapted from Barnlund [17]

The transactional model extends this process to include nonverbal behavioral cues (e.g., facial expressions, posture) and public cues (e.g., environment, culture), and thus, generalizes the model to include all explicit and implicit communication. Understanding the interaction models between humans, including communication modalities, is important in development of HRI.

Communication models demonstrate a variety of means to interact beyond speech incorporating combinations of explicit and implicit modalities. Explicit, or purposeful, communication methods consist of gesture and language. Implicit communication includes unintentional verbal and nonverbal behavioral cues and emotions. Published studies to date investigate explicit communications between robots and humans using auditory, visual, and tactile modalities. However, limited information exists in the literature related to direct application of implicit modalities for HRI [18]. Typically, the goal is to observe the robot's ability to understand the operator, acknowledge the command given, and then execute. However, research in the area Multi-Modal Communication (MMC) aims to improve communication by exchanging "information through a flexible selection of explicit and implicit modalities that enables interactions and influences behaviors, thoughts, and emotions [18]."

MMC emerges as a requirement as current and future robot systems move toward knowledge-based systems interacting with a commander, team members, and/or ambient society. As a result, the term "user" or "operator" is too limited to describe the types of people, and their roles, a robot may interact with during a mission. Following this reasoning, this effort uses the term interactor to include any person a robot interacts with. An interactor is either "active" or "passive" in their communication with a robot. An active interactor is a commander or team member able to give direction and tasking to a robot. A passive interactor is a person within society that a robot does not receive commands from, but may need to interact with through observation or other means.

By understanding robot capabilities and leveraging the study of human-human communication, HRI gains a firm foundation for developing effective interfaces. However, understanding the principles of interface and display design fills additional theoretical HRI gaps.

4 Interface Design

Since an interface serves as the bridge between a system and an interactor, the design must account for human and robot considerations. One example the HRI community can draw from traditional computer interface design is the list of Weinschenk and Barker's [16] twenty laws of interface design. Robotic system developers may derive clear interface requirements and design recommendations by adapting these laws.

The list presented in Table 1 focuses primarily on human factors considerations; however, questions related to where and how the robot will operate require additional consideration. Technical questions addressing the primary use of the robot and the type of hardware and software required arise. Understanding the physical environment (e.g., weather conditions) and the resulting effect on the interface design must be determined. Style guidelines regarding the look and feel of the interface play a role an interactor's perception and possibly performance. Ultimately, the goal is to develop an interface that fits the interactor, robot purpose, and circumstances.

Table 1. 20 Laws of Interface Design Applied to HRI adapted from Weinschenk and Barker [16]

Interface Law	Example
User Control- The interactor must think they control the system.	A speech application supporting interruption of robot operation facilitating perception of control.
Human Limitations- The interface must not overload limitations of human senses, (cognitive, visual, auditory, tactile, and/or motor).	Chunking words or grouping numbers reducing overload of short-term memory, which holds between five to n ine things.
Modal Integrity- The interface must fit the task, adapting modes of communication.	Commanding with speech and confirming with touch via pressing a button.
Accommodation- Match the interface for the interactor and the way they work.	Interface adjusts to support alternative communication modalities between normal and off-normal (covert) tasks.
Linguistic Clarity- The interface must communicate as efficiently as possible.	Interface transmits/receives appropriate terminology for communication mapped to the current context/task.
Aesthetic Integrity- The interface is designed to attract or repel interactor(s).	Using anthropomorphism to encourage interactor to engage system or deter interference from others.
Simplicity- Interface presents elements simply.	Interface facilitates natural interaction methods and common lexicon. Interface presents only necessary information without clutter.
Predictability- The interface behaves in a way such that the interactor can accurately predict what will happen next.	System executes commands consistently (e.g., system always stops when commanded).
Interpretation- The interface must anticipate what the interactor is about to do next.	When presenting a map the interface presents tools related to associated tasks (e.g., route manipulation).
Accuracy- The interface must consist of no error.	System interprets speech commands with accuracy greater than or equal to a human within multiple situations (e.g., noisy, quiet).
Technical Clarity- The interface must have the highest level of fidelity.	Visual interfaces present text and graphics clearly using appropriate fonts. Speech synthesizers articulate clearly and with appropriate dialect population.
Flexibility- The interface must have flexibility and customization capabilities for the user.	Interface employs MMC within dismounted operations. Operator control units supporting customized layouts for individuals or tasks.
Cultural Propriety- The interface must adapt to the customs and expectations of the user.	Interface prioritizes interactions based upon intra-team hierarchy. Interface interprets visual signals from different cultures correctly.
Suitable Tempo- The rate of the interface must match and become suitable to the interactor.	Interface presents information (e.g., speaks) at rate appropriate to the situational context and limitations of human perception.
Consistency- Consistency in an interface is a must.	A speech interface using “Go Forward” with “Go Back” as a corollary command rather than “Previous”.
User Support- The interface must support troubleshooting	Interface supports alternative input methods in the event of speech recognition failure and for system diagnosis. Alternatively, interface supports methods to query interactor for assistance (e.g., robot is disabled).
Precision- The interface must allow the interactor to perform a task exactly.	System responds as expected when given a command. For example, interactor requests information to the right of a target and a robot responds with results to the right of its orientation/location.
Forgiveness- The recovery of interactor actions is required.	Interface supports request for confirmation before performing unrecoverable actions.
Responsiveness- Effective responsiveness from the interface is required.	Interface provides progress indicator when performing complex actions.

Human limitations factor into what requirements a robot needs to properly communicate with its interactor. These limitations may affect the design and the type of communication modes used.

5 Human Capabilities and Limitations

A robot's level of automation depends upon the technology required to accomplish its mission or intended use. It also depends upon the technology available to achieve mission goals. However, understanding how to capitalize on the strengths, and minimize the weaknesses of a robot strikes at the core of mixed-initiative teams. In a broader sense, the term "mixed-initiative" indicates the optimization of role allocation for human and robot team members [19]. Thus, understanding HRI constraints resulting from human capabilities and limitations is necessary.

Capability constraints inherent to the human perceptual system, cognitive processing, and performance boundaries must drive the design, development, and creation of interfaces. Failure to recognize this need jeopardizes the success of interactions that will occur between a human and a robot. An average human maintains quantifiable thresholds useful for guiding the creation of HRI interfaces. We briefly describe cognitive, auditory, tactile, visual, and motor capability constraints of specific interest to this endeavor.

Cognitive aspects of note include memory, decision-making and attention. For example, chunking information serves as a common memorization technique for remembering five to nine items [16]. Research indicates that short and easy to remember commands or gestures improve communication accuracy in addition to efficiency [20]. Additionally, deficits in human decision-making capabilities suggest a need for a robot's interface to embody flexible recovery from human user mistakes. The ability to confirm tasking for high-risk commands would act as a failsafe to ensure correct comprehension of the task. In regards to attention, humans have a timesharing ability which allocates resources between two tasks. Ideal timesharing involves dividing attention between a auditory and visual input [21].

With respect to auditory capabilities, human hearing ranges from 20 to 22,000 Hz and from 0 to 130 dB (the threshold of hearing and pain) [16]. Human ears distinguish the direction of sound up to three degrees apart measured by the timing and strength of the sound each ear receives [22]. Weakness in hearing pertains to measuring distance of sound [23].

An interface that has a visual component presents challenges as well because of human capabilities. The wavelength visible to the human eye ranges from 400 to 700 nm [22]. Rods, receptors for nighttime vision, light sensitivity peaks at 500 nm, whereas cones (daytime vision) comprise of three peak sensitivities, 440, 540, and 565 nm for short, middle, and long wavelengths respectively [30]. Eyes adapt to darkness within thirty minutes of exposure [30]. For visual components, it is best to avoid overstimulation and information overload by eliminating display clutter. Font size and display size must be balanced to ensure clarity of messages and graphics displayed.

An emerging modality for robot-to-human communication is the sense of touch via tactile displays. Tactile displays transmit tactile-icons or "tactons" representing words or phrases [24] through vibro-tactile devices (tactors) typically around the abdomen [25]. Although it is still too early to determine if tactile displays are equivalent in utility to speech or visual interfaces for communication, designers must consider them in situations where other sensory modalities are overtaxed in respect to Wicken's multiple resource theory [26]. Research in advanced cueing applications with tactile displays shows no significant differences in reaction time compared to speech or 3-D Audio [27] alone, and demonstrates improved performance when combined with auditory cues [28]. Tactors on the torso also aid in navigation tasks and reduce workload [29]. Moreover, investigations into subjective workload and tactile displays in tactile cueing displays shows no significant difference in overall workload between tactile and visual cues with both showing lower workload than 3-D audio [27]. Even though the use of tactile displays may still be considered in their infancy, they may play a future role in multi-modal communication systems.

Motor acuity develops with practice. Research [31] [32] concluded humans practicing for short intervals spread out through several days learned efficiently rather than long intervals for fewer days [30]. In relation to tasks practiced, Shea and Morgan [33] concluded randomizing the order of learning the tasks resulted in better retention [30].

With regard to motor limitations, Fitts Law [16] [34] provides important quantifiable metrics for visual interfaces, it states:

$$MT = a + b * \log_2 \left(\frac{2D}{W}\right) \tag{1}$$

Where *MT* defines movement time, *D* the distance of the movement from start to center, *W* the width of the target, and *a*, *b* constants based on type of movement.

This law defines how large to make a target on a visual display, so the user can hit the target accurately. Even though a supervisory control standard would not focus on visual interfaces these constraints play an important role. Such capabilities and limitations impact the communication process between a human and a robot, and therefore are critical to setting boundaries for interface design.

6 Recommended Considerations

The topics presented above point to seven foundational recommendations that require expansion and enhancement.

1. Define the mission
2. Understand and account for the level of interaction supported by current and future robotic assets
3. Define cognitive resources required to support mission objectives and required level of interaction

4. Define the HRI communication protocols in terms of human communication processes
5. Develop an applicable MMC framework for HRI facilitating appropriate selection of communication protocols
6. Develop a set of science-based interface design standards
7. Define an HRI framework to account for human cognitive, auditory, tactile, visual and motor capabilities and limitations

7 Conclusion

Advances in robot sensor, autonomy, intelligence, and mobility are ushering in a new era of mixed-initiative teams. Communication between interactors and robots will be a critical factor in the success or failure of fielded robot platforms. The paper presents four key areas to consider when developing HRI standards focused on supervisory control: level of automation, human-human communication processes, interface design, and constraints based on human capabilities and limitations. Ultimately, the seven recommendations provided require expansion and investigation to fully realize the potential of mixed-initiative teams in operational environments.

References

1. U.S. Congress, National Defense Authorization Act for Fiscal Year 2001, U.S. Congress, Washington (2001)
2. Office of the Secretary of Defense, Unmanned Systems Roadmap: 2007-2032. U.S. Department of Defense (2007)
3. Sheridan, T.B.: Telerobotics, Automation, and Human Supervisory Control. The MIT Press, Cambridge (1992)
4. Rasmussan, J.: Outlines of a Hybrid Model of the Process Plant Operator. Monitoring Behavior and Supervisory Control 1, 371–383 (1976)
5. Ogreten, S., Lackey, S., Nicholson, D.: Recommended Roles for Uninhabited Team Members within Mixed-Initiative Combat Teams. Collaborative Technologies and Systems, 531-536 (2010)
6. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for Types and Levels of Human Interaction. IEEE Transactions on Systems 30(3), 286–297 (2000)
7. Yamauchi, B. M.: Packbot: A versatile platform for military robotics. Unmanned Ground Vehicle Technology VI 5422, (2008)
8. Baxter, J.W., Horn, G.S.: Controlling Teams of Uninhabited Air Vehicles. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, Netherlands, (2005)
9. Mehra, R. K., Boskovic, J. D., Li, S. M.: Autonomous Formation Flying of Multiple UCAVs Under Communication Failure. Position Location and Navigation Symposium, 371-378 (2000).
10. Hildebrandt, E.: RQ-4 Global Hawk. Northrop Grumman (2012)

11. Khurshid, J., Bing-rong, H.: Military robots-A Glimpse From Today and Tomorrow. In: 8th International Conference on Control, Automation, Robotics, and Vision, pp. 771–777 (2004)
12. Crow, W. D.: RQ1L Predator Unmanned Ground Vehicle, U.S. Marine Corps (2002)
13. Valois, J.S., Herman, H., Bares, J., Rice, D.P.: Remote Operation of the Black Knight Unmanned Ground Combat Vehicle, Unmanned System Technology X (2008)
14. Yamauchi, B.: Artist, The Valkyrie.(Art). I Robot (2005)
15. Ott, N.: Artist, Black Knight.(Art). National Robotics Engineering Center, 2007-2012
16. Weinschenk, S., Barker, D.T.: Designing Effective Speech Interfaces. In: Hudson, T. (ed.). John Wiley & Sons, Inc. (2000)
17. Barnlund, Interpersonal Communication: Survey and Studies. Houghton Mifflin, Boston (1986)
18. Lackey, S.J., Barber, D.J., Reinerman-Jones, L., Badler, N., Hudson, I.: Defining Next-Generation Multi-modal Communication in Human-Robot Interaction. In: Human Factors and ERgonomics Society Conference, Las Vegas (2011)
19. Hearst, M., Allen, J., Guinn, C., Horvitz, E.: Mixed-Initiative Interaction: Trends & Controversies, pp. 14–23. IEEE Intelligent Systems (1999)
20. Miller, G.A.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological Review 63, 81–97 (1956)
21. Wickens, C.D., Gordon, S.E., Liu, Y.: An Introduction to Human Factors Engineering. Addison-Wesley Educational Publishers Inc., New York (1998)
22. Bruce, V., Green, P.R., Georgeson, M.A.: Visual Perception: Physiology, Psychology, and Ecology. Psychology Press, New York (2003)
23. Smith, S.W.: The Scientist and Engineer's Guide to Digital Signal Processing. California Technical Publishing, San Diego (2011)
24. Brewster, B., Brown, L.M.: Tactons: Structured Tactile Messages for Non-Visual Information Display. In: Austrailian User Interface Conference, Dunedin, New Zealand (2004)
25. White, T.: Suitable Body Locations and Vibrotactile Cueing Types for Dismounted Soldiers. U.S. Army Research Laboratory, Aberdeen Proving Grounds, MD (2010)
26. Wickens, C.: Multiple Resources and Performance Prediction. Theoretical Issues in Ergonomics Science 3(2), 159–177 (2002)
27. Glumm, M.M., Kehring, K.L., White, T.L.: Effects of Visual and Auditory Cues About Threat Location on Target Acquisition and Attention to Auditory Communications. US Army Research Laboratory, Aberdeen Proving Ground, MD (2005)
28. Gunn, D.V., Warm, J.S., Nelson, W.T., Bolia, R.S., Schumsky, D.A., Corcoran, K.J.: Target acquisition with UAVs: Vigilance displays and advanced cueing interface. Human Factors 47(3), 488–497 (2006)
29. Van Erp, J.B.F., Werkhoven, P.: Validation of Principles for Tactile Navigation Displays. In: Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, San Francisco, CA (2006)
30. Salvendy, G.: Handbook of Human Factors and Ergonomics. John Wiley & Sons, Inc., New York (2012)
31. Lee, T.D., Genovese, E.D.: Distribution of Practice in Motor Skill Aquisition: Learning and Performance Effect Reconsidered. Research Quarterly for Exercise and Sport 59, 277–287 (1988)

32. Lee, T.D., Genovese, E.D.: Distribution of Practice in Motor Skill Aquisition: Different Effects for Discrete and Continuous Task. Research Quarterly for Exercise and Sport 60, 59–65 (1989)
33. Shea, J., Morgan, R.L.: Contextual Effect on Acquisition, Retention, and Transfer of a Motor Skill. Journal of Experimentation Psychology: Human Learning and Memory 5, 179–187 (1979)
34. Sanders, M.S., McCormick, E.J.: Human Factors in Engineering and Design. McGraw-Hill, New York (1993)

An Emotional Framework for a Real-Life Worker Simulation

Emotional Valence Scoring Inside a Workflow Enhancement Simulator

Nicholas H. Müller[1,*] and Martina Truschzinski[2,*]

[1] Institute for Media Research, Chemnitz University of Technology, Chemnitz, Germany
nicholas.mueller@phil.tu-chemnitz.de
[2] Chair of Automation Technology, Chemnitz University of Technology, Chemnitz, Germany
martina.truschzinski@etit.tu-chemnitz.de

Abstract. Within the framework of the project 'The Smart Virtual Worker' we put forward a sound and functioning emotional model which adequately simulates a worker's emotional feelings throughout a typical task in an industrial setting. We restricted the model to represent the basic emotions by Ekman and focused on the implementation of 'joy' and 'anger'. Since emotions are uniquely generated, based on the interpretation of a stimulus by an individual, we linked the genesis of emotions to empirical findings of the sports sciences to infer an emotional reaction. This paper describes the concept of the model from a theoretical and practical point of view as well as the preliminary state of implementation and upcoming steps of the project.

Keywords: emotion framework, work simulation, workflow simulator, emotional valence, emotional model.

1 Introduction

Demographic changes in Germany, as in most of the industrialized countries, will result in an aged workforce [3], [5], [7]. To avoid a shortage of skilled labor, employers have to cope with the aging population by looking for adaptive strategies regarding their workflow in order to keep already qualified employees. E.g. established workplaces will have to be modified so the needs of older, qualified but limited employees, due to their age or medical conditions, are considered. The 'Smart Virtual Worker'(SVW)-project presents an opportunity to easily replicate established workflow parameters inside a virtual simulation to establish alternative routes, storage, or construction methods during the stage of production planning. This will help to keep acquired skills inside a company which would otherwise be lost to competitors or, in case of rare and specialized manufacturers, even be removed from the general workforce due to early retirement.

[*] This work has been funded by the European Union with the European Social Fund (ESF) and by the state of Saxony.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 675–686, 2014.

A key component of the simulation is the consideration of emotional tendencies within an employee while performing a task. Since emotions are very uniquely linked to an individual, as are past experiences, lessons learned and many other aspects such as strength of mind, bodily endurance etc. [4], the challenge of any emotional model is to find a balance between unified emotional display and a generalized reasoning as to allow for a sufficient prognosis of emotions in the general population by the simulation.

1.1 Emotions in the Workplace

Scientific research regarding emotions has been widely focused on a meta-level of emotional classifications and origins [13]. To facilitate the model of this paper, a narrower look at emotional specificities is necessary. Emotions at work are as wide a field of possible research as emotions are in general, but since the goal of the SVW-project is to accurately simulate the emotional stability during a work task, it is necessary to focus on the individual and the situation of being at work. Psychological research in conjunction with a work environment is different from other forms of academic psychological research in general [14]. We subjugate our personal aspirations, as far as possible, during work hours to the needs of a company or a task. Hence there might be a wide array of psychological problems originating from work but a single task within ones measures of capability would result in a fairly steady emotional state. Especially since work-tasks are per se pretty much unemotional, apart from anecdotal reports. The typical work task is structured to be feasible and a worker will handle it while being balanced, although tending to either liking the task at hand or disliking it.

1.2 Robotic Applications

Since the SVW-project contains multiple individual modules, the emotional model is built as a standalone solution. This allows for a much wider array of possible implementation strategies like being an add-on component for already established system architectures. The input from a motion generation module and the integration of the output by any form of artificial intelligence or path planning module are basically the only adjustments necessary while the input from ergonomics can either be disabled or easily replaced by, for example, a movement limitation system of the robot. Furthermore, the ability for understanding an intention is closely linked to an emotional understanding [18]. With this in mind, an emotional action unit like the one presented here could serve as a basis for a system of social-cognitive reasoning, which in turn could ease the co-existence of robots as a household-help and human subjects, regarding the legibility and predictability of a robot's intended action [16]. It would facilitate Human-Computer interactions, e.g. by mimicking emotions like robot pets (Sony Aibo, Tamagotchi etc.) do, or signal a user about the internal working state of the machine by either displaying it by writing it on a display, flashing a light or by producing an emotional sound like low-pitched beeps, which already notifies users today about their rundown battery inside their cell-phones. Although the neurobiological system of a human is neither easily copied onto a machine nor would it be a very practicable approach, some beneficial approaches are available to have a machine 'compute' an emotion without being cognitive aware about it [17]. In addition, there are emotions one possibly would not want to install upon a machine. Robotic 'emotions' should be perceived as a useful new information channel which is beneficial for the suggested role the robot was built to perform. An angry Roomba or a sad car manufacturing robot is most probably not anyone's goal.

1.3 Computational Models

Research in the field of computational models led to a vast number of published models. A good overview is presented in the meta-analysis section of [15] differentiating between emotional models having a rational-, anatomical-, dimensional- or appraisal-related approach. But since the main focus of these models is to adequately resemble the general emotional system of human beings, an adaptation to the SVW-project and its narrow focus on emotional reactions to work-tasks seemed not to be feasible. For instance most of the time the motivational part of many models is quite irrelevant for the worker's emotional state since the motivational goal of a worker is to complete a task in order to get paid, not to fulfill an intrinsic need [14].

2 Elements of the Model

To allow the model to work two types of numerical input are necessary (see Fig.1). First the planned action from the reinforcement learning algorithm for motion generation suggests a work task which has already been assessed regarding its possibility by the actuator module, which impacts the emotional model in three ways. The actuator assessment is characterized as being either feasible, being precarious or alarming.

Since emotions heavily depend on an individual's singular response to outer circumstances and the simulation aspires to be able to make individual recommendations for a better workflow, the emotional model contains an individualizing computation routine. Depending on three assessed factors about the physiology of the worker in question, the impact on the valence scales is adapted. These three factors are constitution, sensitivity, and experience, shortened in the model as C, S and E. We chose those three factors because they represent a feasible reduction of human physiological peculiarities. Constitution allows for an intuitive assessment of the agents strength and endurance, whereas the sensibility works as a damping variable which permits assessing the resilience when confronted with obstacles. The experience is again used as a damping variable to enable the model to, for example, heighten the resilience, since the agent simply 'knows' it has to keep the weight up just for a little longer to accomplish the work task successfully. In addition the model tracks the fatigue of the agent and henceforth the probability of a successful operation, which decreases over time and with growing exhaustion.

Regarding the emotional state, the model right now is focused on a pure valence-based emotional distinction, meaning for the simulated agent to either like the current emotional state or to dislike it. These two states are accordingly labeled as 'joy' and ‘anger’, but please do not think of these valences as being directly linked to these emotions. They are a continuum in which the emotional state tends to being joyful or leans into the other direction. So if asked, a person would report to be feeling okay or not, as emotions are not computed with a distinct level of something. In addition the model integrates another scale which mirrors a sympathetic arousal, which is used as a form of energy the emotional state of the current action invokes in the agent. In addition, following the theory of an emotional transaction by Zillmann [8], [9], [10], this sympathetic arousal enables the transfer of energy between the, currently two, emotional states implemented.

As a result, the emotional model outputs three variables. The artificial intelligence module receives information about the chances for exhaustion and the current emotional valence state, which is either positive or negative, and secondly, the assumed time requirement for a successful in-time work completion is handed over. The reinforcement learning algorithm will then score the emotional output and repeat the procedure by initiating another planned action into the module.

2.1 The Emotional Model - State of Implementation

The emotional model is implemented within the framework of the SVW-project. It is supplied with input from both the ergonomic and motion generation module, while the latter gets its input from the artificial intelligence module. The artificial intelligence module itself is responsible for planning a work process and requests an evaluation of a particular task-related element. The implemented planning is based on a hierarchical reinforcement learning algorithm, which chooses the next possible action based on the paramount emotional and ergonomic evaluation.

2.2 The Agent

Our agent, who simulates the human worker, includes attributes such as sex, weight, height and a resulting BMI score. Furthermore we differentiate between fitness-levels (the worker being either well-trained, considered to be of normal strength or weak), the work-experience, age and a score for sensitivity. These attributes define the required internal state and allow for an upcoming calculation of its unique emotional valence.

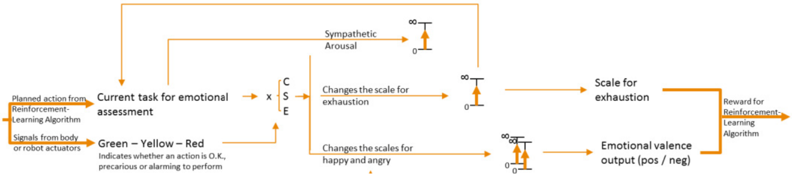

Planned action from Reinforcement-Learning Algorithm
Signals from body or robot actuators
Current task for emotional assessment
Green – Yellow – Red
Indicates whether an action is O.K., precarious or alarming to perform
x
C
S
E
Sympathetic Arousal
∞
0
Changes the scale for exhaustion
∞
0
Changes the scales for happy and angry
∞
∞
0
0
Scale for exhaustion
Emotional valence output (pos / neg)
Reward for Reinforcement-Learning Algorithm

Fig. 1. Emotional model of the ‘Smart Virtual Worker’

1. *The physiological attributes:* Based on the attributes for weight, height, BMI score and fitness, the athleticism of the agent is calcuated. This value enables us to compute how well trained the agent is and as a result, how heavy an object might be to allow even a weak worker to perform. The differences between a weak, a normal and an athletic worker are calculated as a resulting value of capability. So in the model an athletic worker has a capability value of 1.2, which means he is 20% stronger than an assumed normal worker. A very weak worker has a capability value of 0.9, meaning he is 10% weaker than our assumed normal worker (capability of 1.0).
2. Experiences: We assume that over time a worker gains knowledge about the tasks performed, leading to an experience value which defines his familiarity with the task at hand. For example, since he knows from past experiences that the current task includes heavy lifting, but only for a short time, he 'clenches his teeth' and endures this brief moment, compared to an inexperienced worker who might quit the task altogether. Within the model the experience value is set between 0.8 and 1.2, analogous to the fitness calculation. A very inexperienced worker is scored with a value of 0.8, a normal worker with 1.0 and a very experienced worker with a value of 1.2.
3. Sensitivity: The sensitivity value defines how much the worker is affected by any given situation. A worker with a high sensitivity reacts more intensely while a worker with a lower score of sensitivity does not. The model computes the sensitivity value being between 0.8 and 1.2. In this case a very thick-skinned worker has a sensitivity value of 0.8, a normal worker of 1.0 and a very sensitive worker is scored with a value of 1.2.

2.3 The Internal Emotional State

The internal emotional state is calculated on the basis of the described psychological and physiological values. This computation leads to a value which represents the personal appraisal of doing the work and a temporary estimation of the quality of work.

4. The input parameter: The computation of the emotional state is based on all of the described input values and objects within the simulation. The first input is the currently performed action including physical properties, for example: the involved objects and their weight (in kg), the time (in seconds) how long the action to be performed will take, and an ergonomic value. All described variables are either available from the database from the start or will be calculated just in-time from other modules of the project, as e.g. the ergonomic scoring. The ergonomic value ranks the physiological stress of the body. At the moment, this calculated value is based on the established RULA-system [11].
5. Level of activity: In order to compute the emotional valence, our module calculates a level of activity for the currently performed action. This computation is based on the evaluated guidelines of 'BGI 582: Safety and health requirements for transport and storage tasks' by the association 'Vereinigung der Metall-Berufsgenossenschaften' [12].

The output consists of four distinct levels of activity based on the values: carried weight, covered distance, action time, and a separate ergonomic value:

- 1: low level of activity, no handicap, and no overloading
- 2: increased level of activity, impairment by weaker persons is possible
- 3: more increased level of activity, impairment, and overloading of normal persons is possible
- 4: overloading of normal people

Based on these levels of activity, the internal basis for emotional valence now calculates the values for sympathetic arousal, joy, and anger.

6. Sympathetic arousal: The arousal (a) is influenced by the level of activity (l), the current exhaustion value (x_{act}), the experience (p), the rate of arousal adaption (t_a), the time for recovery (t_r) and a moderating value (r). This value lowers the score of sympathetic arousal over time, so if nothing stimulating is happening, the worker's arousal level might even go down to 0, resulting in a sympathetic stability. Also the moderating value cannot be too high, since an emotional state is able to influence another subsequent emotional state [9]. In this case the arousal is calculated as follows:

$$a_i = a_{i-1} + (d_x + d_a(l)) \cdot t_a \cdot p - t_r \cdot r \tag{1}$$

$$d_x = \left(100 - \frac{x_{act}}{100}\right) \cdot d_a(l) \cdot t_a \tag{2}$$

The value d_a is an array which defines the changes of the arousal depending on the level of activity (1). Within the model, the array d_a consists of four states ($d_a = \{0.25, 0.5, 0.75, 1.0\}$). This means, if the level of activity is i, the change of arousal is $d_a[i]$. For example, the change of arousal is 0.25 if the level of activity is 1. The change of arousal is increased by higher activity levels, while our parameters t_a and t_r are based on the length of an action ($t_a = \frac{t_{action}}{60}, t_r = \frac{t_{action}}{30}$). The value of recovery (r) is set to 0.04. Depending on the exhaustion, the term (d_x) increases the arousal, if the activity level is 2 or higher.

7. The values of emotion: The calculated emotional valence (e) is labeled as either joy or anger and is influenced by the level of activity, the level of exhaustion (x_l), the length of activity (t_{action}) and the sensitivity (s) of the agent. In this case the emotional value is calculated as follows:

$$e_{act_i} = e_{act_{i-1}} + (d_{ex}(x_l) + 1) \cdot d_e(l) \cdot s \cdot t_{action} \tag{3}$$

$$e_{act} = \begin{cases} e_{joy}: if\ l = 1 \\ e_{anger}: if\ l = 3,4 \end{cases} \tag{4}$$

$$d_{ex} = \begin{cases} 0: if\ \ x_l = 1 \\ \frac{x_{act}}{2}: if\ x_l = 2 \\ x_{act}: if\ \ x_l = 3,4 \end{cases} \tag{5}$$

Both emotions are scored independently from the other, so if the level of activity is 1.0, the value of joy is increased. If the level of activity is 3.0 or higher, the value of anger is increased

(4). Actions which have an activity value of 2.0 are considered emotionally steady. The change of emotion is defined by the array $d_e = \{0.5, 0.25, 0.75, 1.0\}$. Exhaustion leads to an increase in anger if an action has the activity level 3 or 4

(5). The level of exhaustion (x_l) is dependent on the strength of the worker. We divide strength into four categories: the area of recuperation $(x_l = 1)$, normal workload $(x_l = 2)$, short-time maximum strength $(x_l = 3)$ and evolutionary emergency reserve $(x_l = 4)$ (see Fig. 2). A worker can only perform inside the range from recuperation to short-time maximum strength. The boundaries of these areas depend on the constitution of the agent. A stronger agent has higher thresholds than a normal agent and a normal agent has again higher thresholds than a weaker agent. These limits are currently fixed for the chosen agent model.

8. The value of exhaustion: This value (x) is influenced by the time of activity (t_{action}), the level of exhaustion (x_l), the calculated level of activity (l) and the corresponding score of the recovery parameter. The current exhaustion of an agent is calculated as follows:

$$x_i = x_{i-1} + d_x(l) \cdot x_l \cdot t_x - t_r \cdot r_x \tag{6}$$

Like the sympathetic arousal the exhaustion has a moderating mechanism as well (6). The time of recovery (t_r) and the moderating value (r_x) define how much time an agent needs to rest. If there is no strain on the worker, the moderating mechanism decreases the value of exhaustion down to 0. The array d_x consists of the four states

0.35, 0.45, 0.65 and 0.75. The time rates depend on the length of an activity ($t_x = \frac{t_{action}}{60}, t_r = \frac{t_{action}}{30}$) and the moderating value (r_x) is set to 0.03.

9. The output parameter: Currently the output of our model is the emotional valence of an action which is dependent on the dominating emotion and the level of arousal. The dominating emotion is e_{joy}, if $e_{joy} > e_{anger}$. If $e_{anger} > e_{joy}$ then the resulting dominating emotion is e_{anger}. Otherwise no emotion is considered to be dominant. The emotional valence is positive if the dominating emotion is joy, and negative if the dominating emotion is anger. If no dominant emotion exists, the valence is zero. The value of the emotional valence's change is defined by the calculated arousal of an action.

Demonstration of the Emotional Model.
We defined three typical workers to demonstrate our model:

- A weak worker (male, 60 kg, 1.85m, BMI=17.5, 30 years, capability value =0.9, e=1.0, s=1.0)
- A normal worker (male, 80 kg, 1.85 m, BMI=23, 30 years, capability value =1, e=1.0, s=1.0)
- An athletic worker (male, 90 kg, 1.85 m, BMI=26, 30 years, capability value =1.2, e=1.0, s=1.0)

Also, we defined different tasks to demonstrate the model. The global task is carrying 20 boxes from one point to another. One episode of carrying one box consists of four elementary behaviors: 'walk to the box', 'grab the box with both hands', 'walk with the box to a target' and 'release the box'. Every action has its own parameter such as length of time, object involved and a corresponding ergonomic value. The involved object has the attributes *weight* and *dimension* which are provided by the database of the SVW-project. To demonstrate different stresses and strains, three different boxes with 15 kg, 20 kg, and 30 kg are simulated.

On the basis of the input parameters which are provided by other modules or by the database, the internal state of a simulated worker while performing is calculated. The following images demonstrate the model's output with different agent types carrying the different boxes.

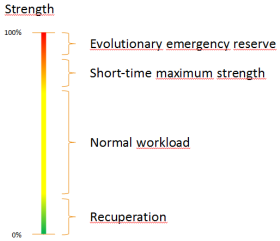


Fig. 2. The strength of an agent is divided into four categories. The limits of these parts depend on the constitution of the agent.

Fig. 3 shows the calculated emotional valence value for an athletic (blue line), normal (red line), and weak worker (green line) executing the task process with a 20 kg box. The different worker types show very different emotional reactions during their work. At first, the athletic worker carries the boxes with ease, which is why its emotional valence value does not change to a negative value before the 16th box (grey lines). At this point the curve progression shows a positive valence value of 11.06 at ‘walk to the box’ while ‘grab’ and ‘carry the box’ result in a negative valence value of -11.38 or -11.81, respectively. This is the result of the exhaustion, which increases the anger factor by a higher increment that the factor joy as increased by the previously easy work. In the next step, if the worker is walking without any load, the value of joy is increased so much that the emotional valence value will be positive again. But during the upcoming episode, the exhaustion level rises beyond the agent threshold and the worker carries the following boxes with a negative emotional reaction.

In comparison, the normal worker carries the first box with different emotional valence values. The elementary behaviors ‘walk to the box’ and ‘release the box’ are easy and increase its value of joy. Our normal worker ‘grabs the box with both hands’ and ‘carries the box’ with a negative valence because for him the weight of the box is laborious. Additionally, after two boxes the worker is so affected by the weights that the recovery period ‘walk to the box’ is insufficient to retrieve a positive valence.

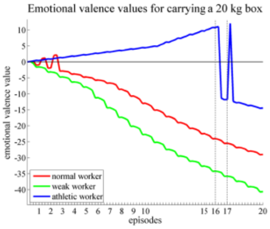


Fig. 3. The emotional valence value for carrying normal and weak worker

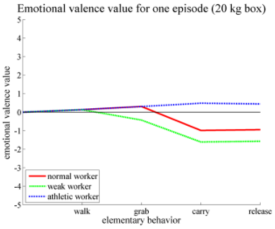


Fig. 4. The first calculated steps for carrying a 20 kg box. One episode consists of the elementary behaviors: walk, grab, carry and release.

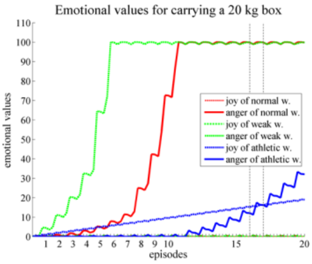


Fig. 5. Emotional values of joy and anger during the task processed by the chosen agents

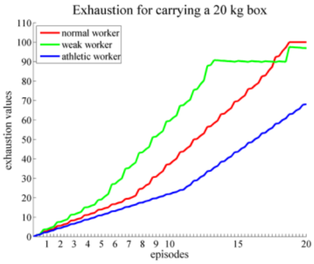


Fig. 6. The exhaustion for the different agent types based on carrying a 20 kg box

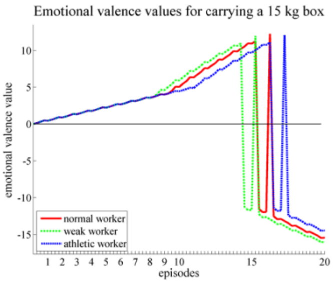


Fig. 7. Emotional valence of different workers for carrying a 15 kg box

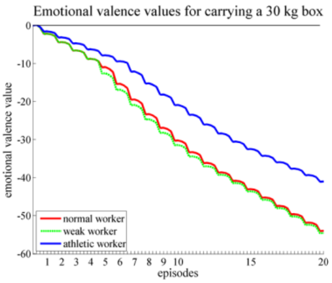


Fig. 8. Emotional valence of different workers for carrying a 30 kg box

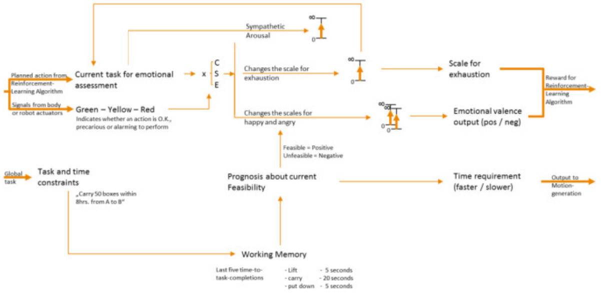

Sympathetic
Arousal
∞
0
Planned action from Reinforcement-Learning Algorithm
Signals from body or robot actuators
Current task for emotional assessment
x
C
S
E
Green – Yellow – Red
Indicates whether an action is O.K., precarious or alarming to perform
Changes the scale for exhaustion
Changes the scales for happy and angry
Scale for exhaustion
Emotional valence output (pos / neg)
Reward for Reinforcement-Learning Algorithm
Feasible = Positive
Unfeasible = Negative
Global task
Task and time constraints
„Carry 50 boxes within 8hrs. from A to B"
Prognosis about current Feasibility
Time requirement (faster / slower)
Output to Motion-generation
Working Memory
Last five time-to-task-completions
- Lift - 5 seconds
- carry - 20 seconds
- put down - 5 seconds

Fig. 9. Enhanced emotional model of the "Smart Virtual Worker"

Furthermore, the worker will carry the next boxes with a negative emotional reaction. All while our weak worker carries all 20 kg boxes with a continuously negative valence, since with his values the boxes are too heavy.

Fig.4 clarifies the different reactions for carrying a box. All workers start with an emotional valence value of zero. After ‘walk to the box’, all of them have a small positive valence value, since this is not a strenuous activity. Following the task ‘grab the box’, our weak worker has a negative valence value because the box is too heavy for him. At this point the normal and athletic workers react with a positive emotional valence but only the athletic worker with his physical strength has the attributes to carry the box easily. He is the only one still performing with a positive valence during the next elementary behavior ‘walk with the box to a target’.

In Fig. 5, it is demonstrated why the emotional valence is switched. As described above, the dominating emotion is the one with the highest value. The weak worker shows a strong rise in the emotional values of anger (dark blue) and only a small rise of joy (light blue). This results in anger being the dominating emotion for the following episodes. The normal worker has a small slope of anger (dark red) and joy (light red). These small slopes enable the model to switch between the two dominating emotions. Since both values are very similar, an emotional switch due to external influences is quite easy to facilitate. For the emotional values of the athletic worker the slope of joy (light green) is high at first but with rising exhaustion levels (see Fig. 6) the anger (dark green) increases. Consequently, the dominating emotion in episode 16 is switched (grey lines). The Fig. 7 shows our three workers carrying a 15 kg box. At first, all simulated workers carry the weight of this box quite easily. But similar to carrying a 20 kg box, the workers are exhausted by the process of carrying. The weak worker is exhausted faster than the normal worker and the normal worker is exhausted faster than the athletic worker. For this reason the dominating emotion is switched for the weak worker at the 15th, for the normal worker at the 16th and for the athletic worker at the 17th box. If the carried weight of the box is 30 kg, all workers have a negative emotional valence while performing the task (see Fig. 8). There is only a very narrow difference between the worker types.

3 Conclusion

The model calculates the emotional valence of different workers performing a predefined task. We demonstrated that the emotional valence and therewith the interpretation of the actual situation depends on the attributes of the machine and the number of repetitions. We have shown that the model can simulate different attributes. In addition, we showed that an elementary behavior like ‘joy’ and ‘anger’ is interpreted differently, depending on the number of boxes carried before.

Our next steps will be to further implement the depicted model in Fig. 1 while attempting to evaluate our preliminary predictions about the individual emotional responses within a real-world scientific experiment by implementing the model into a robot. Furthermore we want to expand the model to include a form of memory-system which would allow for a planning routine (see Fig. 9). With this addition the model opens up a new computing path which is dependent on the global task of the robot. If the task would be to carry 5 boxes from point A to point B, the module is able to track the necessary time for one iteration and hence compute the remaining time for the other four boxes. If however the task is to be completed in less time than anticipated, the robot could either try to compute a different / faster route, accelerate its movements or, if the time-limit is not feasible, report this discrepancy back to the user. This would happen by putting the robot into a less favorable emotional state, which is easily understood by the human operator.

References

1. Gross, J.J.: The Emerging Field of Emotion Regulation: An Integrative Review. Review of General Psychology, 271–299 (1998)
2. McKinnon, R.: An ageing workforce and strategic human resource management: Staffing challenges for social security administrations. International Social Security Review, 91–113 (2010)
3. Rauschenbach, C., Hertel, G.: Age differences in strain and emotional reactivity to stressors in professional careers. Stress & Health: Journal of the International Society for the Investigation of Stress, 48–60 (2011)
4. Reisenzein, R., Meyer, W.-U., Schützwohl, A.: Einführung in die Emotionspsychologie. Verlag Hans Huber, Bern (2003)

5. Schneider, N., Schreiber, S., Wilkes, J., Grandt, M., Schlick, C.M.: Stress & Health: Journal of the International Society for the Investigation of Stress. Behaviour & Information Technology, 319–324 (2008)
6. Smith, C.A., Lazarus, R.S.: Emotion and Adaptation. In: Pervin, L.A. (ed.) Handbook of Personality: Theroy and Research, pp. 609–637. Guilford, New York (1990)
7. WHO. Working together for Health. WHO Press, Geneva (2006)
8. Zillmann, D.: Excitation Transfer in Communication-Mediated Aggressive Behavior. Journal of Experimental Social Psychology, 419–434 (1971)
9. Zillmann, D.: Attribution of Apparent Arousal and Proficiency of Recovery from Sympathetic Activation Affecting Excitation Transfer to Aggressive Behavior. Journal of Experimental Social Psychology, 503–515 (1974)
10. Zillmann, D.: Emotionspsychologische Grundlagen. In: Mangold, R., Vorderer, P., Bente, G. (eds.) Lehrbuch der Medienpsychologie, pp. 101–128. Hogrefe, Göttingen (2004)
11. McAtamney, L., Corlett, E.N.: RULA: A survey method for the investigation of world-related upper limb disorders. Applied Ergonomics 24(2), 91–99 (1993)
12. Frener, P.: Sicherheit und Gesundheitsschutz bei Transport- (2008)
13. und Lagerarbeiten. In: Vereinigung der Metall-Berufsgenossenschaften (Eds.) BGI 582
14. Euler, H.A.: Evolutionäre Psychologie. In: Brandstätter, V., Otto, J.H. (eds.) Handbuch der Allgemeinen Psychologie, pp. 405–421. Hogrefe, Göttingen (2009)
15. Hacker, W.: Allgemeine Arbeitspsychologie. Hogrefe, Bern (2005)
16. Marsella, S., Gratch, J., Petta, P.: Computational Models of Emotion. In: Scherer, K.R., Benziger, T., Roesch, E. (eds.) A Blueprint for an Affectively Competent Agent. Oxford University Press, Oxford (2008)
17. Dragan, A.D., Lee, K.: C. T., Srinivasa, S. S (2013). Legibility and Predictability of Robot Motion. Human-Robot Interaction (March 2013)
18. Arbib, M.A., Fellous, J.-M.: Emotions: from brain to robot. TRENDS in Cognitive Sciences 8(12), 554–561 (2004)
19. Olsson, A., Ochsner, K.N.: The role of social cognition in emotion. TRENDS in Cognitive Sciences 12(2), 65–71 (2007)

Behavioral Persona for Human-Robot Interaction: A Study Based on Pet Robot

Thiago Freitas dos Santos[1], Danilo Gouveia de Castro[1],
Andrey Araujo Masiero[1,2], and Plinio Thomaz Aquino Junior[1]

[1] Centro Universitário da FEI – Fundação Educacional Inaciana Pe. Sabóia de Medeiros,
São Paulo, Brazil
[2] Universidade Metodista de São Paulo, Brazil
{thiagosantos38,d.gouveiacastro,andreymasiero}@gmail.com,
plinio.aquino@fei.edu.br

Abstract. With the advancement of technology robots have become more common in every day applications, like Paro and GOSTAI Jazz for health care or Pleo and Genibo for entertainment. Since these robots are designed to constantly interact with people, during the development process it should be considered how people would feel and behave when they interact with those artifacts. However there might be some issues in collecting this type of data or how to efficiently use it in the development of new features. In this study we report a process for creating Personas that will help in the design of subject-focused applications for robots interactions.

Keywords: User modeling and profiling, Human-Robot Interaction, Personas.

1 Introduction

Human-Robot Interaction (HRI) is a subfield of Human-Computer Interaction (HCI). HRI studies how people behave while interacting with robots and it tries to extract the best result from that. Beside of how well a robot can help a person or how easy it can be used to accomplish a task, it should be considered how that person will react while interacting with it. According to Young et al. [1] the way people interact with robots is very unique and different from their interaction with other technologies and artifacts since robots provoke emotionally charged interactions. Our goal was to address these emotions and the way people behave when they interact with a pet robot in the creation process of new applications.

But there is a problem to make the information about the costumers' profiles, expectations and preferences useful to the development team. The adopted solution was to create Personas which are characters that represent a group of subjects (people that will interact with the robot) based on their characteristics. Those characters help the development process since the team can base on their costumers preferences instead of their own. Some of the methods used for gathering data to create the subjects' profile include: interviews; capturing the people's action while using the system; applying questionnaires.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 687–696, 2014.

Thus, in this study we focus to present the methodological approach for creating Personas to be used in design of new features for robots. In this process we conducted tests with users using a methodological approach based on Koay et.al [2] to collect data. This was obtained from questionnaires, video analysis and a real time feedback given by the participant through a device called Comfort Level Device. For the tests we used the pet robot Sony AIBO ERS-7 also aiming to see how participants would react to it because of its resemblance with a real dog. The results were Personas that address people's personality and their expectations and reactions towards the robot we used, which can be of benefit for the development of new robots' features focusing on the subject. Also we present some analysis about people's behavior relating to this AIBO in comparison with other robots of a different type which were used in Koay et.al [2] study.

This paper goes first with an explanation of Personas and how it has been used on the HRI field (Section 2). Then we begin to explain the process of creation of the Personas starting from the data collection, detailing all the components and techniques that were used (Section 3), after we present the tests with users (Section 4). After that we explain how the obtained results were used with the cluster algorithm Q-SIM and present one of created Personas as an example (Section 5). In the end we discuss the observations on the participants' behavior in comparison with Koay et.al [2] study and we talk about how this study can be helpful for future studies (Section 6).

2 Personas

In psychology, Jung [3] defined Personas as people capability to assume different behaviors depends on scenario or situation at the moment. Cooper et al. [4] faced a problem during Human-Computer Interaction (HCI) projects, which is how to attempt all user diversification on it. Due to that, Cooper adapted Jung's Personas concept to HCI and redefined Personas as hypothetic archetypes of user. This means that each Persona can represent a group of real users. That definition helps designers to reach a biggest number of real users analyzing just a few profiles. Other works specify Personas as fictitious characters once it contains information like a real user as picture; name; demographic and behavior and preference information format like a bio description [5], [6]. Personas have been applied in many HCI project since Cooper with focus on better user experience than before. This entire appliance occurs due to the easy communication about Personas needs between designers. Because of this, some works have been developed it also into Human-Robot Interaction (HRI) with aim to improve robots behaviors during interaction.

However, many HRI researches have been exploring Robot Personas that change the focus. Robot Personas are robots, which assume some profiles designed to get direction between interactions with people. It works like a mental model for robots [7-9]. This kind of approach is interesting, although it is not completely a user-centered approach. It helps to improve the robot interaction, but not considered the user behaviors and the feeling of them about to interact with robots directly. To really keep a great interaction between robots and humans we need to attempt not only for the robot personality, but also for human personality and how these different personalities interact with each other. So, to complete the cycle of interactions between robots and people considering the focus on people, we need to create also People's Personas and analyze how these Personas interact with Robot Personas or just with a specific robot. With this approach in mind, this paper presents an adapted methodology for creating Personas from HCI to HRI [6], [10]. It will help to create more social robots centered on subject.

3 Methodological Approach

From the tests with people until the definition of the personas this study followed a sequence of events illustrated in figure 1. The first step is to conduct user tests to collect data as presented in the section 4. All the data obtained from cameras, CLD and pre/post questionnaires are stored for the post analysis. After that the data from the participants is grouped using the Q-SIM algorithm. With the groups defined the researchers analyze the stored data to identify characteristics of the Personas. With the analyzed data researchers are able to address participants' psychological traits and how their behavior when interacting with the robot to the Personas.

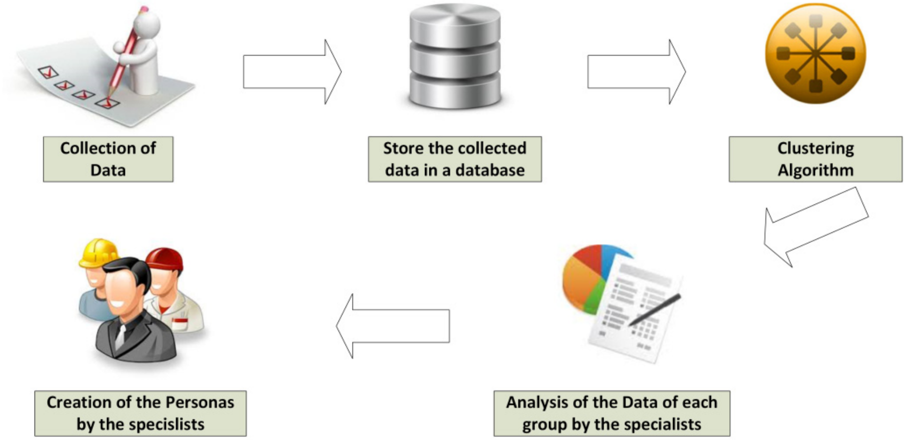

Collection of Data
Store the collected data in a database
Clustering Algorithm
Creation of the Personas by the specislists
Analysis of the Data of each group by the specialists

Fig. 1. Illustration of the five steps of creation process

As mentioned the methodological approach for data collection was based on Koay's study [2] and that was because it would provide researchers the means to collect the require data to create Personas and better take advantage of it. This data was collected from four different sources: Pre-Questionnaire, with questions about the participant's personality (the big five technique), age, genre and previous experience with robots; Comfort Level Device, an application running on a smartphone that participants used during the test to inform if they were comfortable or not; Interaction recorded video, that enabled to see the participants' reactions and what happened at the times that AIBO (which was been controlled by one of the researchers) let them uncomfortable; Post-Questionnaire, with questions about how was the experience of the interactions and in which tasks were more comfortable. The following is a description of the techniques and tools that were used during the tests and creation of the Personas.

3.1 The Wizard-of-Oz

During the tests AIBO was controlled by one of the team members using the Wizard-of-Oz method. This technique can be used to simulate functions and behavior of a robot. Therefore is common used by researchers to test the viability of a system to be implemented and also at studies centered on human behavior (which is the application for this study) [11]. A person plays the role of the "wizard" by remotely controlling the robot while the participants of the test interact with it. It's important to establish a set of actions for the wizard to perform and the person practices it so the interactions fell more naturally. In our study we used the AIBO Entertainment Player software to control the robot and auxiliary camera to give the wizard a better visualization of the environment. Through the Entertainment Player the wizard could control AIBO's movements and have access to its camera, speaker and microphone. During the interactions AIBO was controlled to behave like a dog by responding to commands (i.e. sit, stand up, catch, come here), barking, and perform a dance while playing a music which is a robot like action.

3.2 The Big-Five Technique

One of the parameters that we used to create personas was the participants' psychological traits, and to obtain these we used a tool called Big Five, that is according with [12] "a hierarchical model of personality traits with five broad factors, which represent personality at the broadest level of abstraction". The reason why we choose this tool is because the Big Five framework is the most widely used and extensively researched model of personality by the community and has a considerable support [12]. Besides [13] says that this theory of personality can also be used as a framework to describe and design the personality of products and in particular of robots.

The data used to classify the participant's personality was obtained in the first part of the test, where they had to fill a questionnaire. We used questions from the Big Five which measures five dimensions of people's personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness to Experience. It was used the TIPI (Ten-Item Personality Inventory) as the instrument to collect these data and it contains ten questions about the participants' personality, where the questions used a Likert-scale ranging from one to seven. The TIPI was adopted because it was quickly to answer, so the participant didn't fell bored before the interaction with AIBO and [12] suggest that these very brief instruments can stand as reasonable proxies for longer models (240-item for example, that takes about 45 minutes to be completed).

3.3 Comfort Level Device

To capture the participant's comfort level while interacting with AIBO we used an adaptation of the Comfort Level Device (CLD) that was used in Koay et al. [14]. Our CLD was an application for smartphone which allowed the participant to inform if he or she was or wasn't comfortable during the interaction. It had three buttons: happy face; unhappy face; end task. The button with a happy face meant that the participant was comfortable and the one with the sad face that wasn't. The button at the top of the screen meant that the participant had finished the present task so we could keep control of the comfortable recordings for each task. This information was displayed to the researcher that was operating AIBO and recorded. Before the interaction started the researcher that was conducting the study entered the participant's control number and explained how to use the application.

3.4 Data Clustering

To discover patterns into a database many researches have been use a technique called Data Clustering. This technique works in a simple way, it tries to group information based on similarity rules. Usually, the similarity rule used is the Euclidean distance, but it can be choose others similarity measure [15]. Once Data Clustering is used as a manner to discover groups with similarity, we can use it to help on creation process of Personas, grouping the most similarity user profiles. Many works have been use Data Clustering as a way to identify user profile for HCI projects [16]. Especially in Personas works, some researches use k-means algorithm to help on this process. However, k-means has a problem for creating Personas. Designers not even know how many groups exist into a dataset with user profile information and this is essential information to execute k-means, once it needs to be informed how many groups the designer wants [5], [6], [10].

To solve the problem of k-means in Personas creation process and user profile analysis, Masiero et al. [10] presents a new algorithm for Data Clustering. It calls QSIM (Quality Similarity Clustering). QSIM finds groups in a different manner. Designer informs the minimal desire similarity between element groups. QSIM uses a concept called Related Set to find groups; this concept is disseminated on Case-Reasoning Based studies. In the first results presented, QSIM demonstrated an algorithm with better results for user modeling, at least, than k-means, DBSCAN and Affinity Propagation [10]. Because of this, QSIM was adopted as the main algorithm to guide the methodology of Personas HRI creation presented at this paper. The next section will present the methodology with more details.

4 Tests for Data Collection

The studies were conducted in a laboratory at the university Centro Universitário da FEI; figure 2 explains the settings of the environment.

The participants were students, employers and visitants from an open event that was held at the university. There was a total of 39 participants, 10 children with age ranging from 4 to 12 years old and 29 adults with age ranging from 15 to 43 (from these there were 16 men and 13 women). Each test went through the following sequence of events.

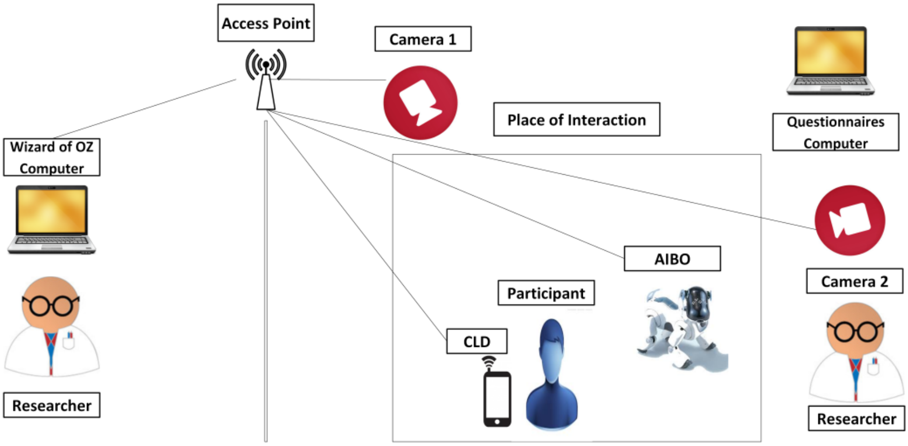


Fig. 2. Environment settings for the user tests

First there was a greeting, where the examiner explained the objectives and procedure of the study to the participant. After giving its consent for the test, the participant answered to a pre-questionnaire, which had the purpose of knowing his or her expectations about interacting with AIBO, profile and personality (the ten questions from the big five technique).

Second the participant was introduced to the CLD and the examiner explained what tasks would be done during the interaction. Before starting each task the participant read its description in loud voice. There were a total of 6 tasks divided in two groups of 3 tasks: no interaction, where the participants didn't give any instruction to the robot; physical interaction, were they had to touch the robot to make it execute the task; voice interaction, when they had to give a voice command to the robot. The first group was tagged as Human in Control (HiC) and the second as Robot in Control (RiC). During the HiC tasks if the participant felt uncomfortable with AIBO it would not move any closer, but during the RiC it wouldn't stop AIBO from getting closer. After the explanation the participant interacted with AIBO performing the tasks listed below:

First Task (No Interaction, HiC) – During this task there were no interaction between AIBO and the participant. The participant just watched AIBO walk by it, and go to the evaluator to get the bone. Second Task (Physical Interaction, HiC) – In this task, the participant waited for AIBO to get close with the bone in its mouth, and the participant had to cuddle the pet robot (in the head or back), so the robot opened its mouth and released the bone for the participant, after that AIBO walked away. Third Task (Voice Interaction, HiC) – Now the participant waited the robot to get close and gave one of these commands to it: Bark; Sit; Lay; Screech head; Wave tail. Fourth Task (No Interaction, RiC) – In this task, AIBO walked until get close to the participant, and then performed a dance. Fifth Task (Physical Interaction, RiC) – After the dance in the fourth task, now the participant "evaluate" the performance, to do that the participant had to cuddle AIBO in its head (if the participant liked the dance) or in its back (if the participant didn't like the dance). And at last AIBO gave a feedback to the participant: the leds in its face got in two colors, green if the participant had cuddle it in its head or read if the cuddle was in its back. Sixth Task (Voice Interaction, RiC) – The last task was like the third one, the participant waited the robot to get close and gave one of these commands to it: Bark; Sit; Lay; Screech head; Wave tail.

In the last part of the test the participant answered to a questionnaire which had the purpose of knowing how comfortable each task was, how easy was to perform the task and if AIBO attended his or her expectations. These questions used a four-point Likert scale. They also needed to elect two tasks where they felt most comfortable (one from the HiC and another from the RiC groups), write a free text about their thoughts on the interaction and finally we invited them to leave a contact to participate from future studies.

5 Creating the Personas

After the tests we separated the participants in groups to define the Personas using Q-SIM with four different percentage values of similarity (20, 40, 60 e 80). The groups were defined by their similarity of personality (big-five technique) and profile (age, gender). After we got those results we chose the one with 80% (see Table 1) of similarity because it was the one that better represented the participants of this study.

Table 1. Groups obtained from Q-SIM with 80% of similarity. Ex (extraversion), Ag (agreeableness), Co (conscientiousness), Ne (neuroticism) and Op (Openness to experience)

Group	Age	Gender	Ex	Ag	Co	Ne	Op
1	7	Female	5.0	4.5	5.0	4.5	6.0
2	11	Male	5.0	4.5	4.0	4.5	4.5
3	18	Male	4.5	5.0	5.5	4.0	5.5
4	23	Female	5.0	5.0	5.5	5.0	5.0
5	41	Male	5.0	4.5	6.0	3.5	6.5

With the groups defined we began analyze the information that was stored from each participant's test and to separate it in their respective groups. Firstly, we interpreted the scores from the Big-Five technique to define their traits of personality. Taking the conscientiousness values for example, it can be said that the Persona from group five is more careful, focused and self-disciplined than the one in the second group. Secondly we used the data from the CLD with the participants' answers in the post questionnaire to determine how comfortable they were during the interactions. Since none of the groups showed significant reporting of being uncomfortable we defined that they all feel comfortable around the robot. Finally we made video analysis of the interactions to be used with the post-questionnaire in the definition of the Personas' behavior. Below we present the Persona created with the information from the fourth group.

Lyanna is 23 years old and she loves dogs. She is an outgoing person that likes the fellowship of other people. Has a lot of energy and is proactive. Besides, she worries about social harmony, is honest, decent and trustful. Prefers to make plans rather them to act spontaneously, also being too self-disciplined. Rarely gets upset and is too calm. She is always looking for new experiences and thinks of a different way than other people. Her expectation for AIBO is that it will behave like a real dog, been capable to respond to her commands and seek for attention to play. She has never interacted with a robot before AIBO, but she had no difficult to perform the tasks with AIBO. During the interaction she kept saying that AIBO was cute and she was enjoying it. Her preferred tasks were the dancing one and the one that she gave voice commands to AIBO. After the test she said that AIBO attended to her expectations and would like to play with it again.

Fig. 3. Lyanna's Persona

6 Insights and Conclusion

Besides of the creation of Personas, during the analysis we observed that the participants of the tests felt more comfortable with AIBO in comparison with the participants that interacted with different types of robots in Koay et al. [2] study. It was reported that participants started to allow the robots to approach closer to them after five weeks of habituation. This opposes to our tests participants' reactions since only seven reported to be uncomfortable through the CLD even with AIBO getting very close to all them since the beginning of the test. In fact the only situation when they felt uncomfortable was when AIBO bumped at them while moving, but they didn't related to be uncomfortable in the post questionnaire. This proves that they weren't uncomfortable with AIBO itself or during the whole interaction but with that specific moment. One even asked if someone ever felt uncomfortable during the tasks and it was surprised when the evaluator answered yes. Other participants also had more particular reactions like a woman who felt so excited that kept touching AIBO constantly, even when she wasn't performing a task that required physical interaction. Also a young boy asked his mother if was possible to change his real dog for AIBO.

Another study [17] conducted to compare people's interaction with an AIBO and a humanoid ASIMO reported that the most visible difference between the participants' attitude towards both robots the way of giving a feedback to the robot; they tended to use expressions like "thank you" to ASIMO while they frequently touched AIBO to give the feedback. That among with the behavior of our participants leads to the conclusion that due to its characteristics, a pet robot makes people feel more comfortable than those with a humanoid or a machine like appearance.

Finally, this study outlines the methodological approach used to create Personas that address human behavior and psychological characteristics to be used in the development of new applications for robots. The required data was collected from different sources to have more complete and effectively results. Although a pet robot was used in this study, as far as we know the methodological approach can be applied to a robot of a different kind by making some minor changes, such as adapting the tasks to ones that match the robot's functionalities.

References

1. Young, J., Sung, J., Voida, A., Sharlin, E.: Evaluating human-robot interaction. International Journal of Social Robotics 3, 53–67 (2011)
2. Koay, K.L., Syrdal, D.S., Walters, M.L., Dautenhahn, K.: Living with Robots: Investigating the Habituation Effect in Participants' Preferences During a Longitudinal Human-Robot Interaction Study. In: IEEE International Conference on Robot and Human Interactive Communication, Jeju, Korea, pp. 564–569 (2007)
3. Jung, C.G.: The archetypes and the collective unconscious (trans: Hull RFC). Princeton University Press, Princeton (1959)
4. Cooper, A., Reimann, R., Cronin, D., Cooper, A.: About Face 3: The Essentials of Interaction Design. Wiley Pub, Indianapolis (2007)
5. Aquino Jr., P.T., Filgueiras, L.V.L.: A expressao da diversidade de usuarios no projeto de interacao com padroes e personas. In: Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems, IHC 2008, pp. 1–10. Brazilian Computer Society, Porto Alegre (2008)
6. Masiero, A.A., Leite, M.G., Filgueiras, L.V.L., Aquino Jr., P.T.: Multidirectional knowledge extraction process for creating behavioral personas. In: Proceedings of the 10th Brazilian Symposium on on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction, IHC+CLIHC 2011, pp. 91–99. Brazilian Computer Society, Porto Alegre (2011)
7. Ljungblad, S., Walter, K., Jacobsson, M., Holmquist, L.E.: Designing Personal Embodied Agents with Personas. In: The 15th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN 2006, pp. 575–580 (2006)
8. Ruckert., J.H., Kahn Jr., P.H., Kanda, T., Ishiguro., H., Shen, S., Gary, H.E.: Designing for sociality in HRI by means of multiple personas in robots. In: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot in-Teraction (HRI 2013), pp. 217–218. IEEE Press, Piscataway (2013)
9. Duque, I., Dautenhahn, K., Koay, K.L., Willcock, L., Christianson, B.: A different approach of using Personas in human-robot interaction: Integrating Personas as computational models to modify robot companions' behaviour. In: ROMAN 2013, pp. 424–429. IEEE (2013)
10. Masiero, A.A., de Carvalho Destro, R., Curioni, O.A., Aquino Junior, P.T.: Automapersona: A process to extract knowledge automatic for improving personas. In: Stephanidis, C. (ed.) Posters, HCII 2013, Part I. CCIS, vol. 373, pp. 61–64. Springer, Heidelberg (2013)
11. Maulsby, D., Greenberg, S., Mander, R.: Prototyping an Intelligent Agent through Wizard of Oz. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 277–284. ACM Press, New York (1993)
12. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality 37(6), 504–528 (2003)
13. Meerbeek, B., Saerbeck, M., Bartneck, C.: Iterative design process for robots with personality. In: Proceedings of the New Frontiers in Human-Robot Interaction, Symposium at the AISB2009 Convention, SSAISB, pp. 94–101. Springer, Berlin (2009)

14. Koay, K., Dautenhahn, K., Woods, S., Walters, M.: Empirical results from using a comfort level device in human-robot interaction studies. In: Proceedings of HRI 2006: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, pp. 194–201. ACM, New York (2006)
15. Witten, I.H., Frank, E., Hall, M.A.: Data mining: Practical machine learning tools and techniques, 3rd edn. Morgan Kaupmann, USA (2011)
16. Pruitt, J., Adlin, T.: The Persona Lifecycle: Keeping People in Mind Throughout Product Design. Morgan Kaufmann Publishers, San Francisco (2005)
17. Austermann, A., Yamada, S., Funakoshi, K., Nakano, M.: How do users interact with a pet-robot and a humanoid. In: Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2010, Atlanta, Georgia, USA, April 10-15, pp. 3727–3732. ACM, New York (2010)

Robotic Border Crosser TNG - Creating an Interactive Mixed Reality

Anke Tallig

Chemnitz University of Technology, Department of Computer Science, Chemnitz, Germany
anke.tallig@informatik.tu-chemnitz.de

Abstract. In this paper is described an interactive mixed reality which is presented by a mobile robot. It explains the structure and functionality of the mixed reality and illustrated, how the combination works with the robot. In addition some evaluation results of the interactive screen are presented. Usage scenario for the interactive mixed reality is the Industriemuseum Chemnitz. This kind of exhibition is suitable for viewing the inner functions of an exhibit to see how this technology works. The view into a technical device can occurs with the help of a public screen which is projected on the exhibits surface. Via an interactive layer it's possible for users to interact with the indicated contents. This interactive projection system is mobile thereby the robot can transport the public screen through the museum and from exhibit to exhibit. This interactive screen contains videos, animations and pictures of the functionality of exhibits. With the assistance of this mobile system the visitor can learn more about the exhibits in general and their specific functionality.

Keywords: Human-Computer Interaction, Human-Robot Interaction, Mixed Reality, Robotic Mediator, Interdisciplinary Collaboration, Blended Museum.

1 Introduction

Why can't we see inside a technical device? Why we can't see how the technology works? When we can see this, we will better understand the technology and the whole process of construction [1].

For example: Most exhibitions contain several technical devices. Visitors to museums have two options: The first is to see the real exhibit in the exhibition environment. The second is to see additional information via an interactive terminal, which is mostly static and far from the exhibit. So the visitors have to split their attention with these technical devices. To bring these split information together is a real problem for this kind of presentation. Because users of the interactive terminal can't see the real exhibit and visitors who look at the real exhibit can't see additional information from this technical device. Sometimes it can also be problematic for visitors, who are interested in technical devices, to understand how the device works. The animated functionality which is shown via the information terminal isn't easy to understand. Then, when they stand before the exhibit they can't imagine where and how this device functions.[7] In museums one new way to acquire additional information is the use of

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 697–706, 2014.

mobile devices [6]. Unfortunately the attention of the visitors isn't on the exhibit anymore but rather on the mobile device – this causes a split attention situation. But a real demonstration can't happened in any case because some of the exhibits are very old, out of order or their functionality can't be shown in a closed room. However the only chance for a comprehensive understanding is to see, how the system works.

So I thought, why not a combination of all these described components, so that the visitors have all of their benefits. They must have an exhibit which they can see, smell and in many cases, touch. The visitors need the "real device" and also additional information on this exhibit and all adjacent subjects. But this isn't needed in central places – Visitors want additional information close to the exhibit. The museum requires a mobile version of an interaction terminal – The *robotic border crosser*. This robot can accompany the visitors through the exhibition. In front of every technical device the robot prepares the interaction interface for the users and explains the handling of the interaction interface.

Advantages of this kind of mobile interface are the presentation of additional information in a non-static way (see above) and the preparation of the interaction possibility. With the robot companion [3] the visitors are prepared for the interaction with technical contents. So the visitors have a guide and also a "friend" on their side, which is a helping hand for them.

The idea and the basic construction of this mobile robot and the projector system [14] came into being within the interdisciplinary research training group *CrossWorlds*. The research training group *"CrossWorlds – Connecting Virtual and Real Social Worlds"* addresses the increase in digitization and its resulting virtualization of processes, communication, environments, and finally of the human counterparts. This research training group is sponsored by the DFG (Deutsche Forschungsgemeinschaft). Goal of this project is to overcome the current constrains of media-mediated communication. Within interdisciplinary research tandems consisting of computer/engineering scientist and social scientist, we study which new progressions that are offered by the connection of virtual and real social worlds.

The connection of the virtual and real world for the *border crosser* project is shown here, fig. 1. This figure describes the dependencies between real world, virtual world and the *border crosser*. The explanation of this nexus and the advantages of a mobile robot with an interactive interface are topic of the paper "Border Crosser" [12].

The major advantage is learning through the help of both worlds. Visitors can marvel at the real exhibit and then they can explore the virtual presented additional information and other cognate subjects. So this robot can be a simple tutoring system [12], which contains the special museums content and both the presentation and the content, are matched for the visitors specific wants and desires.

Most museum robots can't really show the inner working of an exhibit. For example: The construction with all the single parts, the functionality and the usage. Museum robots can traverse through the exhibitions and function as a greeter [5, 11], some are guides [2, 15, 16], single robots interact with touch displays [10] (for navigation and interaction) and still others have a voice recognition to hear the instruction of the visitors [4]. Problem with all of them: They can't explain technical devices so that the description is understandable.

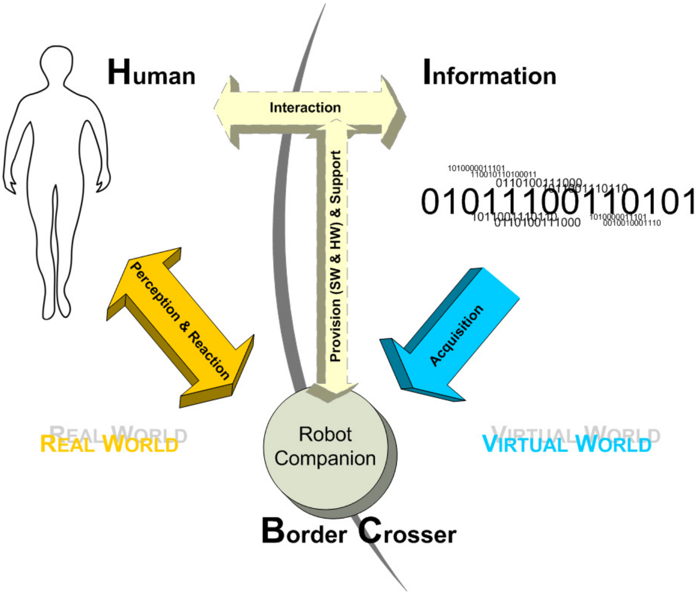


Fig. 1. Nexus of Real and Virtual World

Therefore there is a need for the *robotic border crosser for the next generation (TNG)* which presents details of exhibits in a special technical way.

2 Technical Background

The robot which contains the projection system has to navigate through the exhibition, in robust and secure way. At present this robot is small one which is not very attractive for visitors, because the focus of this project is the projection system and its mobility. For this reason the robot isn't explained in detail. But some points, which are necessary for understanding the projection system, should be presented. The robot must have an average height of an adult. So the projection can happened over the heads of the viewer. One important feature is distance perception. Not only for the secure navigation, this is also relevant for an exact projection on the surface of the exhibit.

2.1 Construction

Before the projection on the surface can happened, some points for technical purposes must be observed:

- projector
- brightness (lumen)
- projection-distance proportion
- wide-angle.

These points affect the distance between the robot and the exhibit surface. The projector, which is used, is a DLP-projector. Because a LED-projector hasn't enough lumen power for a projection in a normal museum area. The brightness of such a projector mustn't be less than 2500 lumen. Museums are places with changing lighting conditions, that's the point for the high brightness. A problem in museums is also the space situation on the premises. Because for the visitors of the exhibition is more than enough space but a projection in front of a technical exhibit requires more space than a single person or a group of people.

Based on the projection-distance proportion and the wide-angle figure 2 accrues. This sketch shows the distances and the total scenario including the position of robot, user and exhibit.

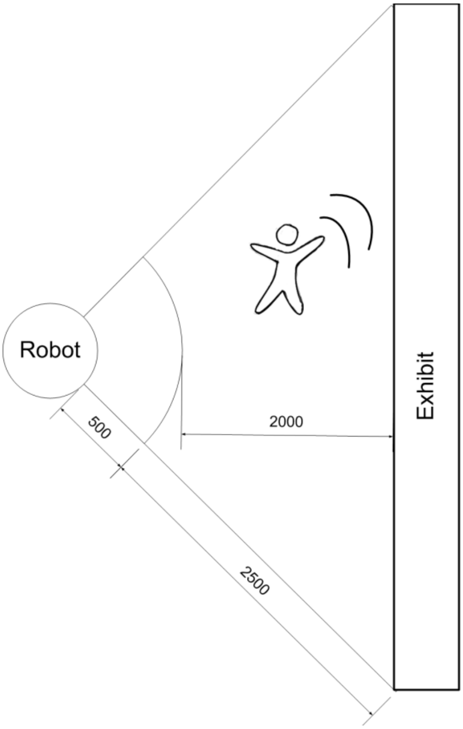


Fig. 2. Sketch of Interaction Distance

This depiction displays the secure distance of the robot, which is 0,5 cm. The interaction area between the robot and the exhibit has a space of 2,0 m length. With these distances a projection of 1,8 m diameter originates.

The space between the robot and the exhibit depends on the height and width of the exhibit. Consequently in front of a small exhibit is less interaction space than in front of large exhibits. Independent of the different distances the projection works. Only the size of the interaction interface depends on space.

The Interaction Pointing Stick. To navigate through the several sub-layers of the interaction interface an infrared pointing stick is necessary. The stick's length is 0,6 m. A button at the handle triggers the interaction event. The handling of this stick seems no problem, because this stick is like a pointing stick in school. The usage of the button works in same way as a mouse button.

As the button on the handle pushed the infrared sender initiates the infrared signal in the stick. On top the projector is the infrared receiver system. After the calibration the receiver calculates a line (l) between the position of it's self (A) and projection (B) on the surface of the exhibit via the infrared signal (C) inside the pointing stick (fig. 3). The extension of the straight line between A and C is the selected item on the interaction interface.

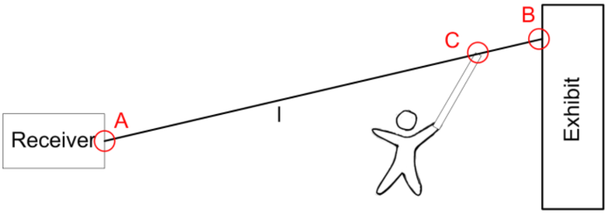


Fig. 3. Scheme of Pointing Stick

A problem of this interaction possibility is the covering of sensor area. If the user stands between the receiver (A) and the pointing stick (C) the sensor has no ability to perceive the position of the selected area. The same problem appears when standing in the projection space. This kinds of deficiencies are discussed under 5 Conclusion, with some possible solutions.

2.2 Projection

The surface structure of the exhibit is the most important item for realistic projection. Surface irregularities of the exhibit must be filtered out of the projection. This happens in such way, that the projection isn't plane – it has to cling on to the exhibit surface. After this mathematical adjustment the projection and the surface of the exhibit appear as a whole (described in figure 4).

The surface of the exhibit and the projection are the first two layers of the mixed reality, which consist of three:

1. surface of the exhibit
2. projection
3. interaction layer.

The interaction layer is the only entity of the projection construction, which is attractive to visitors. Within the first phase this layer shows the same part of the exhibit, which is under the projection. So the user doesn't really see the projection. With the help of a pictogram the user gets an introduction in the handling of this interactive system. By using the pointing stick the user can navigate in the mixed reality.

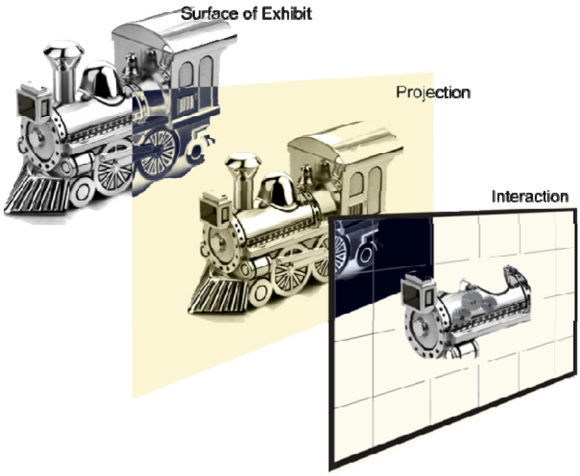


Fig. 4. Layers of Mixed Reality

With the interaction layer the user has the ability to take a look inside the technical device. The projection seems like the real exhibit. If the user selects a area of this projection the cladding of the projected exhibit disappears and a view inside the exhibit is possible. The impression is created that the visitor can take a look inside *real* exhibit. So the user can navigate between the several sub-layers and use the interaction interface like a construction kit.

Every sub-layer shows a different part of the exhibit and its functionality. Is the selected technical section close under the cladding, only the cladding disappears. In this case the user can see the technology on the right place (on that position with which it's in the real exhibit). In the case that the technology is deeper within the exhibit, the selected section is magnified. The magnified section appears as a separate layer, which overlays the normal view. So the user can explore the technology. After observing the technical details the user can push the button on the handle again. In all cases the magnified detail disappears and the normal view of the exhibit is visible.

3 Creating the Mixed Reality

As aforementioned the mixed reality presents the same view as the real exhibit. So the difference between the real and virtual isn't visible. The linkage between the real exhibit and the virtual interface is necessary for comprehensive information. The combination of different types of technology and the real world is the main premise for understanding.[8] Before the user can navigate through the "exhibit construction kid" a small introduction is shown. This introduction appears automatically. The user has the choice of follow the introduction or canceling it. This is important; If we take a look at the different people who want to use this interface, than we can notice that all users have individual preconditions. Users they have more experience with technical devices can cancel the prelude. All other users can follow the instruction of the stick usage and the navigation guide. This part of the interface is intended for all users.[9]

The user prelude is the first part of the simple tutoring system.[13] After the end of the introduction the start view is presented again. All interactive parts of the exhibit are tagged with a flashing frame. The frame blinks from yellow to dark blue and vice versa. This marking version is an outcome of the evaluation, described under point 4. With the help of the pointing stick the user can trigger the interactive parts. After triggering the cladding of the technical device disappears and the technical part is visible.

Figure 5 showed a snippet of the first prototype. It's a Trabant, which is exhibited in the Industriemuseum Chemnitz. One interactive part of this exhibit is the engine of the Trabant. A picture of this engine appears by triggering.

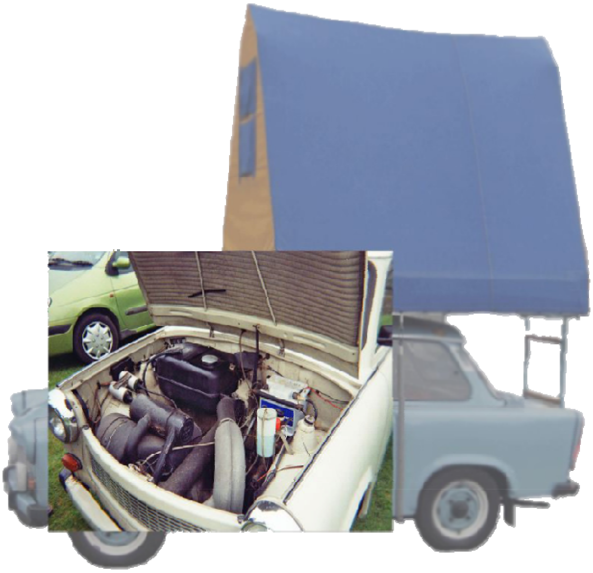


Fig. 5. Snippet of the first Prototype

The interactive interface contains very little text. This interface has a focus on pictures, films and animations. Minimal text and many optical attractions are better for understanding the technical processes.[17]

The user can trigger interactive parts in which order he or she wants. There is no regulation or constraint using the interface. This is also a part of the simple tutoring system.[13] The users have all possibilities. They are independent of time, place and order. So a user has the chance to see everything that is of his or her personal interest. The interactive parts can be trigger unlimited. And depending on the interest of the user the system suggests other similar devices, which the user can view in the exhibition.

At this time, the prototype needs more assistance on graphical presentation. For a presentation close to the real exhibit much more work is necessary.

4 Evaluation Results

During the development of the prototype some evaluations and analysis have happened. For example: One of the first evaluation we have done, is to test the different marks of the interactive parts. The result of this evaluation is depicted in figure 6. It illustrates the percentage of all clicks that are hits or failure. The hit ratio, indicated by color framed interactive parts, is the highest. In the color frame case the frame blinks from yellow to dark blue and vice versa. So the visibility is indicated by all colored backgrounds. The other interaction hint is the area. The hint is marked by a highlighting area. The whole area which is active pulsates. The third version is without any hint. Users have to find out which parts are active.

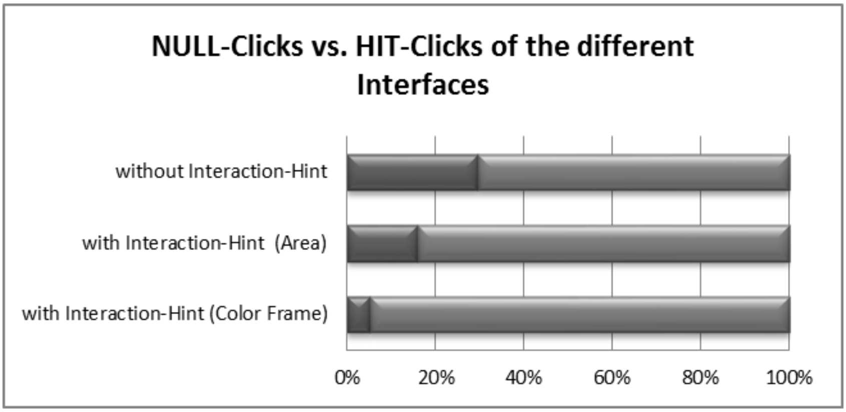


Fig. 6. Result of Different Interaction-Hints

With some more usability tests, the users become fascinated by the interactive interface. The evaluation of the interface are realized in two different areas. The first was a exhibition situation in general, the second the Industriemuseum Chemnitz. 98 percent of the user in the Industriemuseum have been fascinated by the possibility to taking a look in a technical device. Only 56 percent of the users in the normal exhibition area (the exhibition included some oral presentations and information stands) said yes to this interface. But roughly 50 percent of user had different ideas as to what they could do or what they would do with such an interactive interface. They are interested in other presentation areas.

All of the fascinated human guinea pigs wanted more information on this kind of presentation and they wanted more detailed presentation of technical device and information on its background. But all of them wanted also a high-resolution graphic interface.

5 Conclusion

After the evaluation of the interface some new tasks are must to do:

- obtain more information about the exhibits
- more detailed information about the technology and background
- more animations displaying how the technology works
- more exhibits integrated in the tour
- higher-resolution graphic.

Another point to edit is the interaction stick. This is a good and simple device for interaction but it has also a problem. It only works if sensor perceive the infrared stick. If the user stands between the receiver and sender nothing happens. At this time, another technology is being tested – the Kinect sensor. But also this device has advantages and disadvantages. We will see how this system works after the next evaluations.

With this project is being made the first step in a mixed reality in exhibitions with technical devices. Since it is mobile it seems like a private mobile X-ray unit, including a functionality explaining unit. Most of the users have liked this interface and have expressed the desire for such an interactive device.

The next step, which we can do, is prepare a hologram device. With this kind of technology users can see the exhibit from all sides. Desirably would be a sensor system which perceives only the hands of the users. So users can touch the hologram on any part that is interesting to them.

…and the adventure goes on…

References

1. Ballstaedt, S.-P.: Wissensvermittlung die Gestaltung von Lernmaterial. Beltz, Weinheim (1997)
2. Burgard, W., et al.: The Interactive Museum Tour-Guide Robot. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-1998), Madison, Wisconsin (1998)
3. Dautenhahn, K., et al.: What is a Robot Companion – Friend, Assistant or Butler? In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp. 1192–1197 (2005)
4. Faber, F., et al.: The Humanoid Museum Tour Guide Robotinho. In: The 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan (2009)
5. Graf, G., Baum, W., Traub, A., Schraft, R.D.: Konzeption dreier Roboter zur Unterhaltung der Besucher eines Museums. In: VDI-Berichte 1552, pp. 529–536. VDI-Verlag (2000)
6. Initiative D21, T.N.S.: Infratest (pub.), Digitale Gesellschaft 2011, TNS Infratest (2011)
7. Klein, A.: Kunde, Nutzer, Besucher – Besucherforschung und Zielgruppenverständnis. In: FORUM Industriedenkmalpflege und Geschichtskultur 2/04, Unterhalten und Bilden – Anspruch und Wirklichkeit der Industriemuseen, Internationaler Kongress 24 - 26 (2004)

8. Klinkhammer, D., Reiterer, H.: Blended Museum-Vielfältige Besuchererfahrungen durch hybride Vermittlungsstrategien. In: Lucke, U. (ed.) Workshop Proceedings der Tagungen Mensch & Computer, DeLFI 2008 und Cognitive Design 2008, pp. 424–428. Logos-Verlag, Berlin (2008)
9. Lusti, M.: Intelligente tutorielle Systeme Einführung in wissensbasierte Lernsysteme. Oldenbourg Verlag, München (1992)
10. Parlitz, C., Hägele, M., Klein, P., Seifert, J., Dautenhahn, K.: Care-O-bot 3 – Rationale for human-robot interaction design. In: Proceedings of 39th International Symposium on Robotics (ISR), Seul, Korea (2008)
11. Schraft, R.D., Hägele, M., Wegener, K.: Service Roboter Visionen. Carl Hanser Verlag, München (2004)
12. Tallig, A., Hardt, W., Eibl, M.: Border Crosser: A Robot as Mediator between the Virtual and Real World. In: Marcus, A. (ed.) DUXU/HCII 2013, Part III. LNCS, vol. 8014, pp. 411–418. Springer, Heidelberg (2013)
13. Tallig, A.: A Robot Companion as mobile Edutainer. In: CSR-13-04: International Summerworkshop Computer Science 2013 Proceedings of International Summerworkshop, July 17-19. Chemnitz: Chemnitzer Science-Report CSR-13-04 (2013)
14. Tallig, A.: 2012. Grenzgänger-Roboter als Mittler zwischen der virtuellen und realen sozialen Welt. Chemnitz: Chemnitzer Informatik-Berichte CSR-12-05 (2012)
15. Thrun, S., et al.: Probabilistic Algorithms and the Interactive Museum Tour-Guide Robot Minerva. International Journal of Robotics Research 19(11), 972–999 (2000)
16. Thrun, S., et al.: Minerva: A Second-Generation Museum Tour-Guide Robot. In: Proceedings of Robotic and Automation 1999, pp. 1999–2005 (1999)
17. Weidenmann, B.: Wissenserwerb mit Bildern instruktionale Bilder in Printmedien. In: Film/Video & Computerprogrammen. Huber, Bern (1994)

Emotion Transmission System Using a Cellular Phone-Type Teleoperated Robot with a Mobile Projector

Yu Tsuruda, Maiya Hori, Hiroki Yoshimura, and Yoshio Iwai

Graduate School of Engineering, Tottori University
101 Minami 4-chome, Koyama-cho, Tottori 680-8550, Japan

Abstract. We propose an emotion transmission system using a cellular phone-type teleoperated robot with a mobile projector. Elfoid has a soft exterior that provides the look and feel of human skin and is designed to transmit a speaker's presence to their communication partner using a camera and microphone. To transmit the speaker's presence, Elfoid transmits not only the voice of the speaker but also their facial expression as captured by the camera. In this research, facial expressions are recognized by a machine learning technique. Elfoid cannot, however, physically display facial expressions because of its compactness and a lack of sufficiently small actuator motors. The recognized facial expressions are displayed using a mobile projector installed in Elfoid's head to convey emotions. We build a prototype system and experimentally evaluate its subjective usability.

1 Introduction

To communicate with people in remote locations, robots that have human appearance have been developed. Some studies have used humanoid robots for the transmission of human presence. In particular, teleoperated android robots such as Geminoid F and Geminoid HI-1 [1] have appearances similar to an actual person, and were intended to transfer the presence of actual people. These humanoid robots have high degrees of freedom and can transfer human presence. However, they are expensive and limited to a specific individual target. A robot called Telenoid R1 [2] was developed to reduce the number of actuators and costs. Telenoid is not limited to a specific individual target, and is designed to immediately appear and behave as a minimalistic human. A person can easily recognize Telenoid as human; it can be interpreted as male or female, and old or young. With this minimal design, Telenoid allows people to feel as if a distant acquaintance is next to them. Moreover, Telenoid's soft skin and child-like body size make it easy to hold. However, it is difficult to carry around in daily life.

For daily use, a communication medium that is smaller than Telenoid and uses mobile-phone communication technology is now under development. Like a cellular phone, Elfoid is easy to hold in the hand, as shown in Fig. 1. When we use such robots for communication, it is important to convey the facial expressions of a speaker to increase the modality of communication. If the speaker's facial movements are accurately regenerated via these robots, human presence can be

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 707–714, 2014.

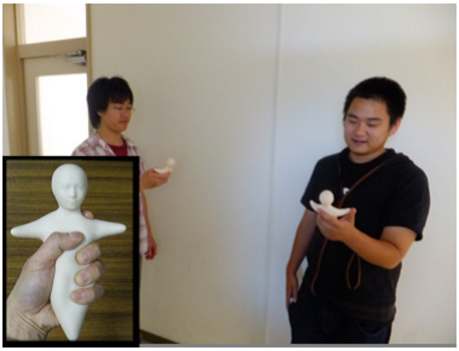


Fig. 1. Elfoid: cellular phone-type teleoperated android

conveyed. Elfoid has a camera within its body and the speaker's facial movements are estimated by conventional face-recognition approaches. However, it is difficult to generate the same expression in robots because a large number of actuators are required. Elfoid cannot perform facial expressions like a human face can because it has a compact design that cannot be intricately activated. That is, since Elfoid's design priority is portability, its modality of communication is less than Telenoid's. For this reason, it is necessary to convey emotions some other way.

2 Generation of Facial Expressions Using Elfoid

2.1 Elfoid: Cellular Phone-Type Teleoperated Android

Elfoid is used as a cellular phone for communication, as shown in Fig. 1. To convey the human presence, Elfoid has the following functions.

- Elfoid has a body that is easy to hold in a person's hand.
- Elfoid's design is recognizable at first glance to be human-like and can be interpreted equally as male or female, and old or young.
- Elfoid has a soft exterior that provides the feel of human skin.
- Elfoid is equipped with a camera and microphone.

Additionally, a mobile projector is mounted in Elfoid's head and facial expressions are generated by projecting images from within the head, as shown in Fig. 2.

2.2 Overview of the Total System

In this research, facial expressions are generated using an Elfoid's head-based mobile projector to convey emotions. Fig. 3 shows an overview of the total system.

First, individual facial images are captured using a camera mounted within Elfoid. Next, the facial region is detected in each captured image and feature points on the face are tracked using the Constrained Local Model (CLM) [3]. Facial expressions are recognized by a machine learning technique using the positions of the feature points. Finally, recognized facial expressions are reproduced using Elfoid's head-based mobile projector.

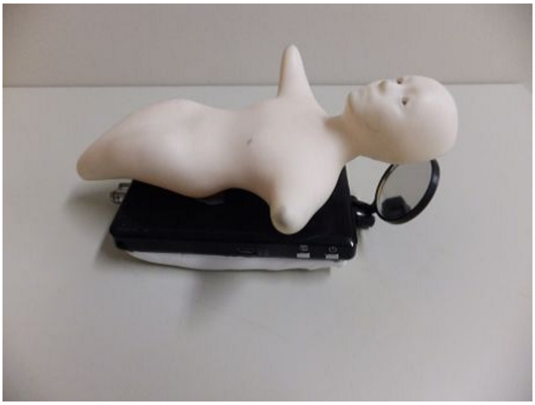
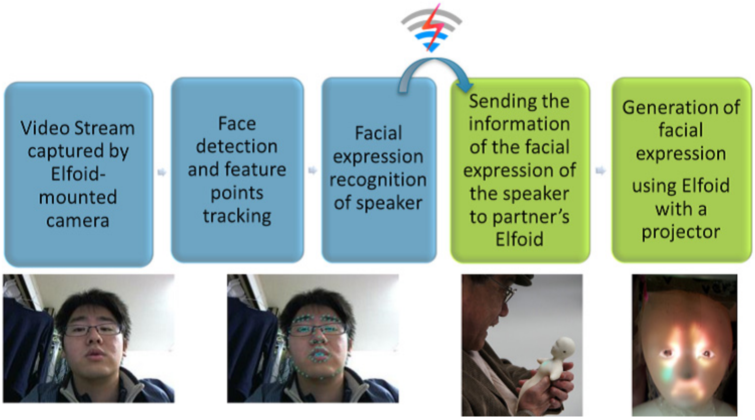


Fig. 2. Prototype system



Video Stream captured by Elfoid-mounted camera
Face detection and feature points tracking
Facial expression recognition of speaker
Sending the information of the facial expression of the speaker to partner's Elfoid
Generation of facial expression using Elfoid with a projector

Fig. 3. Overview of the total system

2.3 Recognition of Facial Expressions

CLM fitting is the search for point distribution model parameters $\mathbf{p}$ that jointly minimize the misalignment error over all feature points. It is formulated as follows

$$\mathcal{Q}(\mathbf{p}) = \mathcal{R}(\mathbf{p}) + \sum_{i=1}^{n} \mathcal{D}_i(\mathbf{x}_i; \mathcal{I}), \tag{1}$$

where $\mathcal{R}$ is a regularization term and $\mathcal{D}_i$ denotes the measure of misalignment for the ith landmark at $\mathbf{x}_i$ in image $\mathcal{I}$. In the CLM framework, the objective is to create a shape model from the parameters $\mathbf{p}$. The misalignment term, $\mathcal{D}_i$, is estimated using the mean-shift technique. This method has low computational complexity and is robust to occlusion. The results of feature point tracking are shown in Fig. 4.

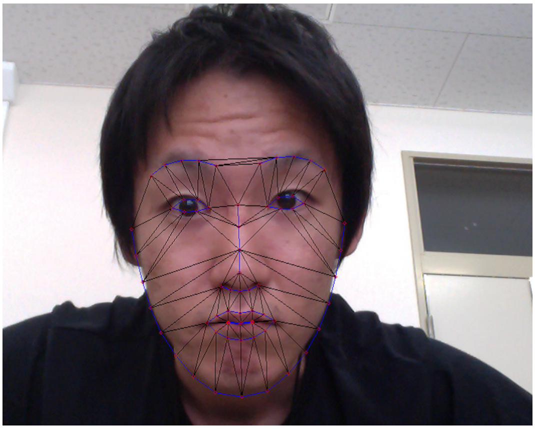


Fig. 4. Results of feature point tracking

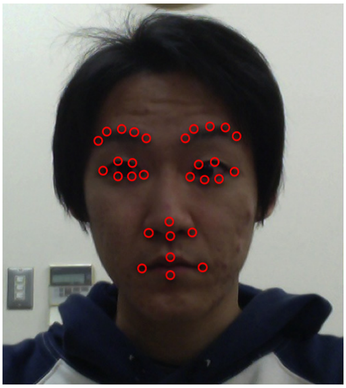


Fig. 5. Feature points used for classification

In this study, six facial expressions that correspond to universal emotions [4], happiness, fear, surprise, sadness, disgust, and anger, are classified using a hierarchical technique similar to [5]. The facial expressions are hierarchically classified by a Support Vector Machine (SVM). Each classifier is implemented beforehand using the estimated positions of feature points. The feature points used for classification are shown in Fig.5. This study is based on the theory that different expressions can be grouped into three categories [6,7] based on the parts of the face that contribute most toward the expression. These categories are shown in Fig. 6 At the first level, we use 31 feature points around the mouth, eyes, eyebrows and nose to discriminate the three expression categories: lip-based, lip-eye-based, and lip-eye-eyebrow-based. After grouping into three categories, each category is divided into two emotion classes. In the lip-based category, four feature points around the mouth are used for expressing happiness or sadness. In the lip-eye-based category, 16 feature points around the mouth and eyes are

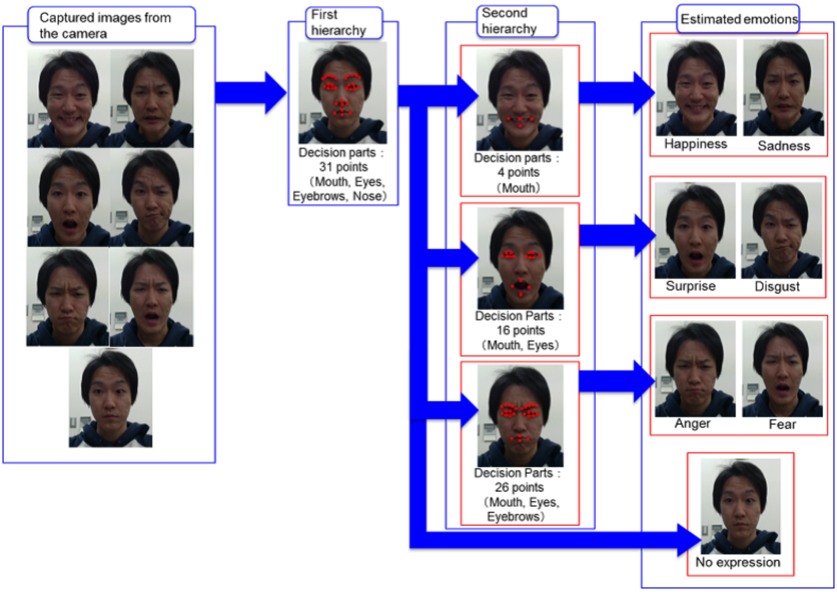

Captured images from the camera
First hierarchy
Decision parts : 31 points (Mouth, Eyes, Eyebrows, Nose)
Second hierarchy
Decision parts : 4 points (Mouth)
Decision Parts : 16 points (Mouth, Eyes)
Decision Parts : 26 points (Mouth, Eyes, Eyebrows)
Estimated emotions
Happiness
Sadness
Surprise
Disgust
Anger
Fear
No expression

Fig. 6. Emotion estimation using a hierarchical technique

used for expressing surprise or disgust. In the lip-eye-eyebrow-based category, 26 feature points around the mouth and eyes and eyebrows are used for expressing anger or fear.

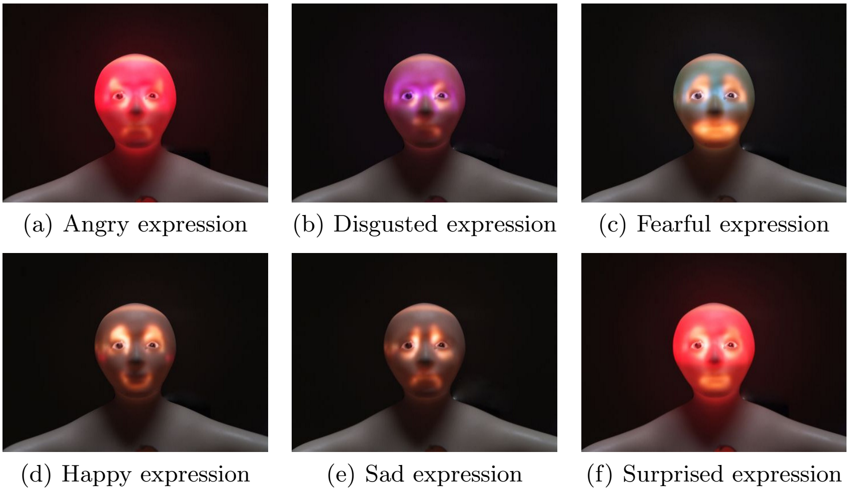
2.4 Generation of Facial Expressions with Elfoid Using Cartoon Techniques

Recognized facial expressions are reproduced using Elfoid's head-based mobile projector. To represent facial expressions, we generate stylized projection patterns using the results of emotion estimations. In this study, the projection patterns are stylized using cartoon techniques [8]. It is widely recognized that cartoons are very effective at expressing emotions and feelings. The movements around the mouth and eyebrows, for example, are exaggerated. The silhouette of face and shapes of eyes are varied by projection effects. Moreover, color stimuli that convey a particular emotion are added.

3 Experiment

3.1 Recognition Rate of Facial Expressions

We conducted an experiment to verify the recognition rate of facial expressions. As training data, we used a total of 8,000 images that consisted of 1,000 images for each facial expression and 2,000 images with no expression. To verify the rate


(a) Angry expression (b) Disgusted expression (c) Fearful expression

(d) Happy expression (e) Sad expression (f) Surprised expression

Fig. 7. Facial expressions generated by Elfoid

of the facial expression recognition, we tested 1,000 images for each expression that were different from the training data. Table 1 shows the facial expression recognition rate results. The results show that this estimation method can be applied to our communication system.

3.2 Subjective Evaluation of the Proposed System

In this experiment, representative face expressions were generated by Elfoid, as shown in Fig. 7. From the experiments also reported in [8] , it is shown that each emotion can be conveyed correctly. Additionally, to verify the validity of this system, we experimentally evaluated its subjective usability. Subjects had a conversation with a communication partner at a remote location using Elfoid. We used an Elfoid that does not project facial expressions as a comparison. We then gave the subjects questionnaires that asked about the satisfaction level

Table 1. Recognition rate for facial expressions (%)

estimation answer	happiness	sadness	surprise	disgust	anger	fear	no expression
happiness	**73.8**	3.0	0.0	0.0	23.2	0.0	0.0
sadness	1.4	**60.1**	0.0	0.0	33.9	0.0	4.6
surprise	0.0	0.0	**82.8**	0.0	17.2	0.0	0.0
disgust	0.0	0.0	0.0	**82.3**	17.7	0.0	0.0
anger	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0
fear	0.0	0.0	1.9	0.0	0.2	**97.9**	0.0
no expression	0.0	0.0	0.0	0.0	10.4	0.0	**89.6**

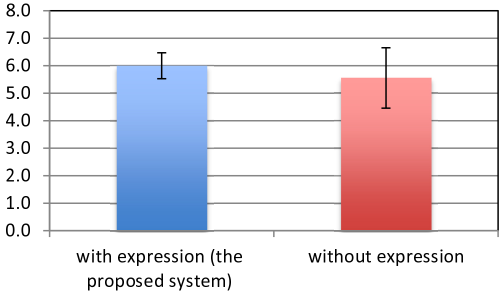


Fig. 8. User's satisfaction with the conversation

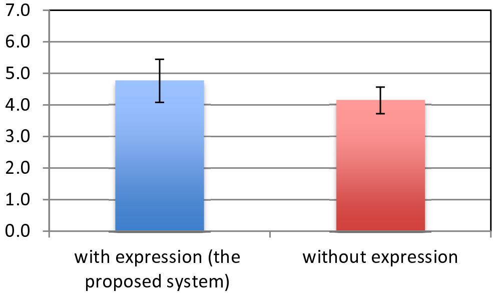


Fig. 9. User's impression of a conversational partner

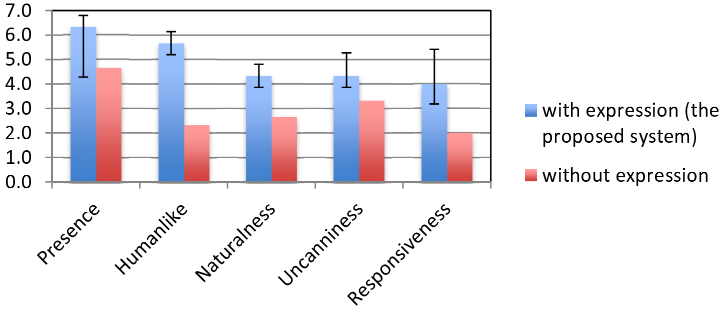


Fig. 10. User's impression of the interface

of the conversation, impression of a conversational partner, and impression of the interface. These items were based on the conventional method for assessing usability [9]. To ascertain the user's impression of the interface, presence, humanlike, naturalness, uncanniness, and responsiveness were investigated.

Figs. 8, 9, and 10 show the experimental results. Higher scores indicate a better impression. According to the results, the proposed system is effective for increasing conversation satisfaction level and the impression of a conversational partner. The evaluation regarding the impression of the interface is higher for the proposed system than the comparison system.

4 Conclusion

We propose an emotion transmission system using a cellular phone-type teleoperated robot with a mobile projector. In this research, facial expressions are recognized by a machine learning technique, and displayed using a mobile projector installed in Elfoid's head to convey emotions. In the experiments, we built a prototype system that generated facial expressions and evaluated the recognition rate of facial expressions and the subjective evaluations of usability. Given the results, we can conclude that the proposed system is effective for increasing conversation satisfaction level and the impression of a conversational partner.

Acknowledgment. This research was supported by the JST CREST (Core Research for Evolutional Science and Technology) research promotion program "Studies on cellphone-type teleoperated androids transmitting human presence."

References

1. Asano, C.B., Ogawa, K., Nishio, S., Ishiguro, H.: Exploring the uncanny valley with geminoid HI-1 in a real-world application. In: Proc. Int'l Conf. of Interfaces and Human Computer Interaction, pp. 121–128 (2010)
2. Ogawa, K., Nishio, S., Koda, K., Balistreri, G., Watanabe, T., Ishiguro, H.: Exploring the natural reaction of young and aged person with Telenoid in a real world. Jour. of Advanced Computational Intelligence and Intelligent Informatics 15(5), 592–597 (2011)
3. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int'l Jour. of Computer Vision 91(2), 200–215 (2011)
4. Ekman, P., Friesen, W.V.: Unmasking the Face. Prentice-Hall, Inc. (1975)
5. Siddiqi, M.H., Lee, S., Lee, Y., Khan, A.M., Truc, P.T.H.: Hierarchical recognition scheme for human facial expression recognition system. Sensors 13, 16682–16713 (2013)
6. Nusseck, M., Cunningham, D.W., Wallraven, C., Bülthoff, H.H.: The contribution of different facial regions to the recognition of conversational expressions. Journal of Vision 8, 1–23 (2008)
7. Schmidt, K.L., Cohn, J.F.: Human facial expressions as adaptations: Evolutionary questions in facial expression resarch. American Journal of Physical Anthoropology 116, 3–24 (2001)
8. Tsuruda, Y., Hori, M., Yoshimura, H., Iwai, Y.: Generation of facial expression emphasized with cartoon techniques using a cellular-phone-type teleoperated robot with a mobile projector. In: Kurosu, M. (ed.) HCII/HCI 2013, Part V. LNCS, vol. 8008, pp. 391–400. Springer, Heidelberg (2013)
9. Sakamoto, D., Kanda, T., Ono, T., Hagita, N., Ishiguro, H.: Android as a telecommunication medium with a human-like presence. In: Proc. of the ACM/IEEE Int'l Conf. on Human-Robot Interaction, pp. 193–200 (2007)

Emotions Recognition

Design of an Emotion Elicitation Framework for Arabic Speakers

Sharifa Alghowinem[1,3], Sarah Alghuwinem[4], Majdah Alshehri[5],
Areej Al-Wabil[5], Roland Goecke[2,1], and Michael Wagner[2,1]

[1] Australian National University, Research School of Computer Science, Canberra, Australia
[2] University of Canberra, Human-Centred Computing Laboratory, Canberra, Australia
[3] Ministry of Higher Education: Kingdom of Saudi Arabia
[4] Princess Noura Bint Abdulrahman University, Social Science College, Riyadh, Saudi Arabia
[5] King Saud University, Human-Computer Interaction Group, Riyadh, Saudi Arabia

`sharifa.alghowinem@anu.edu.au`, `sarah.alghuwinem@gmail.com`,
`{amajdah,aalwabil}@ksu.edu.sa`, `roland.goecke@ieee.org`,
`michael.wagner@canberra.edu.au`

Abstract. The automatic detection of human affective states has been of great interest lately for its applications not only in the field of Human-Computer Interaction, but also for its applications in physiological, neurobiological and sociological studies. Several standardized techniques to elicit emotions have been used, with emotion eliciting movie clips being the most popular. To date, there are only four studies that have been carried out to validate emotional movie clips using three different languages (English, French, Spanish) and cultures (French, Italian, British / American). The context of language and culture is an underexplored area in affective computing. Considering cultural and language differences between Western and Arab countries, it is possible that some of the validated clips, even when dubbed, will not achieve similar results. Given the unique and conservative cultures of the Arab countries, a standardized and validated framework for affect studies is needed in order to be comparable with current studies of different cultures and languages. In this paper, we describe a framework and its prerequisites for eliciting emotions that could be used for affect studies on an Arab population. We present some aspects of Arab culture values that might affect the selection and acceptance of emotion eliciting video clips. Methods for rating and validating Arab emotional clips are presented to derive at a list of clips that could be used in the proposed emotion elicitation framework. A pilot study was conducted to evaluate a basic version of our framework, which showed great potential to succeed in eliciting emotions.

Keywords: Emotion elicitation framework, Arabic emotion data collection, emotional movie clips.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 717–728, 2014.

1 Introduction

Emotions have been widely investigated lately for their importance not only to psychology, neurobiology and sociology, but also for affective computing studies. Affective computing is the study of automatic detection of human emotional states, which has seen much interest lately for its multidisciplinary applications. For example, Human-Computer-Interaction (HCI) is concerned with enhancing the interactions between users and computers by improving the computer's understanding of the user's needs, which includes understanding the user's affective state [1]. In the education field, understanding the emotional state of a student could lead to a more effective presentation style [2]. A current interest in the personalization of commercial products could be enhanced by understanding the client preference based on their mood [3]. Moreover, such understanding of the user's emotions could enhance other applications such as virtual reality and smart surveillance [4]. Such automatic recognition of emotions could also be useful to support psychological studies; for example, to give a baseline of the emotional response of healthy subjects, which could be compared to and used to diagnose mental disorders such as depression [5] or neurodevelopmental disorders such as autism [6].

To study emotions in an efficient, reliable and replicable way, a standardized laboratory setting is needed to induce emotional responses. Studies of emotions in the literature can be divided into simple, discrete, and dimensional emotion representations. Simple emotion representations investigate and compare only positive and negative emotions. Discrete emotion representations use a finite number of distinct classes, such as Ekman's basic universal emotions which are *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* [7]. In dimensional emotion representations, emotions are represented as points in continuous space along multiple dimensions such as valence / pleasantness and arousal / intensity.

Standardized techniques to elicit emotions, such as reading self-referent statements, listening to music, watching film clips, recalling autobiographical information, or combinations of these, have been surveyed in [8]. In addition to these, other techniques have been used, such as odors, emotional imagery, facial expression, and social interactions [9, 10]. Emotion eliciting film clips are widely used in affective studies [11, 12] for their great advantages compared to other techniques, which include the standardization ability, deception reduction, dynamical property, ecological validity, and results replicability [13]. In addition, movie clips proved to induce discrete emotions and could (with the correct method of evaluation) induce dimensional emotions [10]. A limitation, however, of using emotion eliciting film clips is that the selected clips have to be validated to induce the emotions in question. To date, four studies have been conducted to validate movie clips in three different languages and cultures. In [14], a first collection of 12 English-speaking films was shown to elicit six emotional states. Subsequently, [13] presented a collection of 16 English-speaking films that elicited eight different emotional states. Another study validated 70 French speaking movie clips collected from different cultural backgrounds (French, Italian, British and American) [10]. Recently, validated Spanish dubbed clips with the capacity to induce emotions were presented in [9].

Knowing that cultures have a great influence on emotions and their triggers, the same set of elicitation clips could produce different results for different cultures. In [15], it has been concluded that emotional practice (expression and interpretation) of an emotion drastically differs between cultures. Also in [16], the author discussed the cultural influence on emotional responses, concluding that the differences between cultures lie in eliciting certain emotions. The author supports his conclusion with an example of differences between the reactions of northern and southern Americans to the same emotion eliciting stimuli [16]. Considering cultural differences and language between Western countries and Arab countries, it is possible that some of the validated clips, even when dubbed, will not elicit similar responses. Moreover, dubbing might cause subtle differences that could distort the emotion evaluation and some of the clips may not be culturally acceptable by the Arab community. Therefore, there is a need to select and evaluate a set of emotion elicitin clips that are suitable for Arabic speakers in affect studies on an Arab population.

To the best of our knowledge, studies of affect in the Arab world are scarce. Given the unique and conservative culture, a standardized and validated framework for future studies is needed in order to be comparable with current studies of different cultures and languages. In this paper, we describe a framework and its prerequisites for eliciting emotions that could be used for studies of affect on an Arab population. We present some aspects of Arab culture values that might affect the selection and acceptance of emotion eliciting video clips. Then, in line with previous studies, we display validation methods for the selected clips. We present a pilot study conducted in Saudi Arabia to share some observations and to demonstrate the effectiveness of a basic framework.

2 Framework Prerequisites

For the framework to succeed, two preparation steps have to be performed first. Since the main component of our proposed emotion elicitation framework is movie clips, these clips have to be carefully selected and validated to prove their effectiveness in inducing the emotions in question. Figure 1 summarizes the stages for our proposed framework.

2.1 Selecting and Rating Emotional Clips

As mentioned in Section 1, only three previous studies have been conducted to select and rate emotional clips. In [13], 494 participants were divided into 31 groups to group-view 10 clips (different for each group) from a total of 78 that were refined from 250 clips selected by colleagues and film critics to induce eight emotions (anger, fear, sadness, disgust, surprise, amusement, contentment, or natural). After each film, subjects were asked to fill an emotion self-reporting questionnaire [13]. Similarly in [9], 127 Spanish subjects individually watched 10 clips of carefully selected subsets of 57 films to induce seven emotions (anger, fear, sadness, disgust, amusement, tenderness, or natural). A self-report questionnaire

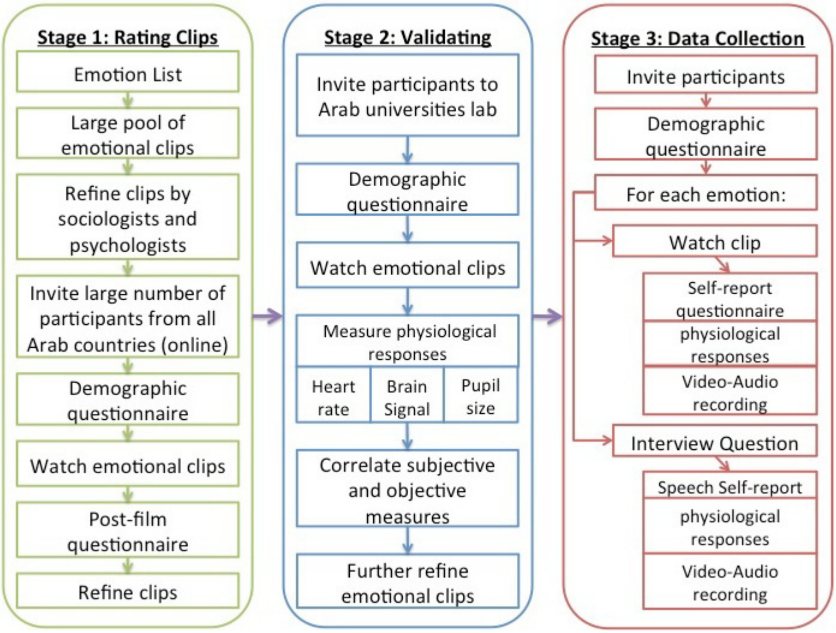

Stage 1: Rating Clips
Emotion List
Large pool of emotional clips
Refine clips by sociologists and psychologists
Invite large number of participants from all Arab countries (online)
Demographic questionnaire
Watch emotional clips
Post-film questionnaire
Refine clips
Stage 2: Validating
Invite participants to Arab universities lab
Demographic questionnaire
Watch emotional clips
Measure physiological responses
Heart rate
Brain Signal
Pupil size
Correlate subjective and objective measures
Further refine emotional clips
Stage 3: Data Collection
Invite participants
Demographic questionnaire
For each emotion:
Watch clip
Self-report questionnaire
physiological responses
Video-Audio recording
Interview Question
Speech Self-report
physiological responses
Video-Audio recording

Fig. 1. Proposed framework and its prerequisites

of dimensional and categorical representations was completed by each subject immediately after watching each clip [9]. For the list of French clips [10], 70 clips were selected from 824 clips by 50 movie experts to measure seven emotions as in [17]. The refined clips were divided to sets of 10 films each, then were viewed by 364 participants, assigned to 7 groups. Emotions were measured by several self-report questionnaires: a 7-point scale of emotional arousal, a differential emotions scale, and a positive and negative affect schedule [10]. In our proposed framework, in line with previous studies, a standardized list of emotions to be induced in Arab population includes anger, fear, sadness, disgust, amusement, surprise, and natural.

The Arab countries in the Middle East and North Africa have unique and conservative cultures, Saudi Arabia being the most conservative one, which makes clip selection criteria tricky. Even though most Arab cultures share the same language and religion, they are diverse for reasons such as history, foreign colonisation, and revolutions [18]. Most of the general characteristics and values in Arab cultures are based on the Islamic religion, emphasising respect for all religions and prophets, honoring and obeying the parents, and following and complying to Islamic religion rules [19]. Arab cultures share general characteristics and values, such as a tribal structure, and honor, chivalry and courage values to defend their tribe and allied tribes, especially protecting women whether or not they are related to them [19, 20]. Not surprisingly, these cultural values have influenced and restricted the media to reflect on Arab population expectation and acceptance [21].

Even though Western media is popular in Arab countries, content undergoes a thoughtful censorship [22] or adjustments [23]. For example, in [13], a clip to elicit happiness, using Jesus' name as fun material, is not accepted in Arab cultures due to a respect for religions and prophets. Another clip in [13] to elicit amusement shows sexual references, which is strictly unacceptable to the Arab community and media. Another important aspect of Arabic clip selection is the huge gender difference in expressing emotions between men and women due to the cultural expectations of both genders [19, 24]. For example, a clip showing a woman being flirted with will induce fear in Arab female viewers, but rage in Arab male viewers, due to valuing honor in Arab cultures [19, 24]. As a consequence, clips that could lead to different responses based on the viewer's gender should be avoided.

Due to the complexity and variability of Arab cultures, social experts and psychologists should review and refine a pool of emotion inducing clips suitable for Arab cultures. Beside cultural selection criteria, a few other criteria should be considered. First, the clips should be relatively short in duration and at the same time should induce emotions without additional background explanation. Second, each clip should induce a specific emotion from the identified list of emotions in question. Cinema and theater production in Arab countries are limited, which increases the quality and number of television production, including television series [22]. Particularly, Arab television drama score the highest viewing rate at 99.7%, followed by religious shows, and news [22]. Although using television series adds the difficulty of finding a short clip without additional background to induce a specific emotion, the high quality and popularity of such series might overcome this issue. Moreover, editing such clips to include minimal background would be sufficient to elicit emotions.

Once a sufficiently large pool of video clips has been selected, a rating scheme of the emotional effect of each video clip should be run on a random and dispersed Arab population. Following previous studies, the clips will be assigned to subset groups, where each subset should contain at least one clip to elicit each emotion. A post-viewing questionnaire should be completed after each clip, covering dimensional and categorical emotions as in [9, 10]. The clip subsets should then be viewed by subjects of different Arab countries and cultures, in gender balanced and age matched groups. Given the large number of Arab countries and the sparse distance between them, the internet could be used to facilitate the rating of the selected clips. Although using the internet is convenient for participants to rate the clips from home in their spare time, the variability in the subjects' mood as well as the variability in continuity (pauses) while watching the clips might affect the rating. Knowing that controlling environment settings could be compromised using the online rating, this step should be done only to rate the selected clips for further validation. In order to reduce outliers and variability caused by the lack of control, the number of participants should be large enough to mitigate this issue. Subjects could be invited with arrangements and collaborations with universities from each country. Before participating, invited subjects should electronically sign a consent form and also fill a demographic questionnaire asking about their age, country, cultural heritage, physical and mental health, etc.

2.2 Validation of Selected Clips

Given the diversity of Arab cultures and the online method proposed to rate the clips, objective emotional response measures of top rated clips are necessary not only to validate the rating, but also to further refine the top rated emotional clips. Once a top rated clips list has been selected, validating this list should be conducted to measure the correlation of self-reported results with their physiological reaction. Most of the emotion eliciting clips in previous studies relied only on the self-report, subjective measures, mentioning the importance of objective validity of their clip selection [9, 10, 13]. Only in [17], the selected emotional clips in [9] were validated by measuring physiological responses using the skin conductance level and heart rate, where a convergence between subjective and objective responses was found.

An extensive literature review on physiological activities in emotion using several cardiovascular, respiratory, and electrodermal measures summarized the effectiveness of such measures as an indication of emotional activities [25]. With the use of current technologies, brain signals, heart rate, skin temperature, and eye activities could be measured in a normal university lab setting. We propose the use of portable multi-channel electroencephalogram (EEG) devices to measure brain signals, the use of skin conductance devices to measure sweat, heart, and respiratory rate, beside blood pressure, and the use of an eye tracking device to measure eye activity and pupil size. The top rated clips from the first stage could be shown to invited participants in a lab setting individually or in groups to measure the physiological activity. Although showing clips to participants in groups could speed up the validation, multiple physiological measure devices might not be available. Another advantage of the group setting is that it might simulate real-word emotion expression; however, in such exploratory studies comparably expressing emotions might be preferred in an individual setting. Based on the objective validation of the emotional clips using physiological measures, a final emotional clips list will be selected. With these physiologically refined clips, a standardized list of Arabic emotion elicitation clips will be produced, which can be used for emotional data collection.

3 Emotion Elicitation Framework Design

Our proposed framework contains induced emotions by emotional clips and spontaneous emotions elicited by asking affective questions in an interview. By this stage, a list of Arabic emotion elicitation clips will be validated and standardized, which will be used as one of the main components of our proposed framework. After each emotion eliciting video clip, a question will be asked designed to arouse that particular emotion. Our proposed framework is described in the following subsections and summarized in Figure 1.

3.1 Participants

To collect a large and rich Arab emotional database, participants selection should cover cultures, genders, ages and socio-economic levels. As mentioned earlier, collaborations with most Arabic universities might cover the diversity of Arab cultures. Ideally, a gender balanced participant cohort would be beneficial for emotional differences between genders. Covering age groups, where participants will be divided into three groups (e.g. $< 25, 30 - 45, > 50$), will be beneficial to psychology and sociology studies to study the effect of age in the physiological and behaviour measures. Finally, having a balance of at least two socio-economic level groups will be constructive to study the sociology differences in affect.

3.2 Recording Environment

Although non-standardized recording environment is more easily accessible, the variability from such environment is challenging and might affect the results. Therefore, a standardized recording environment and settings is preferable in such exploratory studies. Ideally, a highly controlled recording environment would be desirable to get high quality and clean recording. However, finding such recording environment for all locations of Arabic universities might not be feasible. Therefore, a reasonably quiet room with good lighting might be sufficient. Using semi-structured interviews to elicit spontaneous emotions (see Section 3.4) requires an individual recording setting for participants.

3.3 Hardware

Since the database in this framework is intended to be large and rich for multidisciplinary studies, both physiological measures and audio-video emotion expressions will be collected. We propose the use of portable multi-channel electroencephalogram (EEG) devices to measure brain signals, skin conductance devices to measure sweat, heart, and respiratory rate, beside blood pressure, and the use of an eye tracking device to measure eye activity and pupil size. In addition, audio-video recordings of participants' facial and vocal emotional expressions will be made for further affect analysis.

3.4 Protocol

Demographics: Before participating, invited subjects must sign a consent form and also complete a demographic questionnaire. The demographic questionnaire will cover general personal questions such as age, country, and cultural heritage, and general health questions about physical and mental health conditions. Questions about general personality characteristics as well as the subject's economic and social situation, will be useful to rationalize results. The aim for such a questionnaire is to have a wide understanding of differences or similarities in emotional responses.

Watching Movie Clips: The validated clips from the second stage are deemed to elicit the selected emotions. Participants will watch one clip per emotion, starting with a neutral clip to normalize their affect. After watching each clip, participants will be asked to complete a short post-film questionnaire to self-report the intensity of each emotion they felt.

Interview: To induce spontaneous emotions, emotion eliciting questions about events that aroused specific emotions in the subject's life will be asked after watching each clip. That is, after watching the amusement clip, a question about a happy moment in subject's life will be asked. Although it might not be convenient to interview the subjects after each clip, this method is beneficial as each clip will the prepare subject's mood for that particular emotion.

4 Challenges and Opportunities

The large number of Arab countries, the sparse distances between them and the diversity of cultures within each country will introduce several challenges as well as opportunities. Opportunities for collaboration with universities spanning the varied cultures of the Arab world call for careful coordination and planning. Rules and regulations vary in different countries, which need to be taken into consideration when implementating the proposed framework to allow for sufficient time for the study to be carried out in different contexts. For example, video recording females on Saudi Arabia's university campuses is restricted [26], a collaboration with other institutions (e.g. hospitals, private organizations or companies) could overcome this issue. A restricted confidentiality and ethics agreement for such data collection will strengthen the acceptance of this type of research within the Arab population (both males and females) to participate.

5 A Pilot Study

A proof-of-concept version of the framework was designed to elicit positive and negative emotions only using two video clips and two interview questions. The paradigm and initial results are desribed in the following.

5.1 Emotion Elicitation Paradigm

Given the unique culture of Saudi Arabia, and to ensure acceptance of all participants, video clips demonstrating positive (joy) and negative (sad) emotions were selected from classic cartoon animation series dubbed in Arabic, namely: "Heidi" and "Nobody's Boy: Remi", respectively. The clips had an almost similar duration ($\sim 2.5min$). Participants were asked to rate the emotional effect of each clip and whether they have seen the clip before, as that could affect their response. The positive emotion clip had a positive affect rate of 8.27 out of 10, with 10 being the highest positive effect. 57% of the participants had seen the clip before. The negative emotion clip had a negative affect for 6.43 out of 10

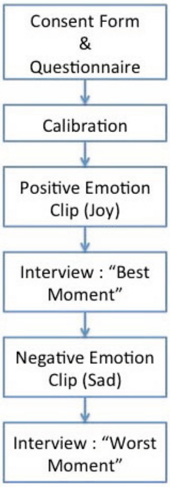

Consent Form
&
Questionnaire
Calibration
Positive Emotion Clip (Joy)
Interview : "Best Moment"
Negative Emotion Clip (Sad)
Interview : "Worst Moment"

Fig. 2. Emotion elicitation paradigm and data collection process for the pilot study

participants, with 10 being the highest negative effect. 7% of the participants had seen the clip before.

Apart from inducing emotions, watching video clips served as a preparation of the participant's affect for the spontaneous emotion elicitation via a semi-structured interview, where participants were interviewed about an emotional event in their life. That is, after watching the positive emotion clip, the participants were asked about the happiest moment in their life. For the negative emotion, after watching the negative emotion clip, participants were asked about the saddest moment in their life. The paradigm of our data collection is shown in Figure 2.

For emotion eliciting clips, self-report was carried out by rating the clips and answering questions about their feelings after watching the clips. For the interview questions, we assume that the questions elicit the emotions, although the answers were not validated for certain emotions. We also physiologically measured emotion elicitation using an eye tracker (see Section 5.3).

5.2 Participants

In this experiment, 71 native Arabic speakers were recruited from a convenience sample (65 females, 6 males). The age ranged from 18 to 41 years ($\mu = 25.6, \sigma = 4.8$). Regular participants' mood and mental state are important for the study. None of the participants reported any mental health disorder (no clinical validation). 72% of the participants reported they were in their normal, neutral mood at the time of recording, 7% always sad, and 22% always happy.

5.3 Recording Environment Settings

We used a Tobii X120 eye tracker attached to a Toshiba Satellite L655 laptop. We used a PowerLite 1880 XGA Epson projector screen as an extended monitor to the laptop to ensure that the participants look at similar coordinates while watching the clips and while talking to the interviewer. While the participants watch the clips, the interviewer leaves the room to reduce distractions and to allow the participant to freely watch the clips. The interviewer enters the room for the interview questions and locates themselves in front of the projector screen. The screen resolution and distance from the projector screen and the eye tracker location were fixed in all sessions. Although we had limited control of light in the recording room, we normalized the extracted features for each segment of each participant to reduce the light variability coming from the video clips themselves and the room light.

5.4 Initial Findings and Observations

Due to ethic restrictions at King Saud University regarding video recordings of participants, observations were made only by the interviewer at the time of the interview and were not recorded. Regarding negative emotions, while watching the clip, 39% of participants rated the clip to have a strong effect (more than 8 out of 10), though only almost 1% cried over the clip. On the other hand, while answering the negative emotion interview question, 70% of the participants cried (including one male participant). Since the negative clip shows a death scene, almost 85% participants talked about their negative emotion during losing a loved person in their life. Other topics included injustice, failure and conflict with a close person. Those late findings indicate that watching the video clips prepared the participant mood for the spontaneous emotions in the interview. Since the number of male participants was not enough to make any reliable gender comparisons, more data needs to be collected.

For the positive emotion, while watching the movie clip, 53% of participants rated the clip to have a strong effect (more than 8 out of 10). In contrast, while answering the positive emotion interview question, only 0.7% of the participants cried while expressing their joy (none of which were males). Our observations indicate that unlike happiness crying, sadness crying was associated with eye contact avoidance. Pupil size measurements indicated dilation activities during emotion expression, with more activity in the spontaneous emotion expression. However, for a reliable conclusion, more participants need to be recorded.

In our work [27], we only analysed eye activities as an indication of the emotional state. In general, the automatic classification results using eye activity were reasonable, giving 66% correct recognition rate on average. With more channels to be included (facial expression, voice analysis, physiological cues, etc.), we anticipate a higher recognition rate. Moreover, while expressing spontaneous emotions, the recognition rate of positive and negative emotions is slightly higher than for induced emotions. This finding indicates that spontaneous emotions might have stronger eye activity patterns than induced emotions. Statistical measures show statistically significant differences in eye activity patterns between positive and negative emotions. We found that pupil dilation size and duration increase while expressing negative emotions. We also found less eye contact due to head rotation. In our previous work [28], we investigated fixation features and found significant differences in fixation duration and count between positive and negative stimuli.

Given that these initial observations and findings are based on a basic version of our overall framework, which only investigated positive and negative emotion elicitation, we anticipate a great potential for our framework to succeed.

6 Conclusions

In this paper, we presented a standardized and validated framework for future studies of emotions that is needed for the unique and conservative Arab cultures. Such framework is important in order to be comparable with current studies in Western cultures and languages. We describe a framework and its prerequisites for eliciting emotions that could be used for affect studies on an Arab population. We present some aspects of Arab cultural values that might affect the selection and acceptance of emotion elicitation video clips. Two main prerequisites must be performed before collecting emotional data using the proposed framework: rating and then validating Arabic emotion eliciting clips. The validation scheme will finalize a list of clips that will elicit emotions to be used in our proposed emotional elicitation framework. Our suggested framework contains both induced emotions by emotion eliciting clips and spontaneous emotions induced by answering affective questions in an interview. After watching each video clip, a question will be asked arousing that particular emotion. We also conducted a pilot study in Saudi Arabia to test the feasibility and effectiveness of our framework on a small scale. Our initial findings and observations are encouraging as they showed the successful elicitation of emotions.

Acknowledgment. The work described in this paper was supported in part by grant RGP-VPP-157.

References

[1] Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. In: Proceedings of the 8th International Conference on Multimodal Interfaces, pp. 239–248. ACM (2006)

[2] Craig, S., Graesser, A., Sullins, J., Gholson, B.: Affect and learning: an exploratory look into the role of affect in learning with autotutor. Journal of Educational Media 29(3), 241–250 (2004)

[3] Zhou, F., Ji, Y., Jiao, R.J.: Affective and cognitive design for mass personalization: Status and prospect. Journal of Intelligent Manufacturing, 1–23 (2012)

[4] Tao, J., Tan, T.: Affective computing: A review. Affective Computing and Intelligent Interaction, 981–995 (2005)

[5] Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. In: Proc. FLAIRS-25, pp. 141–146 (2012)

[6] Hobson, R.P., Ouston, J., Lee, A.: Emotion recognition in autism: Coordinating faces and voices. Psychological Medicine 18(4), 911–923 (1988)

[7] Ekman, P.: An argument for basic emotions. Cognition & Emotion 6(3-4), 169–200 (1992)

[8] Gilet, A.: Mood induction procedures: A critical review. L'Encephale 34(3), 233–239 (2008) (in French)

[9] Fernández, M.C., Pascual, M.J., Soler, R.J., Fernández-Abascal, E.: Spanish validation of an emotion-eliciting set of films. Psicothema 23(4), 778 (2011)

[10] Schaefer, A., Nils, F., Sanchez, X., Philippot, P.: Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition and Emotion 24(7), 1153–1172 (2010)

[11] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)

[12] Jerritta, S., Murugappan, M., Nagarajan, R., Wan, K.: Physiological signals based human emotion Recognition: A review. In: 2011 IEEE 7th International Colloquium on Signal Processing and its Applications (CSPA), pp. 410–415 (March 2011)

[13] Gross, J.J., Levenson, R.W.: Emotion elicitation using films. Cognition & Emotion 9(1), 87–108 (1995)

[14] Philippot, P.: Inducing and assessing differentiated emotion-feeling states in the laboratory. Cognition & Emotion 7(2), 171–193 (1993)

[15] Mesquita, B., Frijda, N.H., Scherer, K.R.: Culture and emotion. Handbook of Cross-Cultural Psychology 2, 255–297 (1997)

[16] Richerson, P.J., Boyd, R.: Not by genes alone: How culture transformed human evolution. University of Chicago Press (2008)

[17] Fernández, C., Pascual, J.C., Soler, J., Elices, M., Portella, M.J., Fernández-Abascal, E.: Physiological Responses Induced by Emotion-Eliciting Films. Applied Psychophysiology and Biofeedback 37(2), 73–79 (2012)

[18] Barakat, H.: Arab society in the twentieth century: Research in changing conditions and relationships, vol. 1. Center for Arab Unity Studies (2000) (Arabic)

[19] Al-Saif, M.: Introduction to the study of Saudi society: Approach in sociology and functional analysis of the community, and scientific lessons in social change and education. Dar Al-Khurajy publication (1997) (Arabic)

[20] Long, D.E.: Culture and customs of Saudi Arabia. Greenwood Publishing Group (2005)

[21] Al-Hasan, A.: The basics of media production and display standards, vol. 1. Safeer Puplication (2010) (Arabic)

[22] Al-Hassan, A.: The Basics of Television Drama Production, vol. 1. Safeer Puplication (2010) (Arabic)

[23] Ayish, M.: Television reality shows in the arab world: The case for a glocalized media ethics. Journalism Studies 12(6), 768–779 (2011)

[24] Nydell, M.K.: Understanding Arabs: A guide for Westerners. Intercultural Pr (2002)

[25] Kreibig, S.D.: Autonomic nervous system activity in emotion: A review. Biological psychology 84(3), 394–421 (2010)

[26] Ministry of Education: Rules and regulations (May 2013)

[27] Alghowinem, S., AlShehri, M., Goecke, R., Wagner, M.: Exploring eye activity as an indication of emotional states using an eye-tracking sensor. In: Chen, L., Kapoor, S., Bhatia, R. (eds.) Intelligent Systems for Science and Information. SCI, vol. 542, pp. 261–276. Springer, Heidelberg (2014)

[28] Alshehri, M., Alghowinem, S.: An exploratory study of detecting emotion states using eye-tracking technology. In: Science and Information Conference (SAI), pp. 428–433. IEEE (2013)

Analysing Emotional Video Using Consumer EEG Hardware

Jeroen de Man

VU University Amsterdam, Department of Computer Science
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
j.de.man@vu.nl

Abstract. Low-cost, easy to use EEG hardware produced for the consumer-market provide interesting possibilities for human-computer interaction in a wide variety of applications. Recent years have produced numerous papers discussing the use of these types of devices in various ways, but only some of this work looks into what these devices can actually measure. In this paper, data is used that has been collected using a Myndplay Brainband, while 30 participants viewed emotional videos eliciting different mental states. This data is analysed by looking at average power in multiple frequency bands and eSense™ values, as well as peaks in the measurements detected throughout the videos. Although average values do not differentiate well between the mental states, peak detection provides some promising results worthy of future research.

Keywords: emotional response, emotion analysis, affective HCI, EEG.

1 Introduction

Electroencephalogram (EEG) is a measurement of electronic activity on the scalp. Recent developments made hardware measuring EEG signals available at low prices. Although most of these devices contain fewer sensors than professional EEG devices, data quality is comparable [1]. More important though, these devices are very easy to use and can therefore be used in a variety of settings. Not surprisingly, more and more devices are appearing aiming at a particular consumer audience. Neurosky was one of the first companies developing a dry sensor EEG device 'to power the user-interface of games, education and research applications'.[1] Besides common EEG measurements, this device also outputs eSense™ values representing levels of attention and meditation [2]. Based on the same chipset, Myndplay developed the Brainband, 'the world's first mind controlled media player'.[2] InteraXon is far along developing the Muse, an EEG device using six dry sensors to 'manage stress and settle your mind'.[3] This is not an exhaustive list and more examples can easily be found.

[1] http://neurosky.com/products-markets/eeg-biosensors/, accessed 04-02-2014.

[2] http://www.myndplay.com/, accessed 04-02-2014.

[3] http://www.interaxon.ca/muse/, accessed 04-02-2014.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 729–738, 2014.

Just as Neurosky mentions research applications, many of these devices with access to the measurement data are appearing in scientific research. In the research project STRESS, a simulation-based training is envisioned to train professionals in high risk jobs to handle stressful incidents and improve their decision making in these situations. A software agent is being developed to analyse the trainee's mental state and provide support by either giving textual feedback or adapting the training scenario whilst running. However, such applications should not be limited to professional use and by using cheap and easy to use devices, in combination with the development of generic techniques, creating a virtual training to cope with everyday stress is envisioned as a feasible next step.

At this point in time, the virtual environment to be used for this research project is in its final stages of development. Therefore, the current work uses video to investigate whether such commercial EEG devices, in this case the Myndplay, can be used in such a training context to detect relevant mental states. The next section provides some background information on scientific research using these types of consumer EEG devices in various methods and applications. Afterwards, the process of data collection is described as well as the method for data analysis. The last two sections cover results and a discussion thereof.

2 Background

An extensive amount of research has been done using professional EEG hardware in a wide variety of contexts and applications. However, with the development of cheap, easy to use EEG hardware, using EEG as a human-computer interface for many types of applications has become plausible. Recent years have brought an increasing number of research papers, using these low-cost consumer EEG devices in a variety of applications. A selection of this work will be described below.

There are however differences between professional and consumer EEG hardware that need to be pointed out before continuing. First and foremost, the number of sensor is very limited on these low-cost devices with often just a single sensor, as is the case in this work as well. Therefore, measurements are made in only one location, which roughly corresponds to the FP1 position. Furthermore, while most professional devices require some sort of conductive gel or paste to be used, so-called dry sensors are used which eliminate this need without reducing data quality [1].

Consumer EEG devices are used in various scientific applications. Among them are applications using the outputted data in a way that is unrelated to the conceptual meaning of the measurement. For example, [3] uses the eSense™ attention value to drive forward a robot. In [4] a game is developed using a similar premise, where both the attention and mediation values are used to manipulate objects in the game. As a last example here, [5] also uses both these eSense™ values in an interactive story where the values are used to reach certain goals depending on the current chapter.

However, from a user-perspective it might be more natural when the action (increasing attention or meditation) matches the action required in the game [6]. For example, in [7] an archery game is used where the player needs to be focussed and relaxed in order to make a good shot. Here, there is a more natural connection between the type of action required to make and the effect it has on the game. However, making such clear mappings from brain activity to particular actions is rather difficult. Other work is using measurements such as attention and meditation more indirectly. In [8], attention is used to manipulate weather conditions during gameplay. A more sophisticated proposal has been made in [9], where attention levels result in particular in-game events aimed at increasing the attention of the player. Another option is to use biofeedback, where the measurements are simply displayed to the user. Training mindfulness is an area where such a technique is used and consumer EEG hardware makes it way in providing these measurements [10, 11].

Another set of research papers focuses on discovering the potential of these devices through experimentation. In [12], a puzzle game is used to evaluate the eSense™ meditation values. They found that in trials where participants showed stressed behaviour, meditation values often dropped below a certain threshold. However, during a routine stress-free task, these measurements never dropped below 40. In another study, a consumer EEG device is used to investigate the possibility of measuring interest during a first-person shooter game [13]. Here, they found that attention levels of individual players spiked simultaneously during gameplay moments common to each player such as killing an enemy.

In this paper, we will add to this last set of research. Using emotional video, it is investigated whether consumer grade EEG devices can be used to detect various mental states. This work is in line with research such as [14], where a more advanced EEG device is used to mark highlights in a video. Furthermore, most related work only focuses on the eSense™ values provided by the Neurosky chipset, but here all frequency bands will be considered as well. Various methods of analysis used in studies described above will be applied, such as looking at average values, lowered and heightened attention and meditation values as well as common peaks across participants in any of the measurements.

3 Method

For this research, the focus is on stressful stimuli, more specifically stressful movies. In the next section, the experimental setup used for data collection is explained. The second section covers the various approaches used for data analysis.

3.1 Data Collection

Data collection was performed simultaneous with an experiment investigating the possibility of inducing anxiety using stressful video material [15]. Here, a group of 30 participants was shown five fragments of video in sequence. Between each video, the participants were asked about the emotion experienced (‘relaxed’, ‘bored’, ‘interested’, ‘excited’, ‘scared’) and its intensity on a scale from 0 (not at all) to 5 (very much). Of the participants, 17 were male and 13 female with an age between 20 and 64 years old (mean age of 33). EEG signals were recorded using the Myndplay Brainband, which is based on the Neurosky chipset. Furthermore, heart rate and skin conductance was measured using the Plux wireless biosensors[4], but are not used in this paper.

[4] `http://www.biosignalsplux.com/`, accessed 04-02-2014.

Figure 1 shows the sequence of videos viewed by the participants. In [15] it is shown that each type of video (beach/documentary/stressful) induces different emotions; relaxed/boring, interesting and exciting/scary respectively. Furthermore, skin conductance shows a clear elevation during the stressful movie, which slowly decays in the subsequent clips. Based on this information, EEG measurements obtained during this experiment are expected to reflect those under different emotional states, thereby providing a relevant dataset for investigating the potential for consumer EEG devices to distinguish such different mental states.

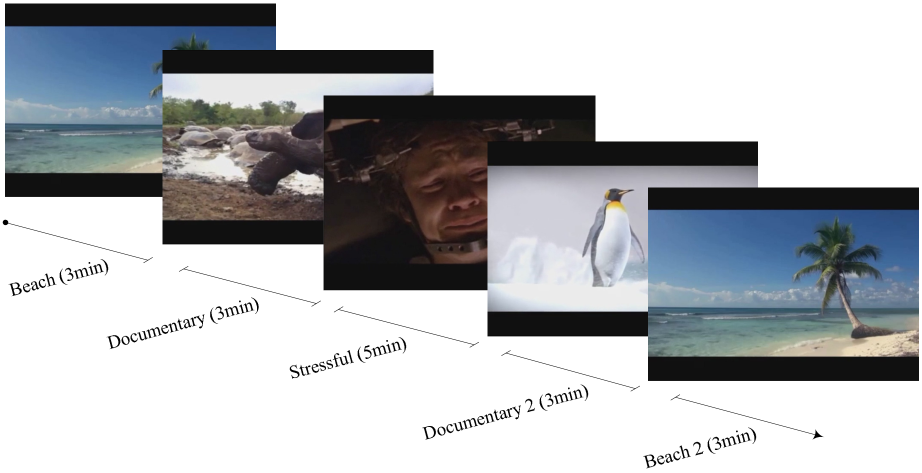


Fig. 1. Sequence of video clips used for data collection

3.2 Analysis

The hardware used provides raw EEG measurements, per second activity in frequency bands ranging from delta waves (1-3Hz) up to mid-gamma (41-50Hz) as well as eSense™ values for attention and meditation [2]. For the current research, the frequency band values from the device itself were used which already have undergone noise-filtering and therefore raw EEG signals have not been used. Although little is known about the basis of the eSense™ values, they are used in related research and are thus also incorporated in this work.

Data analysis will be performed with a focus on two different aspects. First, average values of each variable (frequency bands and eSense™) for each movie will be calculated. These averages will be statistically compared to find any potential markers for particular mental states corresponding to the videos.

A different approach will also be taken, that is by calculating an average and standard deviation for each of the different frequency bands. Using these values, peaks in subsequent movies were identified by finding peaks more than four standard deviations away from the mean value. For the eSense™ values, a slightly different approach is used. These values range between 0 - 100 and are slightly lowered or elevated when the value decreases below 40 or increases above 60 according the documentation. As the current research focuses on negative stimuli, increased values for attention and decreased values for meditation are of our interest. Therefore, values above 60 in both the attention and the inversed meditation[5] value will be considered as peaks and used as such in analyzing the results.

4 Results

The results of this research are discussed in two separate sections as described above. First, the average values over each video clip are compared. Afterwards, a more in-depth analysis of peaks in each signal is performed.

4.1 Mean Activity

Figure 2 shows a collection of bar charts, including standard deviations, for each measurement that is provided by the Brainband. Starting from the top left, the first two graphs represent the average eSense™ values. The remaining eight graphs show values for the different frequency bands that are outputted by the device itself. On visual inspection, it is clear that there is a large variation between the subjects, as could be expected. At this point, no individual tuning or normalisation has been performed. Looking at the average values, both for the eSense™ and frequency bands, no clear effect of the different movies can be distinguished, although some graphs show signs of possible effects.

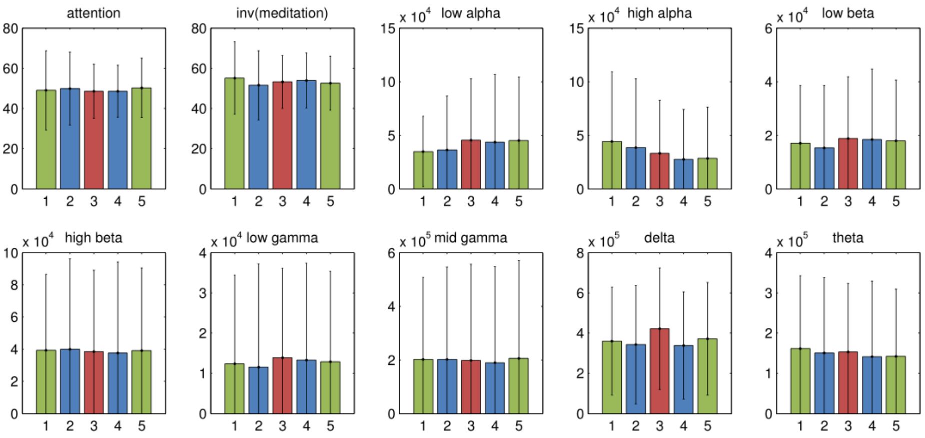


Fig. 2. Average power (and standard deviation) of each variable for video clips 1 to 5

[5] For consistency with the other results, mediation values are inversed by subtracting the actual value from 100. The resulting value, inv(meditation), can be analyzed similarly to the attention values by looking at.

To check the results, a repeated measures ANOVA is performed for each outputted variable. With this measurement, it is calculated whether there is a statistical difference in the mean value of any of the five videos. Unfortunately, in none of the variables, a statistical difference was found with p values of 0.13 (meditation), 0.18 (high alpha) up to 0.90 (high beta). Similar test were performed after normalising the measurements for each participant, resulting in slightly lower p values with a significant differences in the theta waves (F(4,27)=3.05, p=0.02). Post-hoc analysis using a Bonferroni correction showed a significant difference between the second and third clip, that is the first documentary and the stressful movie (t(27)=3.40, p=0.02).

4.2 Peaks

A second method of analysis, peak detection, was used following the method described in Section 3.2. Figure 3 shows the peaks for each of the four videos that followed the first baseline movie. On the x-axis time is represented in seconds, with the stressful movie being 300 seconds long instead of the 180 seconds for the other three videos. The y-axis shows the various frequency bands as well as the attention and meditation. Each 'x' represents a peak for that aspect on that time point. In case of the meditation, the value has been inversed such that each peak represents the minimum of meditation as described above.

As can be seen, there are some clear differences between the different videos, as well as a lack of peaks in the eSense™ values compared to these bands. In the following paragraphs, first the attention and meditation values will be discussed, thereafter the frequency bands. Before discussing all the results, the final paragraphs in this section relates these peaks to specific scenes in the video to get a feel for possible causes underlying the various peaks.

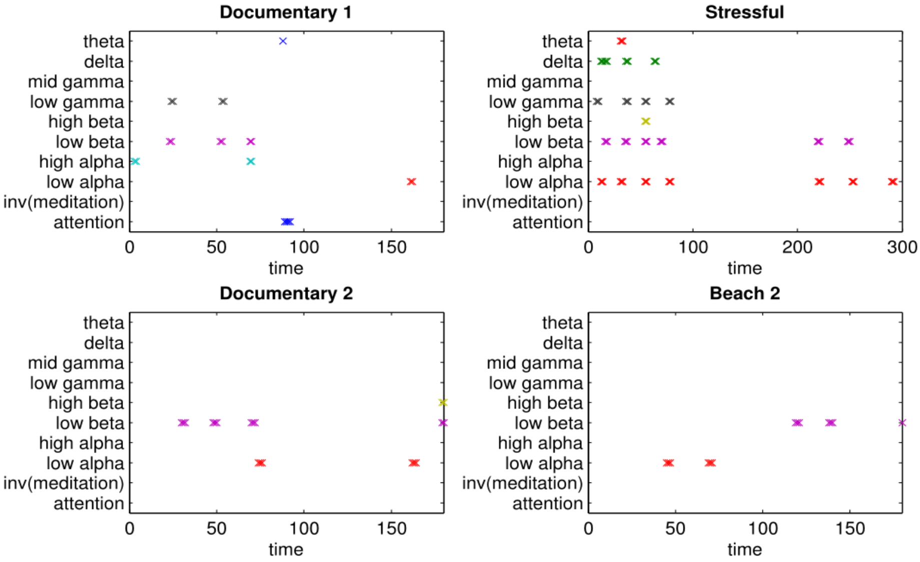


Fig. 3. Measurement peaks in the last 4 clips compared to the first baseline movie

eSense™ Values. The bottom two lines represent moments of slightly increased levels of attention and slightly lowered levels of meditation. Considering the attention level, there is only one point during the neutral video where it is increased. There is also only one point, during the same video, where attention is slightly decreased, but these results are not shown here. Overall, this would imply that during the experiment, participants paid an average amount of attention throughout.

Looking at the mediation values, there are no lowered values in any of the videos. For completeness, increased values of meditation were also checked. Here, each video showed a few moments spread out over the duration of the clip where meditation was increased. Even the stressful movie showed increased values for meditation, however all of these peaks were concentrated around the last minute of the video.

Frequency Bands. Peaks in average power for each of the provided frequency bands are shown in the top rows of each graph in Figure 3. On first glance, it is clear that the stressful video produced many more peaks than the neutral videos and even less peaks are produced in the last clip of a beach. To better grasp the data underlying these peaks, take a look at Figure 4. Here, the average power (line) as well as the standard deviation (area) of the low-beta frequency during the stressful movie are shown. Peaks in the line, which are more than four standard deviations above the average during the first movie, are shown in the figure above.

Although during each video there are several peaks in various frequency bands, there are some clear differences between the stressful movie and the other ones. For one, peaks during the stressful movie appear in more bands simultaneously than in the other videos. Furthermore, there are multiple peaks in the delta band during the stressful movie, whereas none of the other videos contain peaks in this frequency range.

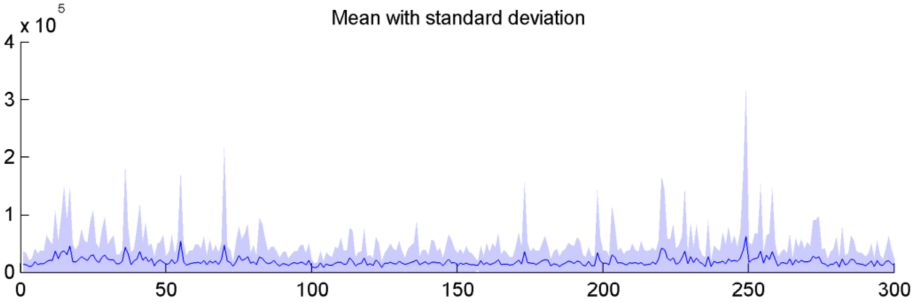


Fig. 4. Average power (and standard deviation) over time in the low beta frequency band

Peaks in Relation to Video Content. In this paragraph, an in-depth examination will be made of the exact content of each video around the peaks. Although this analysis is rather subjective, it might offer some insights into possible reasons for peaks to occur, which in turn can be used for further research.

During the first documentary clip, the two peaks in various bands after 23 and 67 seconds coincide with strange sounds made by giant tortoises. The peaks after 52 and 160 seconds occur when one tortoise tries to climb another tortoise. The peak in attention after roughly 90 seconds occurs together with the introduction of some 'exciting' music. Peaks in the second documentary have less clear relations to the content, with the peak after 30 seconds co-occurring with a penguin walking into an ocean and the two peaks after 50 and 60 seconds marking the beginning and end of an computer generated animation. The clip of the beach was chosen to be as unexciting as possible, and these peaks subsequently cannot be related to any remarkable event in the movie.

The stressful video contains many peaks, among which many occur roughly simultaneously. We consider the contents of the video around 13, 32, 55, 78, 221, 252 and 291 seconds of interest. These moments are based on the peaks in the low alpha range as these always seem to occur at roughly the same time as peaks in any of the other bands. A multitude of these instances coincide with typical moments of fright in the video. There are however a few moments that are more interesting. At 55 seconds, the clip briefly shows a woman huddled up on the floor of a psychiatric hospital. The peak at 221 seconds is after a longer scene with ominous music playing while the camera slowly moves through a house. The peak occurs at the end, when a door is slowly opened, just to reveal another empty room. Each of the other peaks mentioned above occur at moments of fright such as a snake jumping towards the camera or during some explicit horror scene. It has to be said however, that this stressful video contains more of these types of moments, without a visible peak in the measurements.

5 Discussion

Many different results are reported in the previous section. Here, an attempt is made to discuss some of the major points brought forward by these results. First, the lack of any significant findings in the average values over each video is discussed. Secondly, the absence of peaks in the eSense™ values are of our interest, as many related research did successfully use these measurements. Lastly, the promising results using peaks in the frequency bands are discussed further, explaining what future research is needed to better understand and be able to use such measurements.

Looking at the average value for each video resulted in non-significant differences in each of the measured variables. For attention and meditation, differences were very small and variation between and within participants was rather large. This could have obscured any potential differences, however in our opinion it is more likely that on average no differences exist. Possible changes within attention and meditation due to some visual scene seem to come and go rather fast. Therefore, it might be more plausible to expect the short moments where those values might significantly have changed to be averaged out by the longer period in which no change is present. A similar explanation can be given for the lack of significant differences in the frequency bands. No explanation can be given for the one significant effect that was found.

The reasoning above might be why many relevant research using the eSense™ variables not look at the average, but at the periods in which it gets above or below a certain threshold, as was done here in Section 4.2. However, during none of the videos, attention or meditation was even slightly increased or decreased. Thus the question remains why other research was able to use these values to detect changes in attention and/or meditation and no effects were found here. One problem here is that due to the company trademark of these measurements, no information is available on how they are exactly calculated. However, another explanation could be based on the sensor location at FP1 measuring activity in the prefrontal cortex. The data used is obtained while participants viewed emotional video. Possibly this passive nature of the task might involve less activation from the prefrontal cortex, which is commonly associated with higher order cognitive processes that may be required in game-like setups as reported in related literature. However, based on the fact that there are findings made using the frequency bands, it cannot be ruled out that the underlying methods of calculating the eSense™ values might play a role in explaining these findings.

Throughout each of the videos, peaks occurred in various frequency bands. First, it was clear that during the stressful movie, more peaks in more frequencies occurred. Furthermore, for the first documentary and the stressful movie, it was possible to manually find plausible explanations for these peaks. Questions arise however, when it is considered that during the stressful movie similar scenes exist which do not coincide with a peak in the data or that peaks in the second documentary and in the last clip of an empty beach cannot be explained in a similar fashion. With regard to missing peaks for the stressful movie, it could be that these scenes were not stressful enough to cause such a peak. Alternatively, it might be the case that during the movie some form of emotion regulation was applied, thereby suppressing later peaks. On the other hand, peaks without a plausible cause in the videos could be that it is still not exactly known what causes these effects and the reason for the peak to occur is not obvious enough here to notice on visual inspection. This is however less plausible for the last video, which is the same video as participants saw the first time when no peaks were produced.

Thus, although these results are promising, more research is required before such measurements can be used in any application. A tailor-made experiment in order to get more frequent subjective feedback on the mental state of participants can help in understanding the exact nature of the peaks. Furthermore, it is interesting to investigate the effects of active effort, instead of passive viewing of film, on data obtained with these consumer EEG devices. Combining these results, might give valuable insights into when and where these devices can be used for human-computer interaction.

Acknowledgements. Special thanks go out to Marco Stam, for conducting the experiments that provided the data for this research. Furthermore, a thank you is in order for Tibor Bosse and Charlotte Gerritsen for their support and input on this work. This research was supported by funding from the National Initiative Brain and Cognition, coordinated by the Netherlands Organisation for Scientific Research (NWO), under grant agreement No. 056-25-013.

References

1. NeuroSky: Brain Wave Signal (EEG) of NeuroSky, Inc. (2009)
2. NeuroSky'seSense™ Meters and Detection of Mental State, NeuroSky, Inc. (2009)
3. Vourvopoulos, A., Liarokapis, F.: Brain-controlled NXT Robot: Tele-operating a robot through brain electrical activity. In: 2011 Third International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), pp. 140–143. IEEE (May 2011)
4. Diefenbach, P., Bhatt, H., Gupta, A., Lorenz, J., Lyon, P., Stevenson, C., Stratton, P.: Maxwell's Demon: A Study in Brain-Computer Interface Game Development (2004)

5. Yoh, M.S., Kwon, J., Kim, S.: NeuroWander: A BCI game in the form of interactive fairy tale. In: Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing-Adjunct, pp. 389–390. ACM (September 2010)
6. Nacke, L.E., Kalyn, M., Lough, C., Mandryk, R.L.: Biofeedback game design: Using direct and indirect physiological control to enhance game interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 103–112. ACM (May 2011)
7. Lee, K.: Evaluation of Attention and Relaxation Levels of Archers in Shooting Process using Brain Wave Signal Analysis Algorithms. 감성과학 12(3), 341–350 (2009)
8. Cho, O.H., Kim, J.Y., Lee, W.H.: Implement of weather simulation system using EEG for immersion of game play (2013)
9. Rebolledo-Mendez, G., De Freitas, S.: Attention modeling using inputs from a Brain Computer Interface and user-generated data in Second Life. In: The Tenth International Conference on Multimodal Interfaces (ICMI 2008), Crete, Greece (2008)
10. Stinson, B., Arthur, D.: A novel EEG for alpha brain state training, neurobiofeedback and behavior change. Complementary Therapies in Clinical Practice (2013)
11. Kido, T.: Self-Tracking Mindfulness Incorporating a Personal Genome. In: AAAI 2012 Spring Symposium (March 2012)
12. Crowley, K., Sliney, A., Pitt, I., Murphy, D.: Evaluating a brain-computer interface to categorise human emotional response. In: 2010 IEEE 10th International Conference on Advanced Learning Technologies (ICALT), pp. 276–278 (July 2010)
13. Chan, K., Mikami, K., Kondo, K.: Measuring interest in linear single player FPS games. In: ACM SIGGRAPH ASIA 2010 Sketches, p. 3. ACM (December 2010)
14. Funk, M., Glück, H., Pfleiderer, F.: Evaluation der Brain-Computer Interfaces (2012)
15. Bosse, T., Gerritsen, C., de Man, J., Stam, M.: Inducing Anxiety through Video Material. In: Proceedings of the 16th International Conference on Human-Computer Interaction, HCII 2014. Springer Verlag (2014)

Emotracking Digital Art

Isabelle Hupont[1], Eva Cerezo[2], Sandra Baldassarri[2], and Rafael Del-Hoyo[1]

[1] Multimedia Technologies Division, Aragon Institute of Technology, Zaragoza, Spain
{ihupont,rdelhoyo,dabadia}@ita.es
[2] GIGA AffectiveLab, University of Zaragoza, Zaragoza, Spain
{ecerezo,sandra}@unizar.es

Abstract. Art and emotions are intimately related. This work proposes the application to arts of Emotracker, a novel tool that mixes eye tracking technology and facial emotions detection to track user behaviour. This combination offers intuitive and highly visual possibilities of relating eye gaze, emotions and artistic contents. The results obtained after carrying out "5-second emotracking tests" over art illustrations and the use of the gathered information to create real-time artistic effects are presented.

Keywords: affect analysis, gaze, face analysis, digital arts.

1 Introduction

Affective Computing aims at developing intelligent systems able to provide a computer with the ability of recognizing, interpreting and processing human emotions [1]. Since the introduction of the term Affective Computing in the late 1990s, an increasing number of efforts towards automatic human affect extraction have been reported in the literature. Systems able to recognize human emotions from facial expressions, physiological signals, voice, text, etc. have been developed with high accuracy [2].

Independently of the channel -or channels- chosen to detect affect, most works still focus efforts on increasing the success rates in the emotion recognition task. However, other important issues have scarcely been studied, namely how to efficiently visualize the extracted affective information, how to process it to improve the user's experience in different applications or what is the best combination of channels depending on the information sought. In particular, the combination of user eye gaze and facial emotional information has been proved to have a great potential in measuring user perception, impact and/or engagement with digital contents [3].

One of the most subjective perceptual experiences is given by arts. Emotions and art are intimately related [4] and it is, perhaps, the unique and highly variable personal emotional perceptions elicited what makes art so attractive. The study of that perceptions require the interaction between art and science, two fields that, with few notable exceptions have grown in parallel with only counted interactions. In spite of the impact that the scientific study of art could have, it is somehow understandable that such enterprise is only starting to take off [5, 6, 7]. On the one hand, art perception is too subjective and challenging for rigorous scientific exploration. On the other hand,

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 739–749, 2014.

artists may fear that scientists could bring a misleading reductionism that would oversimplify all the aspects involved in the appreciation of art.

In an attempt to bridge these two fields, i.e., using scientific methods to study art, in this work we use Emotracker, a novel tool that mixes eye-tracking technology and facial emotions extraction to track user behaviour. This combination offers intuitive and highly visual possibilities of relating eye gaze, emotions and artistic contents. In particular, the results obtained after carrying out "5-second emotracking tests" over art illustrations and the use of the information gathered to create artistic effects will be shown.

The structure of the paper is the following. Section 2 analyzes the related state of the art. In section 3, the Emotracker tool is presented. Section 4 comprises the description of the "5-second emotracking tests", while section 5 focuses on real-time emotracking data-based artistic effects creation. Finally, in section 6 conclusions and future work are presented.

2 Background

This section explores issues related to the description of affective information, the analysis of eye-movements when looking to images and the use of user behavior data to artistically transform images.

2.1 Description of Affect

Despite the existence of various other models, the categorical and dimensional approaches are the most commonly used models for automatic analysis and prediction of affect.

The most long-standing way that affect has been described by psychologists is in terms of discrete categories, an approach that is rooted in the language of daily life. The most commonly used emotional categories are the six universal emotions proposed by Ekman [8] which include "happiness", "sadness", "fear", "anger", "disgust" and "surprise". The labeling scheme based on category is very intuitive and thus matches peoples' experience. However, human emotions are richer than simple emotional labels and may experiment strong complex variations over time. Those aspects of human affect (complexity and dynamics of emotions) should be captured and described by an ideal affect recognizer.

To overcome the problems cited above, some researchers, such as Whissell [9], Plutchik [10] or Russell [11], prefer to view affective states not independent of one another; rather, related to one another in a systematic manner. They consider emotions as a continuous 2D space whose dimensions are evaluation and activation. The evaluation (also called valence) dimension measures how a human feels, from positive to negative. The activation (also known as arousal) dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive. Unlike the categorical approach, the dimensional approach is attractive because it is able to deal with non-discrete emotions and variations in affective states over time. However, given its continuous (i.e. numerical) nature, the main drawback of this approach is that it does not offer an intuitive understanding of affective information, since people is used to report emotions by means of words.

2.2 Gathering Information from Images: Eye Movements

The analysis of user eye gaze is a fundamental part when studying the impact of an art piece. There are several basic facts about how people look at (unchanging) images [12] that come from human vision studies:

- People can examine only a small part of an image at one time, and so understand images by scanning them using discrete, rapid movements of their eyes, called saccades. While saccades can be initiated voluntarily, they typically proceed in a goal-directed fashion. The motions are performed with remarkable precision and efficiency -the eyes seldom perform wasted motions, and typically land near the best place to gather the desired visual information.
- Saccades are punctuated by stabilizing motions called fixations, which allow the eye to dwell on a particular stationary object. The overwhelming majority of visual processing takes place during fixations. Under normal circumstances, the attention of the viewer is at the fixation location, for at least the bulk of its duration.
- In each individual glance, people look at something -the eyes do not wander randomly.
- In most tasks, the time spent fixating on a particular location or object indicates that processing on that object is taking place. More specifically, fixation duration provides a rough estimate on how much processing is expended in understanding that portion of the image.
- Many other types of movements are possible, such as those involved in smooth pursuit; but it is the saccades and fixations that play the largest role in gathering information from across a static image.

2.3 Using User Behaviour to Create Artistic Image Effects

There are a couple of very interesting works coming from the painterly rendering domain that have been exploring the use of user interaction data to modulate the stylization of images. The ultimate objective of painterly rendering is to create non-photorealistic images by placing brush strokes according to some goals.

Shugrina et al. [13] introduce the term "empathic painting", an interactive painterly rendering whose appearance adapts in real-time to reflect the perceived emotional state of the viewer. They recognize users' emotions from their facial expressions detecting facial action units by applying computer vision techniques; the facial action units are mapped to vectors within the 2D valence-arousal space. Then, applying a non-photorealistic painterly rendering algorithm they generate the frames of painterly animation from a source photograph.

Santella and DeCarlo [12] propose a new approach for the creation of painterly renderings that drives on a model of human perception and is driven by eye-tracking data. The eye-tracking data is used to select and emphasize structures in the image that the user found important. They transform the original image by selecting those perceptual elements that people looked at extensively, using a model of people visual sensitivity. They display the image to the user, let the user look at the images several seconds and then perform meaningful abstraction based on the eye-tracking data (basically fixation points' data).

3 The Emotracker: A Tool for Advanced Human Affect Visualization

This section presents Emotracker, a novel and advanced visual tool able to dynamically relate eye gaze information, affect and contents.

3.1 System Description and Setup

Emotracker is based on the combination of an eye tracker and a facial emotions recognizer. It is built on the top of two commercial APIs we have been widely exploring in our user experience laboratory in the last years:

- Tobii Studio [14] is a software by Tobii© that offers tools for easily creating eye tracking tests and experiments, collecting eye gaze data and making graphical visualizations from them. It has an associated specific hardware, Tobii T60, which is a 17-inch TFT monitor with integrated IR diodes that enable the real-time detection of the user's pupil. The eye tracking process is unobtrusive, allowing natural and large degree of head movement, and any kind of ambient light conditions. Moreover, it doesn't lose robustness, accuracy and precision, regardless of a subject's ethnic background, age, use of glasses, etc.
- FaceReader [15] is a facial emotions recognition software by Noldus©. It is able to analyze in real-time the facial expressions of the user, captured by means of any ordinary webcam, and provide affective information both in categorical and dimensional description levels. FaceReader works with high accuracy and robustness, even in naturalistic settings with any kind of illumination and type of user.

Emotracker has been developed with the aim of going beyond traditional eye-tracking by indicating not only where the user is looking at, but also with which affective state. Figure 1 shows the Emotracker system's setup.

After a brief gaze calibration (see Figure 2), the user is allowed to interact with different types of contents: movies, pictures, web pages or applications. The output from this first analysis is: user's navigation information, user's gaze log and user's facial video. The latter video is then analysed by the FaceReader software to obtain a user's emotional log. This log, in addition to the three precedent ones, is then loaded into Emotracker, opening the door to advanced and meaningful visualizations as it is explained in the next section.

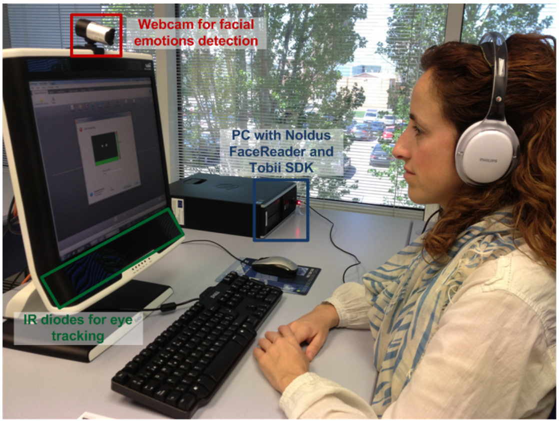

Webcam for facial emotions detection
PC with Noldus FaceReader and Tobii SDK
IR diodes for eye tracking

Fig. 1. Emotracker system setup

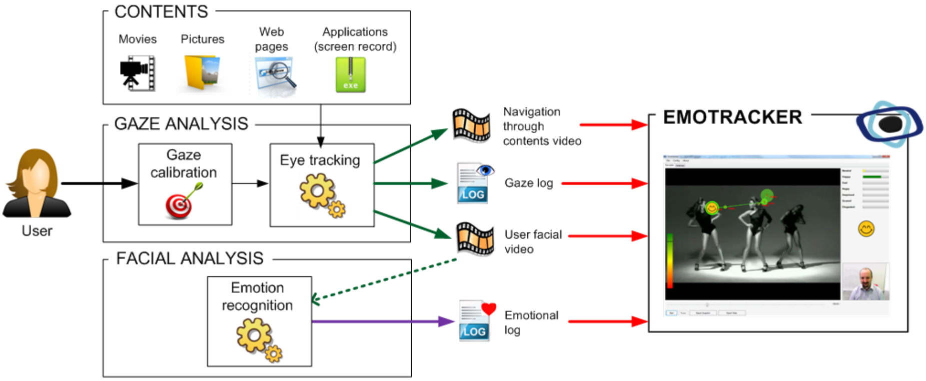


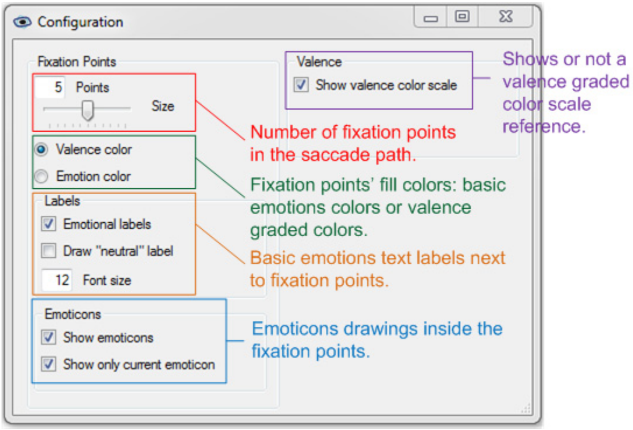
Fig. 2. Emotracker functioning: capturing user's gaze and emotional data

3.2 Emotracker Visualization Capabilities

Emotracker aims to visualize contents, gaze and emotional information at a glance, i.e. in an intuitive and clear way. To accomplish this, it builds visual reports in the form of "emotional heat maps" and "emotional saccade maps".

The "emotional heat map" is a direct unprocessed representation of the user's gaze data (of both eyes), enhanced with the possibility of working with "emotional layers". Each "emotional layer" represents the gaze data associated to a specific basic emotion, so that if a given "emotional layer" is selected only the gaze data associated to this emotion is shown and painted with its corresponding colour. If all the "emotional layers" are selected, the gaze data is filled-in with the colour of the most dominant emotion. This representation is particularly useful when checking whether a given content has elicited a particular emotion (even if non-dominant). The "emotional saccade map" is a dynamic processed representation of gaze data that shows the path formed by the user fixation points: a fixation point is a point the user has been looking at for a minimum amount of time (in milliseconds, configurable).

The initial visualization configuration panel of the tool can be seen in Figure 3, while several examples of the visualization capabilities results obtained with the tool are shown in Figure 4. The main potential of tool is its wide range of customizable representation possibilities. Its interface allows to activate and deactivate different visualization options, both for “emotional saccade maps” and “emotional heat maps”, such as: interest points numerical labels, discrete emotions text labels, drawing smileys inside the fixation points, discrete emotions coloured zones, valence graded colours, etc.



Fixation Points
5 Points
Size
● Valence color
○ Emotion color
Labels
☑ Emotional labels
□ Draw "neutral" label
12 Font size
Emoticons
☑ Show emoticons
☑ Show only current emoticon
Valence
☑ Show valence color scale

Shows or not a valence graded color scale reference.

Number of fixation points in the saccade path.

Fixation points’ fill colors: basic emotions colors or valence graded colors.

Basic emotions text labels next to fixation points.

Emoticons drawings inside the fixation points.

Fig. 3. Emotracker visualization configuration panel

3.3 Validation Issues

The Emotracker tool was validated through a pilot study with 14 naïve users exhaustively detailed in [3]. The objective of the study was investigating whether the emotional and gaze results visualized with the tool were similar with users’ perception. Specifically, the users showed several emotional video sequences in the Emotracker device and then visualized their subsequent emotracking results. They were asked to classify the following statements between 1 (strongly disagree) and 5 (strongly agree):

- The information about the visual fixation points provided by Emotracker correctly represents the path followed by my gaze in the videos (“gaze accuracy”).
- The emotional information provided by Emotracker correctly represents the emotions I felt watching the videos (“emotions accuracy”).
- Emotracker’s results are easy to understand (“intuitiveness”).
- The visualization of the results presented by Emotracker is enough and appropriate (“visualization”).
- During the emotracking session I could forget that I was being filmed and my behaviour was natural (“natural behaviour”).

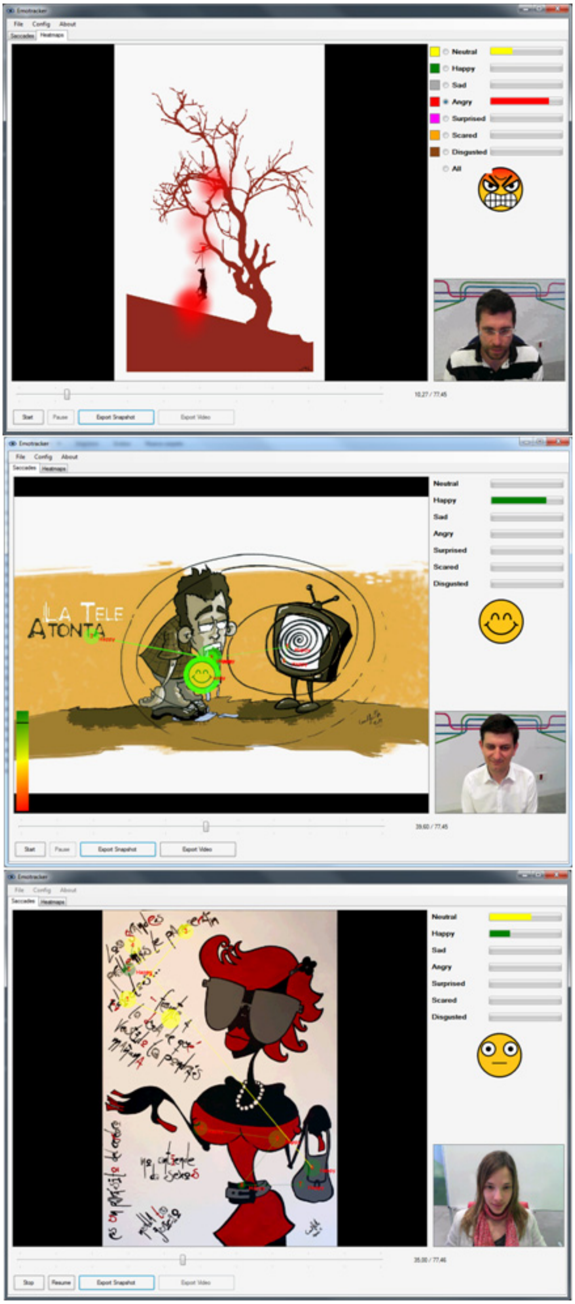


Fig. 4. Visualization capabilities of the Emotracker. Snapshots taken from "5-second emotracking tests" over artistic contents. Top: "emotional heat map" with the "angry layer" selected. Middle: "emotional saccade map" with valence colored fixation points, emoticons and emotions text labels. Bottom: "emotional saccade map" with basic emotions colored fixation points and emotions text labels. Illustrations by Rakel Goodféith.

Figure 5 summarizes the mean scores obtained for each statement. As can be seen, the intuitiveness and visualization capabilities of the tool really satisfied the users. Regarding the accuracy in gaze and emotion detection, users confirm that the visualization of the results in gaze detection are very accurate. However, the results regarding emotion detection vary a little: a greater dispersion in scores appears depending on each user and video sequence type (e.g. terror/tragedy/comedy clips). Finally, it is interesting to point out that most users couldn't completely forget they were being filmed while performing the tests.

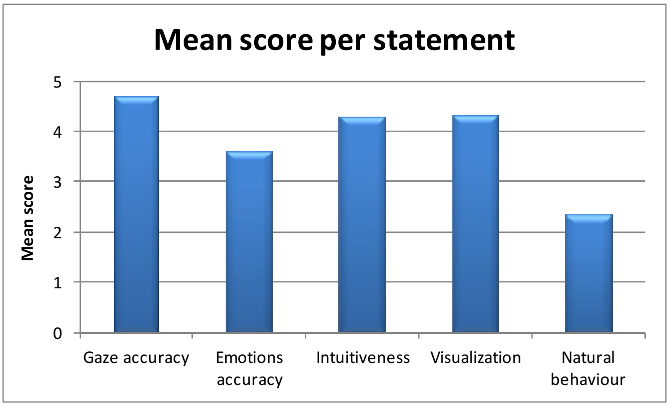


Fig. 5. Mean scores obtained in the pilot user study per evaluated statement

4 "5-Second Emotracking Tests" for Artistic Contents Impact Measurement

One of the most popular usability testing techniques is the so-called "5-second tests" [16]. As the name suggests, the "5-second test" involves showing users a single content, image or page-design for a quick 5 seconds to gather their initial -and therefore more salient- impressions. Five seconds may not seem like a lot of time, but users make important judgments in the first moments they see a content or visit a page. This technique has been traditionally mostly used for websites' usability analysis and users' judgments have generally been collected by directly asking them to write down everything they remember about the page. Eye trackers are a perfect tool to make this process more automatic, measurable and objective.

For that reason, and with the added value of affective information, we have used Emotracker to perform what we call "5-second emotracking tests". We presented a slideshow of 10 illustrations by the Spanish artist Rakel Goodféith to 5 naïve users and measured their reactions with Emotracker. The slideshow was organised so that each painting was shown during 5 seconds, and 10 seconds with black screen lapsed between two different drawings. Some snapshots of the "5-second emotracking tests" are presented in Figure 4.

Emotracking results where then shown to the artist who reported to "find the visualization very helpful and intuitive for understanding the first affective and visual impact of her illustrations on the users". She has also stated that she would be glad to reuse Emotracker, even in earlier stages of the creative process to predict the future impact of her artistic contents.

5 EmotrackingRT: Real-Time Affective Digital Art

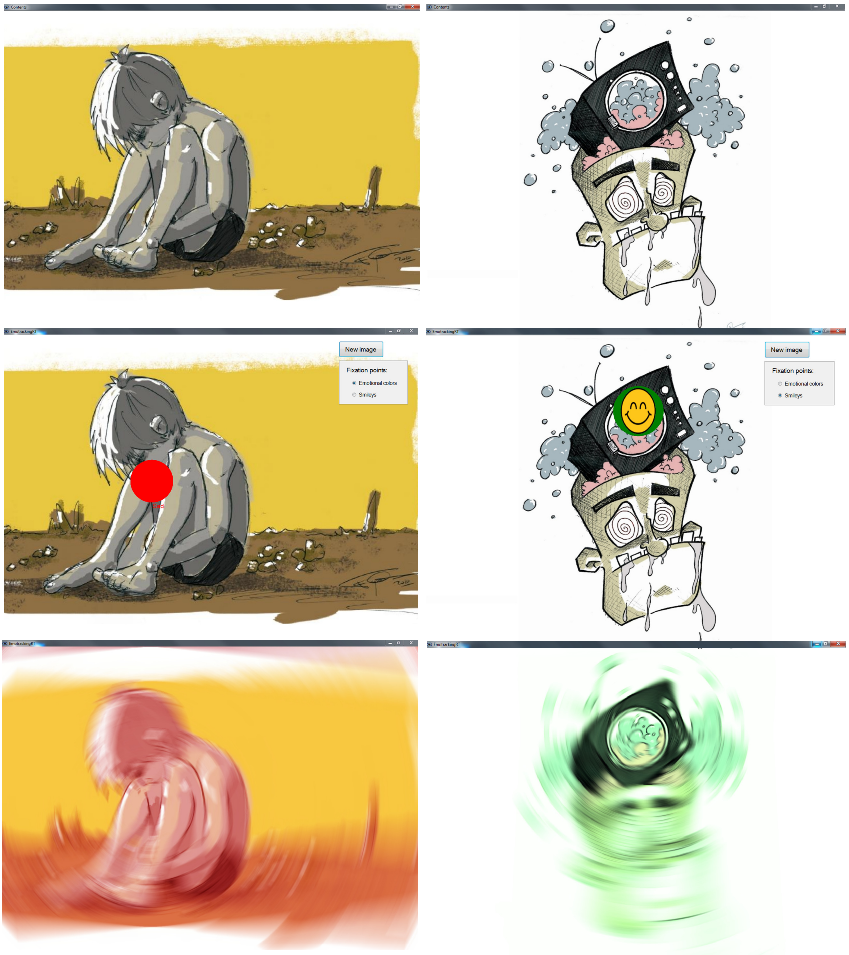
As the initial aim when developing Emotracker has been to provide professionals (raging from marketing to psychology) with a flexible, intuitive tool to analyse the combined information of user's gaze and emotions when interacting with different kinds of content, special focus has not been put in achieving complete on-line real-time processing. But for other types of applications, such as arts, it could be interesting and it is perfectly possible.

In fact we have developed a real-time demo version of Emotracker, called EmotrackingRT, making use of Tobii SDK and FaceReader API. The demo does not account for all the types of visualization options the original off-line Emotracker has, but has been put into operation successfully. EmotrackingRT runs on a single PC with two output displays: the Tobii 17-inch TFT monitor, where the artistic contents themselves are shown to the users and their gaze is tracked, and a second standard monitor where emotracking information is visualized in real-time. In the latter screen, time-growing fixation points with associated emoticons or/and emotional colours are painted and a "new image" button is enabled to switch the artistic contents the user is viewing (Figure 6, middle).

This real-time functioning opens the door to new types of applications based, for example, in the interaction with digital illustrations that adapt colours, apply image filters or make any other kind of artistic effects depending on the user's gaze and emotional data obtained from our tool. To show its potential, we have added to the EmotrackingRT demo a third window, where an artistic effect is applied in real-time to the image being shown by the user: a radial motion blur effect is spread from the current fixation point and the global RGB histogram values of the image are modified depending on the current emotion colour (Figure 6, bottom).

6 Conclusions and Future Work

Art and affect are inherently related. This paper proposes the application of the tool Emotracker to the field of arts. Emotracker is a novel system based on the combination of a facial emotional recognizer and eye tracking technology that allows to plot gaze, emotions and contents in a single map. In this work we have successfully used Emotracker to carry out "5-second emotracking tests" over illustrations and to create real-time affective artistic effects depending on the current user's visual and emotional data. In a near future, we expect to acquire a wider range hardware-independent eye tracking system in order to analyze any kind of artistic contents (e.g. big size real paintings in museums) and create real-time digital interactive emotional art.



Fixation points:
● Emotional colors
○ Smileys

Fixation points:
○ Emotional colors
● Smileys

Fig. 6. EmotrackingRT results. Up: original illustrations by Rakel Goodféith the users are watching in the Tobii 17” TFT monitor. Middle: gaze and emotional information captured and displayed in real-time in a second standard monitor. Bottom: third window where an artistic radial motion blur effect spread from the current fixation point with global RGB histogram modified depending on the current emotion colour is applied.

Acknowledgments. This work has been partly financed by the Spanish “Dirección General de Investigación", contract number TIN2011-24660, by the SISTRONIC group of the Aragon Institute of Technology, by the FEDER Operative Programme for Aragon (2007-2013), and by the European Celtic project QuEEN, contract number IPT-2011-1235-430000. The authors also want to thank Rakel Goodféith for her illustrations (`http://goodfeith.blogspot.com`).

References

1. Picard, R.W.: Affective Computing. The MIT Press (1997)
2. Calvo, R., D'Mello, S.: Affect Detection: An Interdisciplinary Review of Models, Methods and their Applications. IEEE Transactions on Affective Computing 1(1), 18–37 (2010)
3. Hupont, I., Baldassarri, S., Cerezo, E., Del-Hoyo, R.: The Emotracker - Visualizing Contents, Gaze and Emotions at a Glance. In: 5th International Workshop on Affective Interaction in Natural Environments (AFFINE 2013), Geneva, Switzerland (2013) (in press)
4. Tan, E.S.: Emotion, Art, and the Humanities. In: Lewis, M., Haviland-Jones, J.M. (eds.) Handbook of emotions, 2nd edn., pp. 116–134. Guilford Press, New York (2000)
5. Cavanagh, P.: The Artist as a Neuroscientist. Nature 434(7031), 301–307 (2005)
6. Silvia, P.J.: Emotional Responses to Art: From Collation and Arousal to Cognition and Emotion. Review of General Psychology 9, 342–357 (2005)
7. Quiroga, R., Pedreira, C.: How do we See Art: An Eye-Tracker Study. Frontiers in Human Neuroscience 5(98) (2011)
8. Ekman, P., Freisen, W., Ancoli, S.: Facial Signs of Emotional Experience. J. Personality and Social Psychology 39(6), 1125–1134 (1980)
9. Whissell, C.M.: The Dictionary of Affect in Language. In: Emotion: Theory, Research and Experience, vol. 4. Academic (1989)
10. Plutchik, R.: Emotion: A Psychoevolutionary Synthesis. Harper & Row (1980)
11. Russell, J.A.: A Circumplex Model of Affect. J. Pers. Soc. Psychol. 39, 1161–1178 (1980)
12. Santella, A., DeCarlo, D.: Abstracted Painterly Renderings Using Eye-Tracking Data. En. In: Proceedings of the 2nd International Symposium on Non-Photorealistic Animation and Rendering (2002)
13. Shugrina, M., Betke, M., Collomosse, J.: Empathic Painting Interactive Stylization Using Observed Emotional State. In: Proceedings 4th International Symposium on Non-Photorealistic Rendering and Animation (NPAR 2006), pp. 87–96 (2006)
14. Tobii T60, `http://www.tobii.com/en/eye-tracking-research/global/products/hardware/tobii-t60t120-eye-tracker/`
15. Den Uyl, M.J., Van Kuilenburg, H.: The FaceReader: Online Facial Expression Recognition. In: Proceedings of Measuring Behavior, pp. 589–590 (2008)
16. Shari, T., Musica, N.: When Search Meets Web Usability. New Riders (2009)

Estimation of Emotion by Electroencephalography for Music Therapy

Kensyo Kinugasa, Hiroki Yoshimura, Maiya Hori, Satoru Kishida, and Yoshio Iwai

Graduate School of Engineering, Tottori University, 101 Minami 4-chome, Koyama-cho,
Tottori 680-8550, Japan
{s082017,yosimura,hori,kyososhi-fkg,iwai}@ike.tottori-u.ac.jp

Abstract. A system for providing music employing electroencephalography for music therapy is described. Music therapy for the treatment of patients suffering mental illness has been attempted over a period of 20 years. To reduce stress, it is preferable to listen to music that matches a person's emotions. However, it is difficult to know exactly the person's emotion. It is necessary to calibrate the proposed system employing electroencephalography to emotions. We discuss a method of calibration especially used in canonical correlation analysis. Experimental results show that it is possible to roughly estimate feelings. We consider that it is possible to use our system in practice.

Keywords: electroencephalography, music therapy, canonical correlation analysis.

1 Introduction

In recent years, the media has focused on diseases such as depression and mental illness that result from stress in everyday life. It is empirically known that listening to music relaxes and heals the weary body and mind. Additionally, stress dissipates through a feeling of being uplifted. Music therapy that targets patients suffering from mental illness has been attempted for a period of 20 years. To reduce stress, it is preferable for a person to listen to music that matches their emotion [1]–[4]. However, it is difficult to know a person's emotion exactly.

In this study, to solve the above problem, we propose a system that provides music according to the results of electroencephalography. To estimate an emotion employing electroencephalography, it is necessary to perform a calibration. We discuss a method of calibration especially used in canonical correlation analysis.

2 System for Providing Music for Therapy Using Electroencephalography

2.1 Overview of the Music Therapy System

Figure 1 is an overview of the target system in this study. The system is designed to assist in healing and music therapy by providing the song that best matches the

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 750–757, 2014.

emotion of a subject estimated by electroencephalography. Figure 2 shows the flow of processing. First, an electroencephalogram (EEG) is recorded to examine the electrical activity of the brain for a subject. The EEG obtained is analyzed to estimate the emotion. The subject is provided with music that corresponds to the estimated emotion. This cycle is repeated.

The problem here is how to estimate the emotion from the EEG. First, there is a need to calibrate the relationship between the emotion and EEG for the system. Calibration requires an index expressing numerical emotions. We use the V–A plane of Russell as an indicator of emotion as shown in fig. 3. Emotion is expressed by an arousal value (vertical axis) and valence value (horizontal axis). This is described further in Section 2.2.

In this research, the subject listens to music for which the V–A value is known. At the same time, the EEG is recorded. By analyzing these data, a mapping of emotion and the EEG is realized. Canonical correlation analysis is employed. The estimation of emotion through this analysis is described in Chapter 2.5.

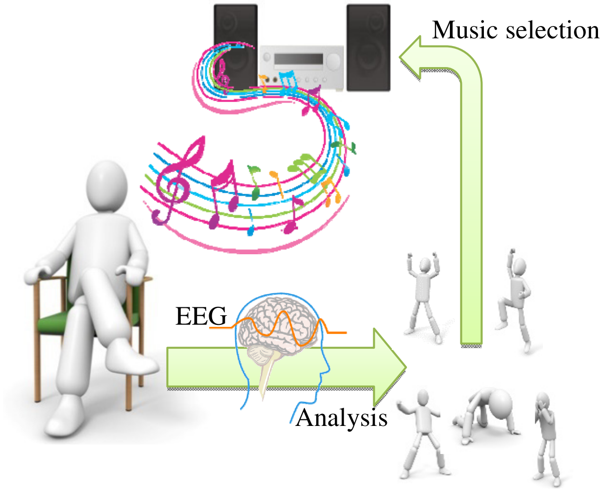


Fig. 1. Overview of our system of providing music employing electroencephalography for music therapy

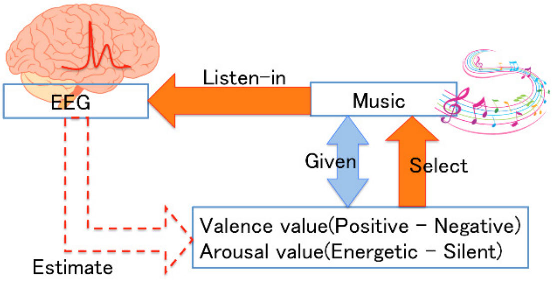

Listen-in
EEG
Music
Given
Select
Valence value(Positive – Negative)
Arousal value(Energetic – Silent)
Estimate

Fig. 2. Processing flow

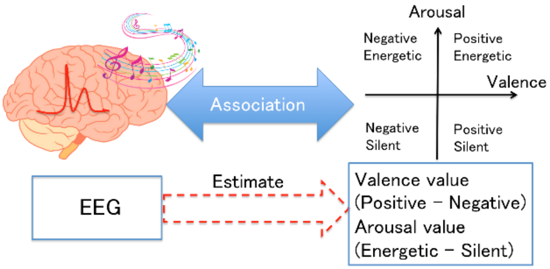


Fig. 3. Emotion indicator

2.2 V–A plane of Russell

Russell defined emotion on two axes of an arousal value and valence value. Figure 4 shows the V–A plane [5]. Emotions are classified into four quadrants generated by the two axes as shown in the figure. The pleasant-aroused quadrant includes emotions such as excitement and joy, the unpleasant-aroused quadrant emotions such as worry and anger, the unpleasant-unaroused quadrant emotions such as melancholy and sadness, and the pleasant-unaroused quadrant emotions such as satisfaction and relaxation.

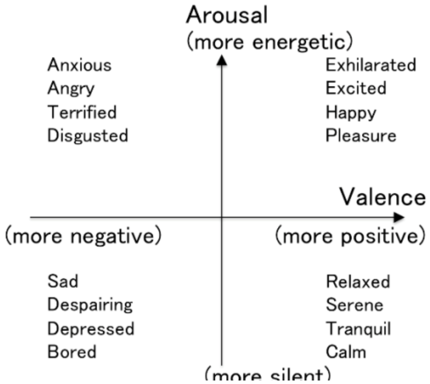


Fig. 4. V–A plane of Russell

2.3 Music Used for Emotion Occurrence

Correlation of the emotion estimated from the EEG requires indicators of emotion evoked by the music, which are discussed below. We used MoodSwing Lite Music as a music database [6]. The database comprises arousal and valence values for 10 to 20 subjects listening to music as determined by a subjective questionnaire. The values are given for 15-second sections of approximately 240 songs. The present study took 20 songs—five songs per quadrant—from the database as shown in fig. 5.

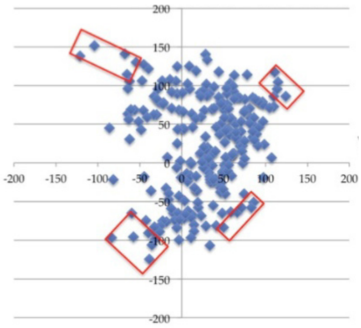


Fig. 5. Arousal and valence values when listening to music from MoodSwing

2.4 EEG data

The EEG was obtained using 14 electrode poles (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) employing the internationally standardized 10-20 system. The EEG was divided into four bands using a band-pass filter. The four bands were 1–4 Hz (δ wave), 4–7 Hz (θ wave), 7–13 Hz (α wave) and 13–30 Hz (β wave). We took the set of values of the average power per second of each band.

2.5 Emotion Estimation Employing Canonical Correlation Analysis

Canonical correlation analysis is a multivariate technique of determining the relationships between groups of variables in a data set. The data set is split into two groups, let us say groups U and V, according to common characteristics. The purpose of canonical correlation analysis is then to find the relationship between U and V; i.e., we ask whether some form of U can represent V. In this study, U is the EEG that is divided into four bands, and V is the V–A values of the music.

Formula (1) is obtained by canonical correlation analysis.

$$\begin{cases} s_1(t) = \sum_{m=1}^{M} a_m u_m(t) \\ s_1^{'}(t) = \begin{pmatrix} d_{Valence} & d_{Arousal} \end{pmatrix} \begin{pmatrix} v_{Valence}(t) \\ v_{Arousal}(t) \end{pmatrix} \end{cases}$$

$$\begin{cases} s_2(t) = \sum_{m=1}^{M} b_m u_m(t) \\ s_2^{'}(t) = \begin{pmatrix} e_{Valence} & e_{Arousal} \end{pmatrix} \begin{pmatrix} v_{Valence}(t) \\ v_{Arousal}(t) \end{pmatrix} \end{cases} \quad (1)$$

$M = E \times W$ ($E = 14$: number of electrodes, $W = 4$: number of bands)

Here a_m represents the linear combination coefficients of the first canonical correlation score on brain waves (14 poles × 4 bands), b_m represents the linear combination coefficients of the second canonical correlation score on brain waves (14 poles × 4 bands), d_{Valence} and d_{Arousal} are the linear combination coefficients of the first canonical correlation score on V-A values of music, $u_{\text{m}}(t)$ is the potential of brain waves at time t, and $v_{\text{m}}(t)$ represents the V-A values of music at time t. Additionally, from the canonical correlation analysis, canonical correlation coefficients C_1 and C_2 are obtained. The relationship of the canonical correlation coefficients and canonical variable scores is expressed by equation (2).

$$\begin{cases} s_1(t) = C_1 s_1^{'}(t) \\ s_2(t) = C_2 s_2^{'}(t) \end{cases}$$

$$(Corr(s_1(t), s_1^{'}(t)) \approx C_1, Corr(s_2(t), s_2^{'}(t)) \approx C_2) \tag{2}$$

v_{Valence} corresponding to the valence value and v_{Arousal} corresponding to the arousal value are estimated using equations (1) and (2).

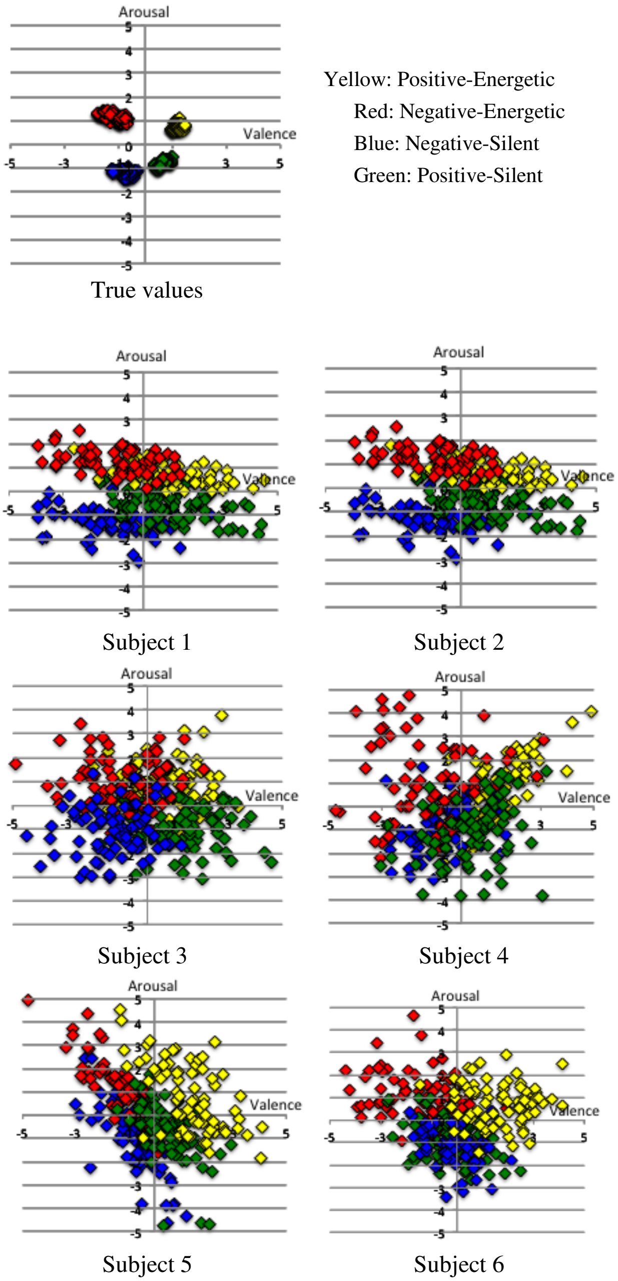
3 Experiment

Estimation performance was examined to calibrate the emotion estimation system employing canonical correlation analysis. An experiment was carried out on 10 healthy men and women (eight males and two females aged 21–25 years). Subjects listened to music with known V–A values for 15 seconds after silence for a period of 15 seconds. This was performed as many as 20 times, while the EEG was measured. The subjects had their eyes closed and wore an eye mask. We used the Emotiv EPOC manufactured by Emotiv Corporation as an EEG measuring device.

Table 2 gives the first canonical correlation coefficient, second canonical correlation coefficient and estimation error for the 10 subjects. The correlation canonical coefficients are high; in particular, the first correlation canonical coefficient is at least 0.7. The coefficients are strongly related to the EEG and emotions that are evoked by music are shown.

Table 1. First canonical correlation coefficient, second canonical correlation coefficient and estimated error

	First canonical correlation coefficient C_1	Second canonical correlation coefficient C_2	Estimation error
Subject 1	0.89	0.55	1.32
Subject 2	0.73	0.55	1.50
Subject 3	0.72	0.60	1.46
Subject 4	0.66	0.53	1.65
Subject 5	0.81	0.54	1.41
Subject 6	0.73	0.60	1.43
Subject 7	0.79	0.54	1.48
Subject 8	0.79	0.58	1.40
Subject 9	0.78	0.55	1.40
Subject 10	0.72	0.64	1.33
Average	0.76	0.57	1.44



True values

Subject 1

Subject 2

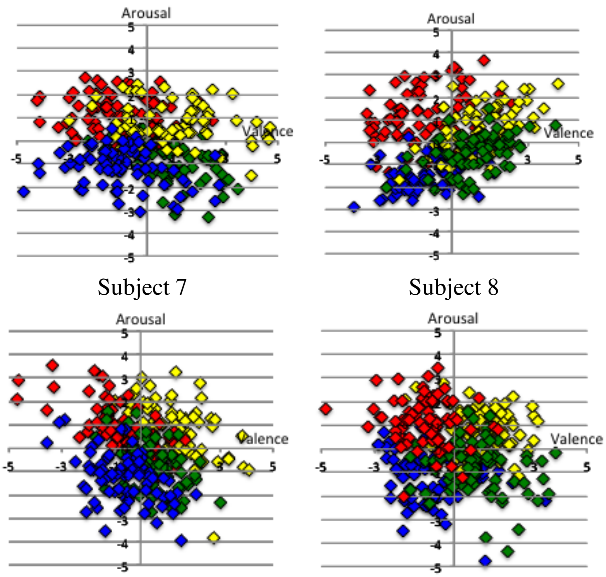
Subject 3

Subject 4

Subject 5

Subject 6

Fig. 6. True values and distributions of estimated values



Subject 7

Subject 8

Subject 9

Subject 10

Fig. 6. (*Continued*)

Figure 6 shows the distributions of the estimated values and the true values. Estimations of V–A values were more widely distributed than the true values. This is because they are approximated directly using the approximate straight line obtained from the canonical correlation analysis. We thus expect the generalization performance to fall. Therefore, for the estimation of 15-second segments of music that were not used in the experiment, which quadrants the segments fall into is determined by majority decision. Table 2 presents the estimation results. Music is presented for only 15 seconds. The accuracy rate was about 60% overall.

Table 2. The estimation results of whether the music is which quadrant by majority decision

Estimation Presentation	Positive - Energetic	Negative - Energetic	Negative - Silent	Positive – Silent
Positive - Energetic	67	15	9	9
Negative - Energetic	20	55	15	10
Negative - Silent	5	7	68	20
Positive - Silent	12	16	27	45

4 Conclusion

In this study, we proposed a system for providing music for therapy. The system estimates emotion from an EEG, and presents the music that best matches the emotion. Because of differences among individuals, the EEG needs to be calibrated for each individual. Experiments on calibration were carried out, and it was found to be possible to roughly estimate feelings. We consider that our system can be applied in practice.

References

1. Zillmann, D.: Mood Management: Using Entertainment to Full Advantage. In: Donohew, L., Sypher, H.E., Higgins, E.T. (eds.) Communication, Social Cognition, and Affect, pp. 147–171. Lawrence Elbaum, New Jersey (1988)
2. Konecni, V.J., Crozier, J.B., Doob, A.N.: Anger and Expression of Aggression: Effects on Aesthetic Preference. Scientific Aesthetics Sciences de l'Art 1, 47–55 (1976)
3. Arnett, J.J.: Adolescents and Heavy Metal Music: From Mouth to Metal Heads. Youth and Society 23, 76–98 (1991)
4. Arnett, J.J.: Metal Heads: Heavy Metal Music and Adolescent Alienation. Westview, Oxford (1995)
5. Russell, J.A.: A Circumflex Model of Affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
6. Kim, Y.E., Schmidt, E.M., Emelle, L.: Moodswings: A Collaborative Game for Music Mood Label Collection. In: ISMIR, pp. 231–236 (2008)

Evaluating User's Emotional Experience in HCI: The PhysiOBS Approach

Alexandros Liapis, Nikos Karousos, Christos Katsanos, and Michalis Xenos

Hellenic Open University, School of Science and Technology
Parodos Aristotelous 18, Patra Greece, 26 335
{aliapis,karousos,ckatsanos,xenos}@eap.gr

Abstract. As computing is changing parameters, apart from effectiveness and efficiency in human-computer interaction, such as emotion have become more relevant than before. In this paper, a new tool-based evaluation approach of user's emotional experience during human-computer interaction is presented. The proposed approach combines user's physiological signals, observation data and self-reported data in an innovative tool (PhysiOBS) that allows continuous and multiple emotional states analysis. To the best of our knowledge, such an approach that effectively combines all these user-generated data in the context of user's emotional experience evaluation does not exist. Results from a preliminary evaluation study of the tool were rather encouraging revealing that the proposed approach can provide valuable insights to user experience practitioners.

Keywords: User Emotional Experience, Human Computer Interaction, Evaluation, Physiological Signals, Emotions.

1 Introduction

People's daily interaction with technology, including personal computers, tablets, and mobile phones, has increased the need for usability. Although the available technology is rather mature, interaction with it can still be frustrating [1-2]. Thus, evaluating and designing for user emotional experience (UEX) is growing in importance.

So far usability evaluation studies are mainly focusing on task-related metrics, such as task success rate and time-on-task. Such metrics are an important indicator of users' performance, but lack in qualitative insight about other factors of user experience [3], such as emotions. Questionnaires, interviews and video analysis can provide such qualitative data, but these methods are time consuming and prone to subjectivity [4-6]. In an attempt to address these limitations, new and innovative approaches such as facial expression recognition [7] and speech tone and keystroke analysis [8-9] have been introduced. Towards the same direction, collecting and analyzing data from users' physiology (e.g. heart rate, respiration, skin conductance) is also a promising approach. Physiological signals are directly connected with emotions [10] and their study could result in the establishment of new user-centered design techniques.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 758–767, 2014.

Emotions recognition and analysis (Fig. 1) are gaining interest in the human-computer interaction (HCI) domain, and especially in usability evaluation studies [11]. Existing approaches to interpret physiological signals in terms of emotions suffer from two important limitations. First, the recognition success of existing physiological signals datasets [12-13] used for emotion analysis relies on contexts that induce intense emotions, such as watching a scary movie, listening to a favorite song, major hardware failures and gaming. However, identification of subtle emotions is of more interest in typical HCI tasks and remains an open research topic. Additional, existing approaches in the HCI domain, attempt to recognize one or two emotions [14-17], thus ignoring any other emotions that users may have felt during an interaction session, which might lead in serious misunderstandings of UEX. Mandryk and Atkins [18] have proposed a psychophysiological method that can continuously monitor and also recognize more than one emotional state. However, it targets a specific domain (i.e. gaming and entertainment) and its application remains rather challenging for a practitioner.

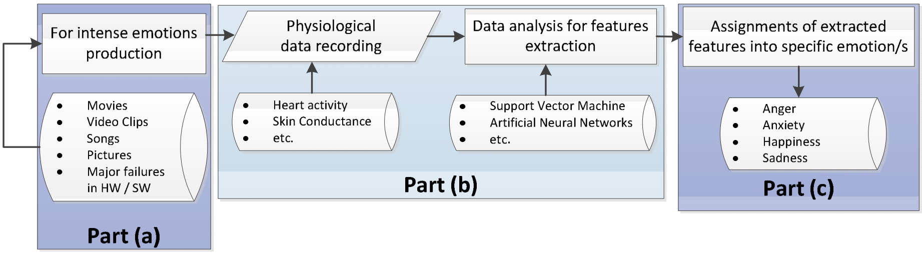

For intense emotions production
Movies
Video Clips
Songs
Pictures
Major failures in HW / SW
Part (a)
Physiological data recording
Heart activity
Skin Conductance
etc.
Data analysis for features extraction
Support Vector Machine
Artificial Neural Networks
etc.
Part (b)
Assignments of extracted features into specific emotion/s
Anger
Anxiety
Happiness
Sadness
Part (c)

Fig. 1. Interpretation of physiological signals into emotions. (a) emotions induction, (b) physiological data recording and analysis, (c) interpretation of extracted features into emotions.

In this paper, a new tool-based evaluation approach for evaluating UEX during HCI is presented. The proposed approach combines user's physiological signals (e.g. heart rate, blood volume pressure, skin conductance), observation data (e.g. users' face recording, screen recording) and self-reported data (e.g. responses in questionnaires, interviews) in an innovative tool (PhysiOBS) that allows continuous and multiple emotional states analysis. To the best of our knowledge, such an approach that effectively combines all these user-generated data in the context of UEX evaluation does not exist. PhysiOBS supports continuous and in-depth evaluation of UEX in a straightforward and easy way. It also combines multiple data sources for both subtle and intense emotions recognition.

The rest of the paper is structured as follows. First, a brief background on physiological signals and emotion analysis is provided. Next, the proposed tool-based approach is delineated along with a presentation of research papers that mention the need for such approach in the HCI domain. In addition, results from a preliminary study in which practitioners used PhysiOBS to evaluate UEX are also presented. The paper concludes with a discussion of the implications of the presented work and directions for future research.

2 Background: Physiological Signals and Emotions

Changes in both the external and internal of a user's body can be measured through physiological signals. Physiological measurements along with other traditional metrics such as questionnaires and interviews have been used in many HCI studies [15-17] offering a new perspective in usability evaluation.

This section describes the advantages and disadvantages of physiological signals along with their relation with emotions and emotion structure theories.

2.1 Advantages and Disadvantages of Physiological Signals

Physiological signals are derived from vital organs, such as the heart and brain. Some of the most-widely used physiological signals are the following:

1. **Heart rate:** measures the electrical activity of the heart;
2. **Skin conductivity** (Sweat)**:** measures the resistance of the skin and it is one of the most well-studied physiological signals in the literature;
3. **Muscle tension:** measures the electrical activity generated by muscles;
4. **Respiration rate:** measures the stretch amount of a person's chest. It is a metric that needs to be treated carefully because it can be affected by cardiac function.

Special and sophisticated sensors systems (e.g. Electroencephalograph, Galvanometer and Electrocardiograph) have been developed in order to support researchers and practitioners in both data acquisition and analysis. In addition characteristics such as objectivity, multidimensionality, unobtrusiveness and continuity [19-22] reinforced the use of physiological signals and made them a valuable asset for usability studies.

However, physiological measurements have some limitations [23]. First, data acquisition depends on specialized and costly devices. Second, physiological measurements can be noisy because of various external factors such as fluctuations in room temperature, user's general health condition and environment humidity levels. Application of filters can alleviate such issues, but only to a certain extent. Finally, the experimental conditions along with sensors attachment can also affect users' physiological signals.

2.2 Emotion Theories and Physiological Signals

The human body is a complex system that reacts to various external stimuli. These stimuli can affect a person's psychology causing a variety of emotional states such as happiness, enthusiasm, frustration or boredom.

William James and Carle Lange theory, known as the "James-Lange theory of emotion", refers to emotions as an interpretation of a psychological state which can be identified through physiological signals [24]. According to this theory, an external stimulus leads to a physiological reaction. The psychological reaction depends on how one interprets these physical reactions. For instance, a walk in the woods and an unexpected encounter with a wild animal can increase one's heart beats and trigger a body tremble reaction. In James-Lange theory, interpretation of physical reactions would conclude that the person is afraid "I am trembling, therefore I am afraid". To this direction [25-26] were the first who studied the relations between physiological signals and emotions, concluding to four types of relations:

1. **one-to-one relation:** one physiological signal is capable to define a unique emotion;
2. **many-to-one relation:** more than one physiological signals are needed in order to define an emotion;
3. **one-to-many relation:** a physiological signal is associated with more than one emotions;
4. **many-to-many relation:** several physiological signals are associated with several emotions.

So far, the last relation is the most plausible and has been adopted by several scientific domains such as HCI.

James-Lange theory of emotion was questioned by [27], who proposed the "two-factor theory of emotion". According to the latter, emotions are neither purely physical nor purely cognitive reactions, but a combination of both. The theory posits that physical reactions must be interpreted along with the situation that someone is facing. Therefore, a fast pounding heart could be interpreted as anxiety, if someone is taking part in exams and as fear if someone encounters a wild animal. To the same direction, our tool-based approach offers to evaluators four views: user's screen capture, user's face recording, user's physiological signals and user's self-reported data. Having available all these perspectives at the same time, UEX evaluation may be more reliable than considering only physical reactions.

2.3 Emotion Structure Approaches

Two main theories-approaches of emotion structure have been established in the literature. A discrete approach supported by Ekman [28] and a dimensional approach supported by [29]. Ekman's approach uses six discrete emotional states: anger, fear, sadness, enjoyment, disgust and surprise. These emotional states can be recognized in all cultures and are gender-independent. By contrast, in the approach proposed in [29], emotions can be characterized using two dimensions: Valence and Arousal. Facial muscular activity and unsolved tasks (high arousal – low valence) have been found to be negatively correlated [30], and this fostered the use of the Valence-Arousal space in emotion analysis. Finally, it should be noted that both approaches are still used by researchers and practitioners, forming two schools of thought. Thus, our tool-based approach takes into consideration both emotion structures theories, offering appropriate supportive mechanisms to evaluators.

3 The PhysiOBS Approach

In this section, a new tool-based evaluation approach of UEX during human-computer interaction is presented. A researcher can combine user's physiological signals (e.g. heart rate, blood volume pressure, skin conductance), observation data (e.g. users' face recording, screen recording) and self-reported data (e.g. questionnaires, interviews) through an innovative tool (PhysiOBS) that allows continuous and multiple emotional states analysis. PhysiOBS supports continuous and in-depth evaluation of UEX in a straightforward and easy way. It also combines multiple data sources, for both subtle and intense emotions recognition.

3.1 The Need for an Emotion Oriented Evaluation Tool

Approaching physiological signals from the perspective of an additional evaluation parameter, Wilson and Sasse [16] used them in order to assess video and audio quality of multimedia conferencing (MMC) systems. Their evaluation approach used three dimensions a) stress (user cost), b) satisfaction and c) performance. The weightiness of each dimension depends on the purpose of MMC use. Physiological measurements analysis revealed that physiological responses can be detected even in degradation of both video and audio quality.

Lin et al. [14] used physiological signals for stress detection and correlated them with traditional usability metrics. Experiment participants' were instructed to complete three stages of a video game as quickly as possible and with a minimum number of mistakes. Each stage offered a different difficulty level. Data analysis revealed a positive correlation between physiological signals and game difficulty.

Ward and Marden [15] examined whether physiological measurements can be used instead of traditional metrics in web usability studies. In a between-subject study, two groups of users performed a website navigation scenario. The first group navigated in a well-designed website and the second one in an extremely bad-designed website. During their navigation, three physiological signals (skin conductivity, blood volume pressure, and heart rate) were recorded. Results didn't reveal significant differences between the two groups, but did reveal differences between individuals.

Along the same direction, [31] related physiological signals to traditional metrics used in web usability evaluation. In a within-subject study, a group of 42 subjects took part in a website navigation scenario. In one trial users had at their disposal navigation help from an artificial face in cases of navigation problems, whereas in the other trial they weren't provided with this help. Physiological measurements from participants that used the artificial face didn't reveal any significant variations.

A common characteristic of all these studies is the use of physiological signals in combination with other methods, such as questionnaires, interviews and video analysis, to collect data about the user experience. To this end, a holistic approach that can combine and support analysis of these user-generated data in an effective way would be of great value for practitioners.

In a different perspective, [32] used physiological signals in a domain called "Affective computing". Affective computing can create new ways of communication between humans and machines, by enabling machines to respond to users' emotions. To this end, users' emotion recognition is a prerequisite. Piccard et al. [32] achieved to recognize eight emotions with high levels of accuracy using physiological signals from one actor in a twenty days experiment.

Schreirer et al. [17] used a hacked computer mouse (random delays were introduced) with the aim to evoke intensive frustration episodes to participants. While participants used this hacked mouse, two physiological signals were collected (skin conductivity and blood volume pressure). Using a Hidden Markov Model as a feature extraction technique, they succeeded to automatically detect frustration events. This method gave a 50% accuracy level in frustration detection for 21 out of 24 users.

Mandryk et al. [18] used physiological signals in order to detect users' emotions while engaged with entertainment technologies. Participants played a video game against the computer, a friend and a stranger. In all three conditions, their physiological signals were continuously recorded and a fuzzy logic system converted them to a Valence-Arousal space. Then, a new fuzzy logic system was used in order to convert the Valence-Arousal space into discrete emotions.

The above studies based their emotion detection success on contexts that induce intense emotions, such as hardware failures and gaming. However, identification of subtle emotions is also of interest in typical HCI tasks, and remains an open research topic.

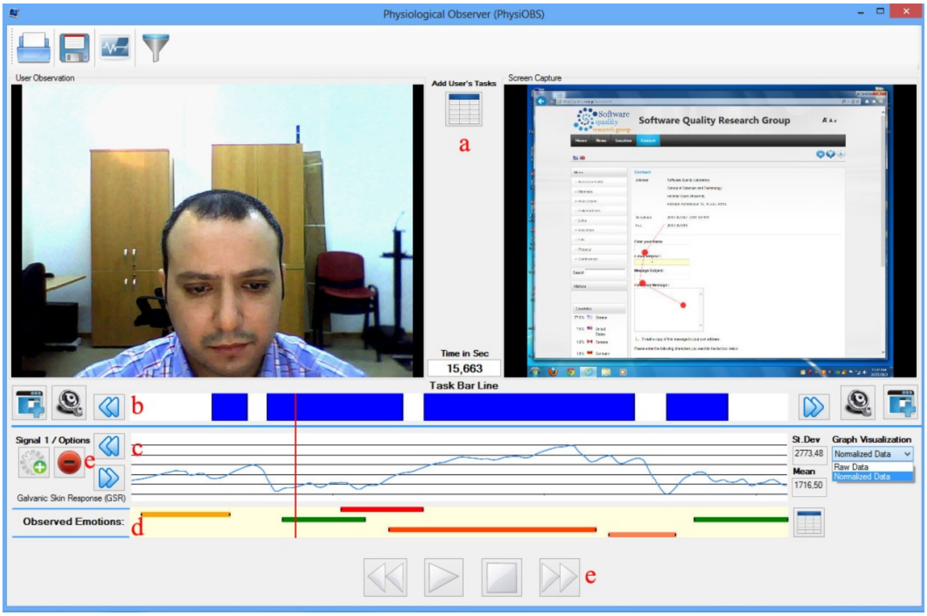
3.2 PhysiOBS Interface and a Typical Usage Scenario

PhysiOBS is a Windows-based tool and has been developed in C#. The aim of the tool is to support researchers and practitioners in the demanding task of UEX data analysis. PhysiOBS is meant to be used as a tool for post-study analysis, and thus requires, as a prerequisite, all users' data sources appropriately synchronized. PhysiOBS will be soon available for download at http://quality.eap.gr/en. In the following, the main interface of the tool (Fig. 2) along with a typical UEX evaluation scenario, are presented.

First, the evaluator adds at least one video (user's or screen recording). If both user's video and screen recording are available, the evaluator can watch them concurrently (Fig. 2, part a). In the example presented in Fig 2 (part a), the screen recording video also includes eye fixations and saccadic movements as captured by an eye-tracker. Thereafter, the evaluator can perform a typical video observation analysis supported by the tool functionality. For instance, the user of PhysiOBS can define tasks/subtasks and assign them to specific time periods (Fig. 2, part b). Extra information such as duration, result (successful/unsuccessful) and general evaluator's comments can be added for each user's task.

More importantly, user's physiological signals can also be inserted into PhysiOBS (Fig. 2, part c). Embedded semiautomatic processes, such as signal normalization and statistical analyses reported in [33], can be applied to provide a general overview of each physiological signal, basic characteristics and identified areas of potential emotional interest. Research-based guidelines [34] to complementary support emotional state identification are also provided to the evaluator. In specific, the evaluator assigns an emotion to a user-defined time period from a list of basic emotions [28] with specific associated characteristics, such as facial expressions and body movements [35-36]. To this direction an extra report, produced by user's answers analysis about their emotional state, is also provided to evaluators.

The result of the analysis process is represented as a series of emotional periods and emotion transitions (Fig. 2, part d). Color coding denotes different identified emotions (e.g. red: anger, coral: anxiety) and can be adjusted through the tool interface. In addition, all emotional periods can be optionally enriched with user's self-reported data in the form of comments. The evaluators' sense-making of the available data is supported by navigation controls (Fig. 2, part e), which are synchronized across all available views. An also important functionality provided by the tool is the save/load project option. The evaluator can save each participant's evaluation in order to edit it later.



Time in Sec 15,663

St.Dev 2773,48

Mean 1716,50

Graph Visualization Normalized Data

Fig. 2. PhysiOBS: a tool that combines physiological measurements, observation and self-reported data. (a) concurrent view of user's video and screen recordings, (b) task/subtasks view, (c) physiological signal(s) view, (d) observed emotions view with adjustable color coding, (e) navigation controls, synchronized across all available views.

3.3 PhysiOBS Preliminary Evaluation Results

Results from a preliminary evaluation study that was conducted in the Software Quality Research Laboratory (http://quality.eap.gr/en/lab) involving five HCI experts and two software practitioners were rather encouraging. Participants had at their disposal a set of user-generated data collected through a previous usability study and were asked to perform an analysis using PhysiOBS. The study indicated that PhysiOBS use can decrease time and perceived effort required to evaluate UEX from user-generated data. Furthermore, practitioners using PhysiOBS reported that the tool enabled a more in-depth UEX evaluation. In specific, participants confirmed that the simultaneous use of all available data sources can contribute different insights in the context of UEX evaluation. Study participants also provided feedback on the tool. For instance, two participants argued that for emotion selection they would prefer a wizard-like functionality with embedded help, instead of a simple menu.

4 Conclusions and Future Work

For many usability evaluation studies, emotions are no longer a supplementary parameter but a must. This paper presented PhysiOBS, an innovative tool-based evaluation approach of user's emotional experience during human-computer interaction. PhysiOBS combines multiple data sources in order to support continuous and in-depth evaluation of UEX in a straightforward and easy way. Results from a preliminary study that involved five practitioners revealed that the proposed tool-based approach could provide valuable help in such evaluations, offering an in-depth analysis of UEX.

One of our future aims is to provide additional automation in the emotion identification process based on physiological measurements of participants involved in typical HCI tasks. To this end, we are already planning studies to produce such emotionally-labeled datasets. Towards a more rigorous evaluation of our proposed tool-based approach, one future aim is to conduct a between-subjects study in which one group of evaluators will be provided with PhysiOBS to analyze user study data while the other will follow its working practices, and then compare findings between the two groups.

Acknowledgments. This paper has been co-financed by the European Union (European Social Fund – ESF) and Greek National funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) (Funding Program: "Hellenic Open University").

References

1. Lazar, J., Jones, A., Shneiderman, B.: Workplace user frustration with computers: An exploratory investigation of the causes and severity. Behaviour & Information Technology 25(2007), 239–251 (2007)
2. Igbaria, M., Chakrabarti, A.: Computer anxiety and attitudes towards microcomputer use. Behaviour & Information Technology 9, 229–241 (1990)
3. Lazar, J., Feng, J.H., Hochheise, H.: Research Methods in Human-Computer-Interaction. John Wiley & Sons Ltd. (2010)
4. Sanderson, P.M., Fisher, C.: Exploratory Sequential Data Analysis: Foundations. Human-Computer Interaction 9, 251–317 (1994)
5. Carter, K., Anderson, B.: Can video research escape the technology? Some reflections on the problems and possibilities of A.V. research. In: SIGCHI Bulletin, vol. 27, pp. 112–114 (1989)

6. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin 111, 256–274 (1992)
7. Bartneck, C., Lyons, M.J.: HCI and the face: Towards an art of the soluble. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part I, HCII 2007. LNCS, vol. 4550, pp. 20–29. Springer, Heidelberg (2007)
8. Kim, J.-H., André, E.: Emotion recognition using physiological and speech signal in short-term observation. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Weber, M. (eds.) PIT 2006. LNCS (LNAI), vol. 4021, pp. 53–64. Springer, Heidelberg (2006)
9. Epp, C., Lippold, M., Mandryk, R.L.: Identifying emotional states using keystroke dynamics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011), pp. 715–724. ACM, New York (2011)
10. Wagner, J., Jonghwa, K., Andre, E.: From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In: IEEE International Conference on Multimedia and Expo, pp. 940–943. IEEE press (2005)
11. Experience Dynamics, `http://dev.experiencedynamics.com/blog`
12. DEAP DataSet, `http://www.eecs.qmul.ac.uk/mmv/datasets/deap/`
13. MIT Media Lab, `http://affect.media.mit.edu/share-data.php`
14. Lin, T., Hu, W.: Do Physiological Data Relate To Traditional Usability Indexes? In: 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, pp. 1–10 (2005)
15. Ward, R.D., Marsden, P.M.: Physiological responses to different WEB page designs. International Journal of Human Computer Studies 59, 199–212 (2003)
16. Wilson, G., Sasse, M.A.: Do users always know what's good for them utilising physiological responses to assess media quality. In: McDonald, S., Waern, Y., Cockton, G. (eds.) People and Computers XIV-Usability or Else, pp. 327–339. Springer, London Limited (2000)
17. Scheirer, J., Fernandez, R., Klein, J., Picard, R.W.: Frustrating the user on purpose: A step toward building an affective computer. Interacting with Computers 14, 93–118 (2002)
18. Mandryk, R.L., Atkins, M.S.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Human-Computer Studies 65, 329–347 (2007)
19. Kivikangas, J.M., Ekman, I., Chanel, G., Järvelä, S., Salminen, M., Cowley, B., Henttonen, P., Ravaja, N.: Review on Psychophysiological Methods in Game Research. In: Proceedings of 1st Nordic DiGRA. DiGRA (2010)
20. Kramer, A.F.: Physiological Metrics of Mental Workload: A Review of Recent Progress. In: Damos, D.L. (ed.) Multiple-Task Performance, pp. 279–328. Taylor and Francis, London (1991)
21. Ikehara, C.S., Crosby, M.: Physiological Measures Used for Identification of Cognitive States and Continuous Authentication, Atlanta, USA (2010)
22. Bellur, S., Sundar, S.S.: Psychophysiological responses to media interfaces. In: ACM SIGCHI, Atlanta, GA,USA (2010)
23. Calhoun, B.H., Lach, J., Stankovic, J., Wentzloff, D.D., Whitehouse, K., Barth, A.T., Brown, J.K., Li, Q., Oh, S., Roberts, N.E., Zhang, Y.: Body sensor networks: A holistic approach from silicon to users. In: Proceedings of the IEEE, pp. 91–106 (2012)
24. James, W.: Discussion: The physical basis of emotion. Psychological Review 1, 516–529 (1894)
25. Cacioppo, J.T., Tassinary, L.G.: Inferring psychological significance from physiological signals. American Psychologist 45, 16–28 (1990)

26. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: Psychophysiological science. In: Handbook of Psychophysiology. Cambridge University Press, Cambridge (2000)
27. Schachter, S.: The Interaction of Cognitive and Physiological Determinants of Emotional State. Advances in Experimental Psychology 1, 49–80 (1964)
28. Ekman, P.: An argument for basic emotions. Cognition and Emotion 6, 169–200 (1993)
29. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathololy 17, 715–734 (2005)
30. Branco, P., Firth, P., Encarnac¸ao, L.M., Bonato, P.: Faces of emotion in human–computer interaction. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1236–1239. ACM Press, New York (2005)
31. Foglia, P., Prete, C.A., Zanda, M.: Relating GSR Signals to traditional Usability Metrics: A Case Study with an anthropomorphic Web Assistant. In: IEEE International Instrumentation & Measurement Technology Conference, Victoria, Canada, pp. 1814–1819 (2008)
32. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis And Machine Intelligence 23, 1175–1191 (2001)
33. Vyzas, E., Picard, R.W.: Affective Pattern Classification, In: Symposium, Emotional and Intelligent: The Tangled Knot of Cognition (October 1998)
34. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications 30, 1334–1345 (2007)
35. Fridlund, A.J., Izard, C.E.: Electromyographic Studies of Facial Expressions of Emotions and Patterns of Emotions. In: Cacioppo, J.T., Petty, R.E. (eds.) Social Psychophysiology: A Sourcebook. Guilford Press, New York (1983)
36. Bartneck, C., Lyons, M.J.: HCI and the face: Towards an art of the soluble. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part I, HCII 2007. LNCS, vol. 4550, pp. 20–29. Springer, Heidelberg (2007)

Proposal for the Model of Occurrence of Negative Response toward Humanlike Agent Based on Brain Function by Qualitative Reasoning

Yoshimasa Tawatsuji[1,*], Keiichi Muramatsu[2], and Tatsunori Matsui[2]

[1] Graduate School of Human Sciences, Waseda University, Saitama Japan
wats-kkoerverfay@akane.waseda.jp
[2] Faculty of Human Sciences, Waseda University, Saitama Japan
kei-mura@ruri.waseda.jp, matsui-t@waseda.jp

Abstract. For designing the rounded communication between human and agent, humanlike appearance of agent can contribute to human understandability towards their intension. However, the excessive humanlike-ness can cause human to feel repulsive against the agent, which is well known as the uncanny valley. In this study, we propose a model providing an explanation for how the human negative response is fomred, based on the brain regions and its function, including the amygdala, hippocampus, cortex and striatum. This model is described with quantitative reasoning and simulated. The results indicate that as human observes a humanlike agent, the emotion goes negative and the brain regions were more activated in comparison with the case human observes a person.

Keywords: Human Agent Interaction, uncanny valley, brain function, qualitative reasoning.

1 Introduction

The research field of human agent interaction has much interest in the equipment of an appropriate appearance to an agent [1]. The agent is referred as a robot in the real world or a computer program with its appearance. Then the uncanny valley [2] is a crucial issue in these studies. In general, human feels more familiar toward an agent as its appearance gets more humanlike, however, the familiarity drastically decreases when the agent gets considerably similar to but slightly different from a real person as illustrated in Fig.1. Several studies have dealt with this issue, but it is still uncertain how the human negative response can be elicited to a highly humanlike agent, and no common explanation for its mechanism to occur has been provided.

We hypothesized that human responds to a humanlike agent as if it is human and also non-human, and the contradiction of two responses elicits human negative response to the agent, which causes human to feel eerie toward the humanlike agent. In this study, we proposed the processing model how the human negative response against humanlike agent occurs.

[*] Corresponding Author.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 768–778, 2014.

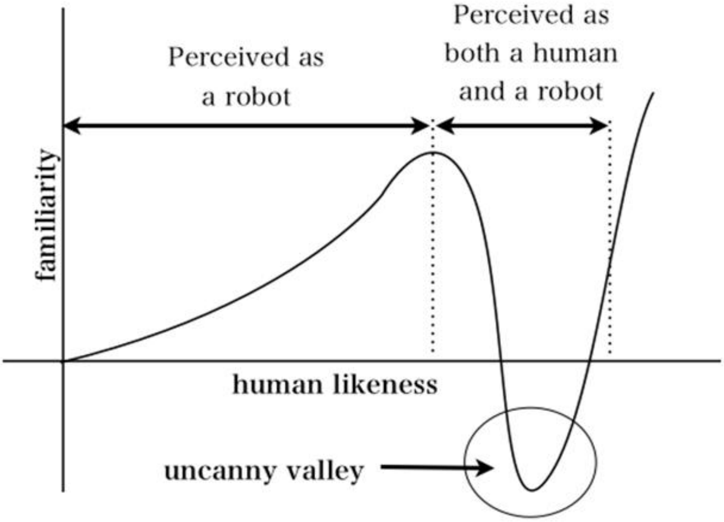


Fig. 1. Basic concept of the uncanny valley (partially altered from that in [2])

2 Experimental Method

Based on our hypothesis, the experiment was conducted to verify which information of a face people pay attention to when judging whether the face was human or not.

2.1 Procedures

This experiment used five pictures of faces of (a) a doll, (b) a CG-modeled human image fairly similar to real human, (d) another image highly similar, (c) an android robot, 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.

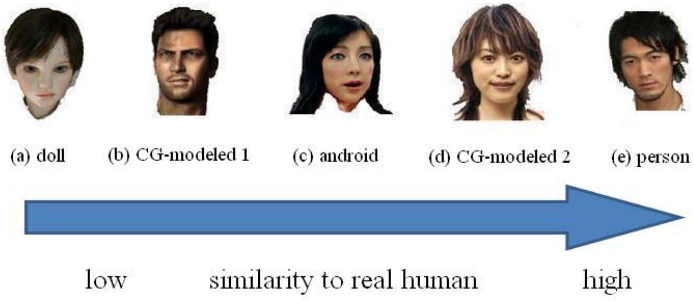


Fig. 2. Five faces used in experiment

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 on the center of the monitor. To control the initial location of eye fixations of the participants, a white page where a cross was depicted at its center 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 [3], 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 tracking data, to calculate how long the participants gazed at each face 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 2. Finally, total time length of eye fixations on each area was calculated by adding time of eye fixations on each of comprised blocks. 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.

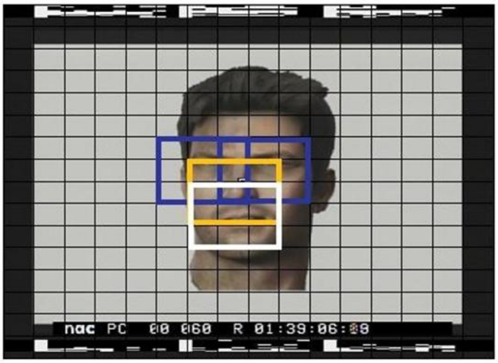


Fig. 3. Areas of right eye, left eye, nose and mouth

2.3 Judgment of human/non-human to each face

Twenty one undergraduates (18 males and 3 females) participated in the experiment. 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. Thus eye fixations on these three faces, the CG-modeled 1 judged as non-human, the CG-modeled 2 judged as human but actually non-human and the person judged as human, were compared to study the differences in perceptual process of human and non-human.

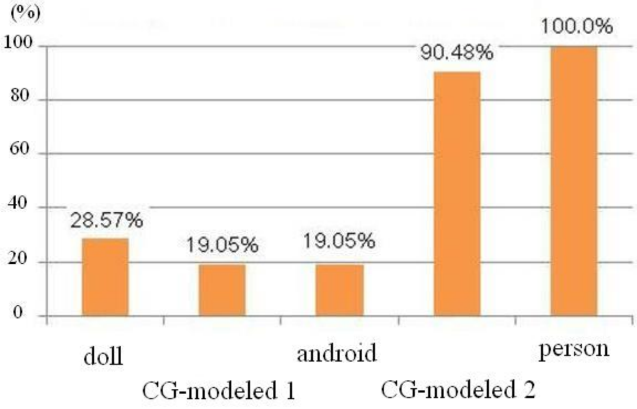


Fig. 4. Proportions of participants who judged each face was human

2.4 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).

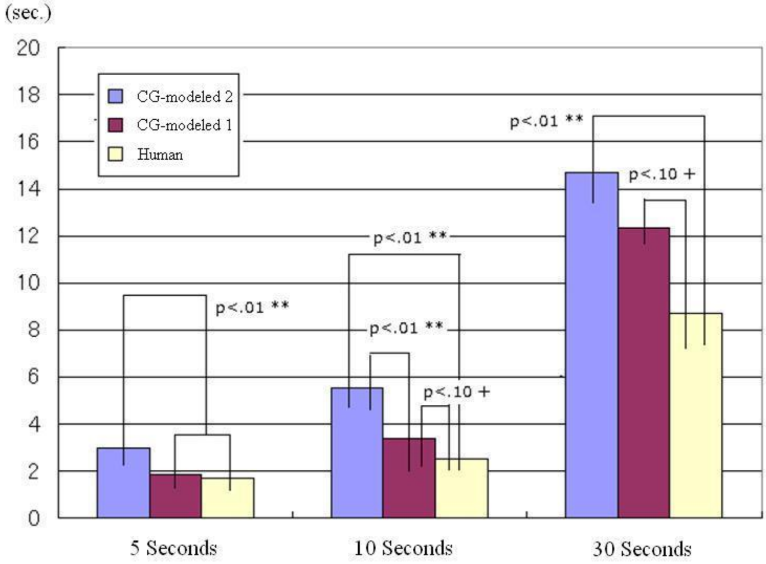


Fig. 5. Averages of time length of eye fixations on the right eyes of each face

2.5 Discussion

The result of judgment of the participants indicated that they judged the CG-modeled 1 as non-human and the CG-modeled 2 and person as human. As 30 seconds passed, time length of eye fixation on the right eye area of the person was shorter than that of both CG-modeled 1 and 2. These results are consistent with the report by Minato et al. [4]. 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. 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. These results must imply that the participants had once perceived 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. Therefore, it can be assumed that when people observe a humanlike agent, they initially perceive it as human, and they then perceive it as non-human. Thus, perceptual processes of a humanlike agent can be considered to include the two steps.

3 Proposal for the Model Interpreting the Occurrence of Human Negative Response

In this section, the model is proposed which provides an explanation for how human negative response is elicited, based on the brain function and perceptual process in two steps towards a humanlike agent

3.1 Relationship between Human Negative Response and Amygdala

According to the report by Steckenfinger et al.[5], Macaque monkey elicited the negative response and they didn't gaze at a CG modeled monkey which appearance looked like its species. Seyama et al. pointed out that human felt eerie against a humanlike agent with the abnormal features of the eye [6]. These studies provide two important suggestions. Eye movement on the agent's eye reflect the human positive or negative emotional response toward it. In addition, the negative response is not peculiar to human beings and its mechanism to generate the negative response can be commonly shared among the species. Therefore, the function of human brain region, especially phylogenetically old one, is instrumental in comprehending of how human elicits the negative response against a humanlike agent.

Amygdala is a brain region which plays an important role of the emotional processing. The processing is expected to have dual processing route called "dual pathway of emotion," consisting of "low-road" and "high-road" [7]. At the former route, the perceived stimuli is sent from the thalamus directly to the amygdala. At the later route, it is sent indirectly to the amygdala via the cortex. Accordingly, low-road is considered as imprecise and rapid processing, and high-load is considered as elaborate and slow processing. The function of these routes accounts for how the human negative response was generated.

When an observer perceives a humanlike agent, the low road rapidly processes it and the observer quickly makes, by instinct, an emotional response to the agent as if it were human: for example, an observer turns his or her eye toward the agent's face. And subsequently, high road processes it to notice that it is not human, which also makes an emotional response based on the cognitive processing. These two emotional responses generate the contradiction, which causes the observer to elicit the negative response against the humanlike agent. The model is depicted as Fig. 6.

3.2 Proposal for Advanced Model by Integration with Previous Model

Moore proposed that a Bayesian model for a psychological phenomenon, perceptual magnet effect, should be applied to generate the curve of the uncanny valley [8]. The model calculated how correctly human identified a humanlike agent as non-human and a person as human. Then, at the point where the human looked at the highly humanlike agent, there occurs a dip in the judging rate, provided a certain contribution of prior knowledge.

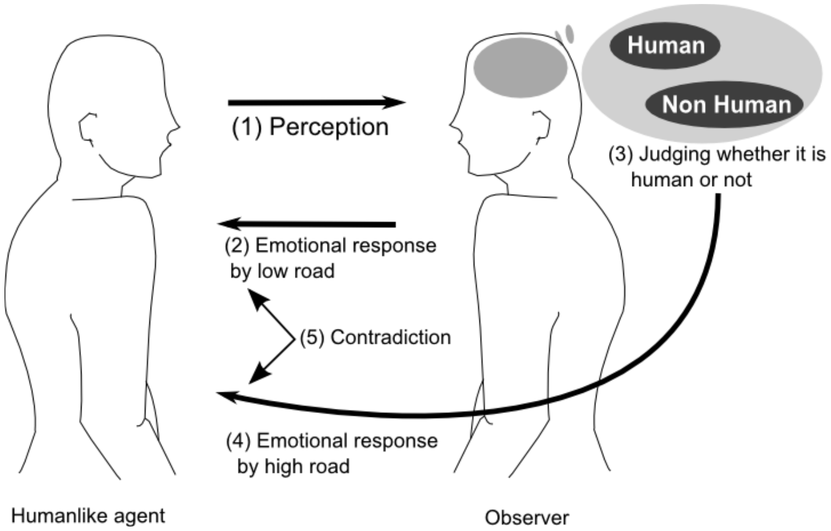

Human
Non Human
(1) Perception
(3) Judging whether it is human or not
(2) Emotional response by low road
(5) Contradiction
(4) Emotional response by high road
Humanlike agent
Observer

Fig. 6. Model flow indicating how human negative response against humanlike agent is elicited

In order to make our model more advanced, the Moore's model and our model were integrated with each other. However, architecture of two models is described in different level: the one is based on the amygdalar function and the other is based on human perceive processing. In integrating two models, the Moore's model should be reinterpreted from the perspective of brain function. Guenther et al. constructed a neural model for the occurrence of the perceptual magnet effect, focusing on the cortex and thalamus [8]. This study suggests that Moore's model is supposed to be a functional model of cortex which to account for human categorization and also of hippocampus which preserves the memories as a prior information.

3.3 Reward System Controlling the Eye Movement

Amygdalar activation lets a human make an emotional response based on approach or avoidance. Especially when an observer encounters a humanlike agent, the eye movements reflect the response, such as turning his/her eyes on the agent or away from the agent. These emotional responses can be accountable with the reinforcement learning. Reinforcement learning is supposed to be related with the reward system and striatum plays an important role in it. Striatum has connections with the amygdala and the cortex. Accordingly, the striatum plays an important role to adjust the emotional response in accordance with the differential between the expectation processed in the cortex and the reward processed in the amygdala as emotional evaluation.

4 Simulation

In this section, the model is expressed in the framework of the qualitative reasoning, and the results of simulation is introduced.

4.1 Qualitative Reasoning

The emotional model that we propose is based on the connections and the functional mechanism in the relationships among the wide range of brain regions. In this point, the qualitative reasoning is useful to simulate the change of human emotional response toward an agent. Qualitative reasoning requires to describe the system with the qualitative relationships between variables, such as “large-small,” “monotonic increase or monotonic decrease,” and so forth, and allows to show the dynamical change of the system according to a change of a variable. Thus we describe the model with the qualitative reasoning and simulate how the human emotional response alters in the time series with the qualitative simulator STELLA produced by isee system.

4.2 Qualitative Description of the Model

There is two qualitative descriptions, the connections of the brain regions and the functions of each region. Connections between the regions are descripted referring to the perceptron. Let $\mathrm{a}(X_i)$, $\omega_{ij}(\mathrm{i} \neq \mathrm{j})$ and $\theta_i$be the degree of activity of a brain region X_i, the strength of connection between the regions X_i, X_j, and the threshold of a brain region X_i, where the activity fulfills

$$\mathrm{a}\left(X_j\right) = \mathrm{sign}\left(\sum_{i\in\{1,\cdots,n\}} \omega_{ij} a(X_i) - \theta_j\right), \tag{1}$$

where $\mathrm{i} = 1,2,\cdots,\mathrm{n}$ and, in this study, $\mathrm{n} = 4$. 4 denotes the number of the regions considered in our model, which means the thalamus, the amygdala (including the hippocampus), the cortex, and the striatum. In this study, let ω_{ij} be 0.3 and let θ_i be 0.5. Fig. 7 indicates the model of the connections between the thalamus and the cortex in STELLA. If the thalamus is activated and the activity of the thalamus gets over a threshold, “opener thalamus to cortex” allows to connect the thalamus with the cortex, and then the cortex is activated.

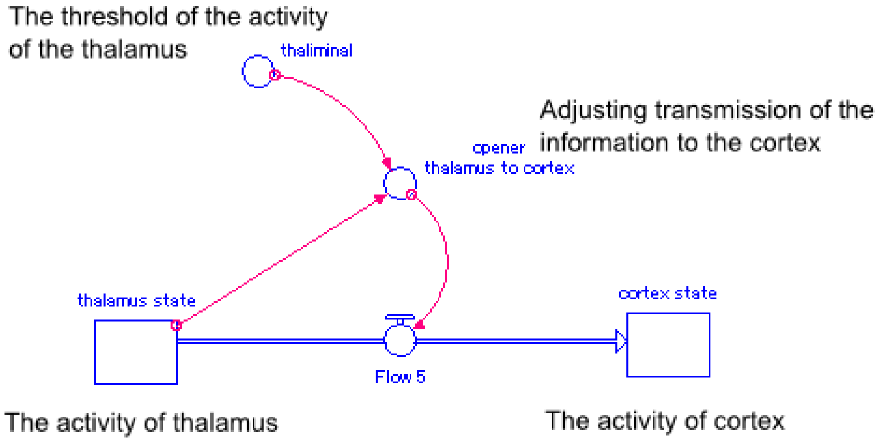


Fig. 7. Model in STELLA indicating the connection between the thalamus and the cortex

The function of the each brain region is given as follow; the thalamus perceives a stimuli, the amygdala and hippocampus evaluates it, the cortex estimates the expected value, and striatum adjusts the emotional response. The amygdala provides the qualitative value as a reward allowing the striatum to make emotional response. The cortex provides the expected value allowing the striatum to conclude how strongly the emotional response is conducted. The hippocampus is supposed to let the evaluation $\mathrm{v} \in \mathrm{Q}$, accumulated value of the rewards, come close to the expected value that the cortex estimates $\kappa \in \mathrm{Q}$,

$$\frac{dv}{dt} = (\kappa - v)v \tag{2}$$

Q denotes the set of qualitative values. The striatum enhances the elicitation rate of the emotional response (turning the eyes to the agent) in accordance with the differential between the reward and the expected value. These functions are activated provided that the activity of the corresponded regions are positive.

Fig. 8 shows the whole model implemented in STELLA. The model consist of two parts, Connection part and Evaluation part. In this study, we assumed the situation that an observer looks at a person and a humanlike agent, and the input of the system is how close the eye width and sclera width of a humanlike agent are to that of human eyes. That both Eye width and sclera width are 0.7 indicates that the observer looks at a humanlike agent. That both of them are 1 indicates that the observer looks at a person. In model, as the input of eye width and sclera is given, the evaluation gets positive by the mechanism of the row load, and increases or decreases in accordance with the value of the input.

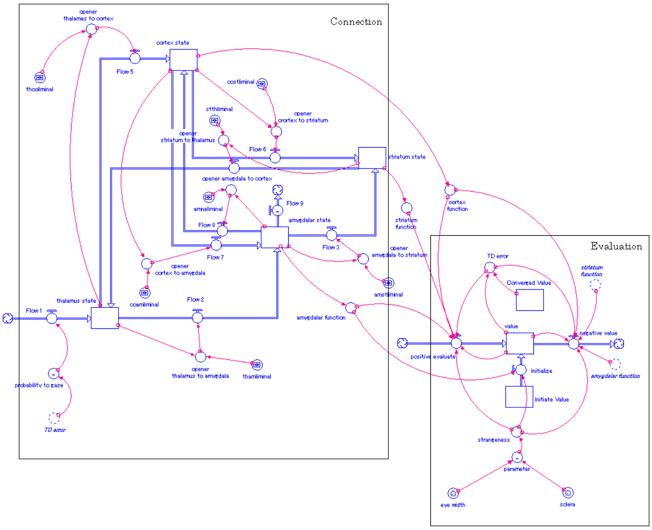


Fig. 8. The model implemented in STELLA

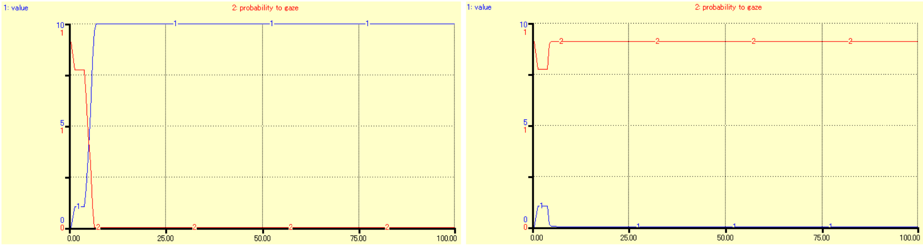


Fig. 9. Observer's evaluation (blue line) and elicitation rate (red line) of emotional response when looking at a person (left side) and a humanlike agent with abnormal eye (right side)

4.3 Simulation of the Human Emotional Response

Fig.9 indicates the time-series changes of human emotional evaluation and elicitation rate of emotional response toward a humanlike agent or a person. The evaluation converged to the expected value when the observer looks at a person. On the other hand, the evaluation once got positive by the response of the low road but soon it decreased to the zero when the observer looked at a humanlike agent. As for the elicitation rate of emotional response, the observer gradually turned his or her eyes away the person whereas the observer continued to turn his or her eyes to the humanlike agent. This suggests that as time passed, the observer turns his or her eye to the humanlike agent longer than to the person. This results can provide the explanation for the results of the experiments that we conducted. Fig. 10 shows the time-series change of activity of the amygdala. In comparison with observing a person, amygdala continues to be activated when observer looked at a humanlike agent. The human negative response is considered to be much related with the divergence of the direction of evaluation.

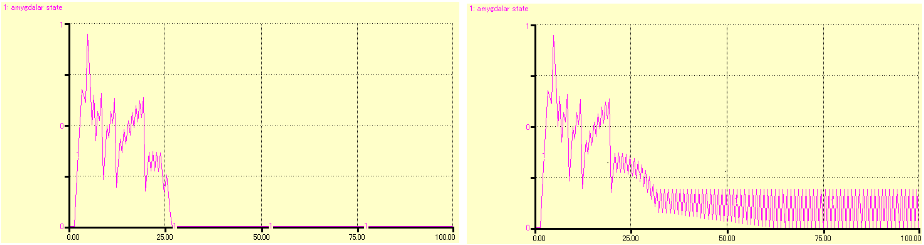


Fig. 10. Amygdalar activity when stimuli (left: a person, right: a humanlike agent) is input

5 Conclusion

In this study, the model interpreting the mechanism of how human negative response toward a humanlike agent elicits were proposed, based on the brain function of emotionally evaluation with qualitative reasoning method. We focused on the function of the brain regions, especially amygdala, cortex, striatum and hippocampus, which are related to compose human emotional response. This model suggested that human initially responds to a humanlike agent as well as to a person by the low road but subsequently, human identified it as non-human to make an emotional response by the high road. The contradiction of these responses is reflected in the human eye movements.

As the model includes some assumptions, it is an important future work to clarify the verification of them, referring to the knowledge of neuroscience. In addition, it is required to clarify the relationship between human negative response and human eerie feeling against a humanlike agent.

References

1. Yamada, S.: Originality in Human-Agent Interaction. Journal of the Japanese Society for Artificial Intelligence 24(6), 810–817
2. Mori, M.: Uncanny Valley. Energy 7(4), 33–35 (1970)
3. Yarbus, A.L.: Eye Movements and Vision. Prenum Press (1967)
4. Minato, T., Shimada, M., Ishiguro, H., Itakura, S.: Development of an Android Robot for Studying Human-Robot Interaction. In: Orchard, B., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS (LNAI), vol. 3029, pp. 424–434. Springer, Heidelberg (2004)
5. LeDoux, J.E.: The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon & Schuster (1998)
6. Steckenfinger, S.A., Ghazanfar, A.A.: Monkey visual behavior falls into the uncanny valley. Proc. of the National Academy of Sciences 16(43), 18362–18366 (2009)
7. Seyama, J., Negayama, R.S.: The uncanny valley: Effect of realism on the impression of artificial human faces. Presence: Teleoperators and Virtual Environments 16(4), 337–351 (2007)
8. Moore, R.K.: A Bayesian explanation of the ‘Uncanny Valley’ effect and related psychological phenomena. Scientific Reports (2) (2012)
9. Guenther, F.H., Bohland, J.W.: Learning sound categories: A neural model and supporting experiments. Journal of Acoustical Society of Japan 58(7), 441–449 (2002)

Current and New Research Perspectives on Dynamic Facial Emotion Detection in Emotional Interface

Tessa-Karina Tews[1], Michael Oehl[1], Helmut Faasch[1], and Taro Kanno[2]

[1] Leuphana University Lueneburg, Institute of Experimental Industrial Psychology,
Wilschenbrucher Weg 84a, 21335 Lueneburg, Germany
tews@leuphana.de, {oehl,faasch}@uni.leuphana.de
[2] The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
kanno@sys.t.u-tokyo.ac.jp

Abstract. In recent years there has been an increasing interdisciplinary exchange between psychology and computer science in the field of recognizing emotions for future-oriented Human-Computer and Human-Machine Interfaces. Although affective computing research has made enormous progress in automatically recognizing facial expressions, it has not yet been fully clarified how algorithms can learn to encode or decode a human face in a real environment. Consequently, our research focuses on the detection of emotions or affective states in a Human-Machine setting. In contrast to other approaches, we use a psychology driven approach trying to minimize complex computations by using a simple dot-based feature extraction method. We suggest a new approach within, but not limited to, a Human-Machine Interface context which detects emotions by analyzing the dynamic change in facial expressions. In order to compare our approach, we discuss our software with respect to other developed facial expression studies in context of its application in a chat environment. Our approach indicates promising results that the program could accurately detect emotions. Implications for further research as well as for applied issues in many areas of Human-Computer Interaction, particularly for affective and social computing, will be discussed and outlined.

Keywords: Emotional Interfaces, Affective Computing, Facial Expression, Human Machine Interface.

1 Introduction

Recently, there has been a promising increase in interdisciplinary exchange between psychology and computer science in the field of recognizing emotions [1-2]. Therefore, our study attempts to present previous developed affective communication methods and compare them with a possible usage in reliable chat emotional interface software assisted by visible muscle movements in terms of emotion detection from mimics, i.e., the user's facial expression. The psychology driven idea behind detecting a user's facial expression is based on knowledge of previous psychological studies in the field of facial expression detection [2-3].

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 779–787, 2014.

From the beginning chat was predominantly used for casual internet communication. In the meantime it has become a common and well-established form of communication. Among its advantages are its easy access, its low-tech affordances, high usability and fast and flexible communication processes. A variant of chat, SMS messaging, has met the public with unexpected popularity. Nevertheless, due to its media properties, chat-based communication frequently suffers from deficits due to incoherence of contributions, lack of coordination and problems of awareness concerning social awareness as well as awareness of context and available knowledge [4]. These shortcomings of conventional chat-based communication pose severe problems. In this paper we focus on affective or emotional aspects auf chat-based communication. We claim that these restrictions can be overcome and propose that extending the medium 'chat' with appropriate strategies of affective computing embedded in the chat environment can actually improve the chat discourse. In this context the effects of emotions in the chat environment have not been researched comprehensively yet. Computer-mediated communication usually does not contain information about the emotional state of a user during typing. Reports on the influence of emotions are mostly based on observations or interviews and do not compare to empirical methods. During interviewing, computer-mediated communication studies point out several comments in logs that participants were frustrated with the interview agent's responses [5].

As a concrete example for implementing and discussing our approach in chat-based environments, we refer to a chat-based interview agent developed by Kanno and colleagues [6-8]. This chat-based interviewer agent was developed to deviate knowledge from the user. The concept is to create a kind of chat program that automatically generates questions and responses to the answers from an interviewee. The basic technique behind this is so-called artificial non-intelligence, i.e., the agent makes responses based on keywords identification and relatively simple rules like ELIZA [9]. Also a set of concepts was implemented and an interview guide as a database for this interviewer agent. The first prototype was developed and tested [10]. However, human users sometimes found the responses from the agent unfitting. To overcome the deficiencies of chat-based interview agents, our psychology driven approach might be applied as a possible chat-based emotional interface to reduce the number of unexpected responses and hence the irritations of the user, i.e., the interviewee.

This paper is about how emotions can be detected, and what kind of methods can be applied to extract users' responses or affective states in a possible emotional interface within a chat environment. Further, this paper will give a brief overview of recent developments of affective methods in order to recognize the user's response. After the development section, we will introduce our method of processing the facial responses, e.g., raised eyebrows, of the user and show early results [3] and [11]. Furthermore, our first results lead to a discussion on additional applications and limitations that frames an attempted approach of emotion detection in chat-based environments.

2 Emotion Recognition Method Development

The first step to recognizing affective aspects of communication in Human-Computer Interaction, such as in a chat-based interviews, involves detecting relevant emotional information, e.g., raised eyebrows, from non-relevant. During this process, the main problem is how to quantify this information in order to enable a computer to recognize the meaning in the data [12]. To relate our approach to other studies, we present several approaches and a variety of methods for analyzing communications: Affective Dialog Systems [12-14], Sentence-Based Emotion Recognition [15], and Multimodal Chat Emotion Recognition [16]. All these methods will be described in the following section.

2.1 Affective Dialog System (DS)

Finding relevant information is the basis of affective communication. Affective Dialog Systems can classify the information and they are an important tool for studying affect and social aspects in online communication. An Affective Dialog System is a social intelligence model, i.e. agents that handle affective responses with the help of psychological theories of personality, emotion, and Human-Computer Interaction [12].

Morishima and colleagues [12] argued that agent's socially appropriate affective communication provides a new dimension for collaborative learning systems. In our case, for example, an interview chat-agent can more efficiently interview the user.

Turkle [13] pointed even out that online communication has a huge effect on users' social and psychological perceptions and behavior and even their self-concepts. Skowron and colleagues [14] indicated that an affective system can influence the user in terms of chatting enjoyment, dialog coherence, and realism. Furthermore, the variants of the affective system strengthen the chatting enjoyment and emotional connection. For this reason, an Affective Dialog System can be an important input in the field of developing a dynamic questionnaire. However, the direct feedback, such as facial expressions, will not be included in affective text-based research and could be a drawback during interviewing a user.

2.2 Sentence-Based Emotion Recognition

To further classify emotions in the context of computer-mediated communication, a promising approach is textual emotion recognition. Krcadinac and colleagues [15] present an approach that analyses on the sentence level based on the standard Ekman emotion classification [2]. The developed algorithm reads a text sentence in a chat as an input and sorts it to the six emotional states defined by Ekman (i.e., happiness, sadness, anger, fear, disgust, and surprise). To study emotions in computer-mediated communication, a keyword-spotting method was developed by Krcadinac and colleagues based on a free, open source library software system Synesketch which includes a WordNet-based word lexicon; a lexicon of emoticons, common abbreviations and; colloquialisms, and a set of heuristic rules. During their study each of the

214 participants needed to rate 20 sentences randomly taken from the corpus to one for each emotional status.

The results indicated a high accuracy (~80%) that can lead to further promising future research and applications. However, this approach has two drawbacks one is that the corpus is quite reduced to a basic sentence level with relatively unambiguous emotional type and cannot compared with a fluent chat communication, such as in a dynamic questionnaire. Second, it cannot recognizing neutral as a separate type. In our research we try to include our feature extraction method so that a communication can be analyzed in direct feedback of the participants face.

2.3 Multimodal Chat Emotion Recognition

However, as the use of emoticons and text-only analyses suggests, communication without nonverbal analysis, such as facial expressions, can be monotonous [16]. Another possible way to include emotions in the chat environment is to create a 3D Avatar by extracting the facial actions of each participant with real-time facial expression analysis techniques and research on synthesizing facial expressions and text-to-speech capabilities. Chandrasiri and his colleagues have created a piece of software that creates a 3D facial animation of agents [16]. Their system includes visual, auditory, and primary interfaces to communicate as one multimodal chat interaction. Participants can represent themselves as predefined agents. During the experiment, for example, a user showed facial expressions while typing text in the chat. The represented 3D agent will speak the message aloud while it repeats the recognized facial expression and also replay the synthesized voice with proper emotional pronunciation.

The biggest advantage of the software is that the visual data exchange requires only low bandwidth and, therefore, works in real time. The disadvantage is that the software need a person-specific initialization and several interfaces. Furthermore, in recent years videoconferencing tools have become more popular in our daily lives, e.g., Skype and MSN. The user might be more convinced to use simple live stream video communication software rather than an avatar software for getting more direct facial expression feedback.

3 Software

In relation to the above mentioned approaches, this paper suggests a new psychology driven dynamic approach to detect emotions. In 1979 Bassili [17] suggested that even with minimal information about the spatial arrangement of features, participants can recognize facial expressions. Another interesting approach was presented by Kaiser and colleagues [18]. Here, in order to reduce the complexity, small dots were placed on the participant's face which were then detected and analyzed on videotapes. The dots allow the algorithm to determine the underlying muscle activity.

Earlier approaches attempted to detect a set of typical emotions, such as happiness, surprise, anger, and fear [19]. In contrast to those studies, we focus on the detection of emotion or no emotion present with the less complex feature dots tracking method [3], [11], and [20]. Thus, we developed a computer program that might be able to detect ten blue dots placed on the participant's face. The positions of the points were derived from earlier psychological studies, investigating facial muscle movement with the help of Electromyography (EMG) [21], as well as observation of human mimics [22]. Our software locates the blue dots by searching each frame of the video, line by line. We selected blue as the color for the dots, since blue is present only minimally in the color-spectrum of the human face.

Within the chat environment, our analysis might help to improve the interview communication of a chat agent and a user by detecting the unexpected responses of the chat agent with a facial expression detection.

3.1 Software Development

In our previous studies we recorded several videos of the participants' face [3] and [11]. The study was performed with N = 59 (40 female) participants with an average age of M = 23.39 years (SD = 4.51), who acted the emotions of anger, happiness and no expression. The recorded videos were rated for further analyses by independent raters with regard to the emotional content of the facial expressions.

In a further step the facial expression videos of the participants were analyzed and the area (A-D) and distances (1-3) between the selected blue dots were calculated (see Figure 1).

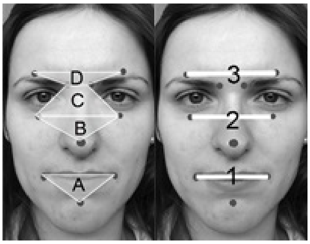


Fig. 1. Faces with dots and their areas (A-D) and distances (1-3)

Then the algorithm summed up each area and distance and then calculated the arithmetic mean with their variance. Consequently, the variance clearly reveals the motion of each area or distance. The following table shows the ideal state of the development of each emotion over time (see Table 1).

For example, the area of “A” shrinks during expressing the emotion anger (closed mouth) and grows during showing the emotion anger (open mouth), since the participant opens their mouth. The preliminary results in Tables 2-3 result from how similar the values correspond to the ideal values in Table 1. An up-arrow means the area or distance is growing larger and a down-arrow means the area or distance is getting smaller over time. The minus signs represent no detectable dot movements, i.e., the participant is in a neutral state. Additionally, we split angry emotions into open and closed mouth types for a more efficient analysis. Preliminary results showed significant changes in the mouth area enclosed by the selected blue dots when participants experienced anger with an open and closed mouth. Finally, the results of the variation of each area and distance are displayed in tables.

Table 1. Dynamic variation of each area (A-D) and distance (1-3) in relation to each emotion

	A	B	C	D	1	2	3
Anger (open mouth)	↑	↑	—	↓	↓	—	↓
Anger (closed mouth)	↓	↓	↑	↓	↓	↓	↓
Happy	↑	↓	↑	↓	↑	↓	↕
Neutral	—	—	—	—	—	—	—

3.2 Software Testing

We reported first software testing results in Tews [3] and [11] within psychological facial expression studies, mostly in an automobile context. Due to technical dropouts only n = 10 subjects' data could be analyzed. Because of the small database, we refrained from using inferential statistics and our results are only displaying descriptive statistics. The tables 2 and 3 show the results of the angry with open (oM) and closed mouth (cM) and no facial expressions video analysis. We excluded the results of the happiness condition, because they were similar to the angry emotions results. To standardize the results, the average of all emotion feature values are set to 100% so that the deviation from each emotion can be expressed as a percentage. The emotion values are shown on the x-axis, describing how strongly each emotion was expressed. The participants' relations are shown in separate columns and displays each participant. For the angry facial expression videos, the raw data of the participants does not show any clear results (see Table 2).

Table 2. Results of the angry facial expression videos

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
Anger (oM)	34	38	8	1	0	2	4	2	8	8
Anger (cM)	35	33	8	1	28	2	0	1	8	2
Happy	29	27	14	1	28	2	2	1	14	12
Neutral	2	2	70	97	44	94	93	96	70	78

The values were broadly spread because of the mimic activities caused by the emotions. In contrast to participants (P0) and (P1), participant (P3) showed only few emotional expressions (see Table 2).

Table 3. Results of the no facial expression videos

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
Anger (oM)	**0**	**0**	**0**	**9**	**3**	**0**	**0**	**1**	**1**	**0**
Anger (cM)	**1**	**2**	**1**	**4**	**3**	**1**	**1**	**0**	**1**	**0**
Happy	**1**	**0**	**1**	**9**	**2**	**1**	**1**	**0**	**1**	**1**
Neutral	**98**	**98**	**98**	**78**	**92**	**98**	**98**	**98**	**97**	**99**

Compared to the emotional facial expression results, the neutral values are explicit, because of lack of movement in the face, as it is expected from a neutral face (see Table 3). The values are concentrated within the neutral emotion row. Another interesting point is the noise-induced error, especially participant (P3) shows the typical error with few values within the emotion states.

In conclusion, our study yielded the promising result that our approach was able to distinguish between an emotional state and no emotional expression. This might be used for example as a chat-based emotional interface for an interview agent to improve the communication with a user.

4 Conclusion

Chat-based internet communication has become a common form of communication. Nevertheless, chat-based communication frequently suffers from deficits due problems of awareness concerning social perceptions as well as context and available knowledge [4]. To address those problems, in this paper we focused on a possible chat-based emotional interface for an interview agent to improve the communication with a user. By reducing the number of unexpected responses, the chat-based interview agent can, for example, adapt and response to the user more dynamically for a more efficient communication. As a concrete example for implementing and discussing our approach in chat-based environments, we referred to a chat-based interview agent developed by Kanno and colleagues [6-8] and [10].

Within this paper we presented previously developed affective communication methods and compared them with a possible usage in reliable dynamic interface assisted by visible muscle movements, i.e., Affective Dialog Systems [12-14], Sentence-Based Emotion Recognition [15], and Multimodal Chat Emotion Recognition [16].

In contrast to the involved multimodal interface method, Multimodal Chat Emotion Recognition, our approach tries to reduce the complexity of affective detection by extracting the features with a simplified feature dots detection method. Though the Affective Dialog Systems and Sentence-Based Emotion Recognition methods were less involved, the main drawback was the text-only analysis without the direct feedback of the participants face.

To detect the affective expressions with our new psychology driven dynamic approach, we placed ten dots on the face of the participant. By analyzing the movement of blue dots, our software can help to distinguish the participants' facial expressions by discriminating the neutral and the emotional state. The new measurement, the dynamic variance of areas and distances was implemented to distinguish the participants' states.

Results showed that the variance of an area and distance defined by distinct dots can support the affective detection. Though our study has some limitations, our methods indicate promising results that our program could be tested in the chat environment. Our future research will also include the extended collection of data of affective expressions in the chat environment, in relation to the responses of a chat agent.

Acknowledgments. This research was supported by JSPS Pre/Postdoctoral Fellowship and FS-Nr. 2006.63 from the 'Arbeitsgruppe Innovative Projekte' (AGIP) of the Ministry of Science and Culture, Lower Saxony, Germany. Many thanks go to Mark Hammonds for his support in style and editing of English language.

References

1. Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., Meyer, J.-J.C.: Computational Modeling of Emotion:Toward Improving the Inter- and Intradisciplinary Exchange. IEEE Transactions on Affective Computing 4(3), 246–266 (2013)
2. Ekman, P., Friesen, W.V., Hager, J.: The Facial Action Coding System (FACS): A technique for the measurement of facial action, Palo Alto (1978)
3. Tews, T.-K., Oehl, M., Siebert, F.W., Höger, R., Faasch, H.: Emotional human-machine interaction: Cues from facial expressions. In: Smith, M.J., Salvendy, G. (eds.) Human Interface, HCII 2011, Part I. LNCS, vol. 6771, pp. 641–650. Springer, Heidelberg (2011)
4. Oehl, M., Pfister, H.-R.: E-Collaborative Knowledge Construction in Chat Environments. In: Ertl, B. (ed.) E-Collaborative Knowledge Construction: Learning from Computer-Supported and Virtual Environments, pp. 54–72. IGI Global, New York (2010)
5. Kumar, R., Rosé, C.P., Wang, Y.C., Joshi, M., Robinson, A.: Tutorial dialogue as adaptive collaborative learning support. Frontiers in Artificial Intelligence and Applications 158, 383 (2007)
6. Ochi, Y., Kanno, T., Furuta, K.: An Interview Agent for Cognitive Task Analysis. In: Proceedings of Human-Agent Interaction Symposium 2010, 3C-1 (2010) (in Japanese)
7. Ochi, Y., Kanno, T., Furuta, K.: An Interviewer Agent for Cognitive Task Analysis. In: Proceedings of Human Interface Symposium 2011, 1441L, pp. 381-390 (2011) (in Japanese)
8. Kanno, T., Ochi, Y., Chou, T., Furuta, K.: Service Cognition Probe Techniques. In: Proceedings of the 3rd Symposium on Systems Innovation, pp. 51–53 (2011) (in Japanese)
9. 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)

10. Kanno, T., Uetshuhara, M., Furuta, K.: Interviewer agent for cognitive task analysis. In: Stephanidis, C., Antona, M. (eds.) UAHCI/HCII 2013, Part I. LNCS, vol. 8009, pp. 40–49. Springer, Heidelberg (2013)
11. Tews, T.-K., Oehl, M., Faasch, H., Kanno, T.: A Survey of Dynamic Facial Emotion Detection in Emotional Car Interfaces. In: FAST-Zero 2013 Proceedings, JSAE Paper 20134654, No.TS2-5-4, JSAE - Society of Automotive Engineers of Japan, Tokyo (2013)
12. Morishima, Y., Nakajima, H., Brave, S., Yamada, R., Maldonado, H., Nass, C., Kawaji, S.: The role of affect and sociality in the agent-based collaborative learning system. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 265–275. Springer, Heidelberg (2004)
13. Turkle, S.: The second self. In: Simon & Schuster, New York (1984)
14. Skowron, M., Theunis, M., Rank, S., Kappas, A.: Affect and Social Processes in Online Communication—Experiments with an Affective Dialog System. IEEE Transactions on Affective Computing 4(3), 267–279 (2013)
15. Krcadinac, U., Pasquier, P., Jovanovic, J., Devedzic, V.: Synesketch: An Open Source Library for Sentence-Based Emotion Recognition. IEEE Transactions on Affective Computing 4(3), 312–325 (2013)
16. Chandrasiri, N.P., Naemura, T., Ishizuka, M., Harashima, H., Barakonyi, I.: Internet communication using real-time facial expression analysis and synthesis. IEEE Multimedia 11(3), 20–29 (2004)
17. Bassili, J.N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology 37, 2049–2058 (1979)
18. Kaiser, S., Wehrle, T.: Automated coding of facial behavior in humancomputer interactions with FACS. Journal of Nonverbal Behavior 16, 67–83 (1992)
19. Tian, Y.-I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(2), 97–115 (2001)
20. Tews, T.-K., Oehl, M., Siebert, F.W., Höger, R.: Emotional Interfaces in Cars: Cues from Facial Expressions. In: de Waard, D., Gérard, N., Onnasch, L., Wiczorek, R., Manzey, D. (eds.) Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society Europe Chapter, Human Centred Automation, pp. 111–122. Shaker Publishing, Maastricht (2011)
21. De Luca, C.J.: The use of surface electromyography in biomechanics. Journal of applied biomechanics 13, 135–163 (1997)
22. Cohn, J.F., Ekman, P.: Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In: Harrigan, J.A., Rosenthal, R., Scherer, K. (eds.) Handbook of Nonverbal Behavior Research Methods in the Affective Sciences, pp. 9–64. The Oxford University Press, New York (2005)

Evaluation of Graceful Movement in Virtual Fitting through Expressed Emotional Response and Emotion Expressed via Physiology Measures

Wan Adilah Wan Adnan[1], Nor Laila Md. Noor[1], Fauzi Mohd Saman[1],
Siti Nurnabillah Zailani[1], and Wan Norizan Wan Hashim[2]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Malaysia
[2] Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak,
Kuching, Malaysia
adilah@tmsk.uitm.edu.my

Abstract. Graceful interaction is a form of interaction that incorporates quality movement that can invoke the emotional appeal of users engaged with it. However method of evaluation of the quality graceful interaction has not been discussed. As we argue that graceful interaction can evoke emotion, we explore the use of possible instruments to evaluate graceful interaction based on the valence–arousal model. To measure emotional response of arousal the response is using verbal and non-verbal instruments. The former is based on self-report emotions and the later through autonomic measures of emotion via bio-physical measures of skin conductance. We conducted an experiment with six participants who were given the tasks to perform movement tasks in virtual fitting using three different virtual fitting room (VFR) applications available on e-commerce fashion retailing websites. The selection of the VFR applications was based on the presence of two identified graceful interaction design elements, which are tempo and sequence as prescribed by the graceful interaction design model. While performing the tasks, each participant's physiology measure of emotional response was recorded using the tool BioGraph Infiniti. Upon completion, the participants were requested to report their emotional response in an instrument constructed based on the valence arousal model. Finally each participant was also interviewed to state the VFR applications they preferred. The analysis of each type emotional response were made and the findings showed the congruence between the verbally expressed emotional response and physiology measure of emotional response in performing graceful interaction tasks. This suggests that the evaluation of graceful interaction can be made by the use of verbally and non-verbally expressed emotional respond instruments.

Keywords: aesthetic experience, graceful interaction, emotional design, virtual task, physiological measure, human computer interaction.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 788–797, 2014.

1 Introduction

Early definition of interaction design was centred on the notion of "shaping of interactive systems with particular emphasis on their use qualities" and was extended to include the use context where the computer is part of the mediated activity system [1]. Good interaction design encompasses three quality perspectives: constructional quality for the structure; ethical quality for the function and aesthetic quality for the form [2]. When considering the overall quality of the interactive product experience, HCI research has covered the users' perspective on the role of aesthetics for systems usefulness and aesthetic appeal. It has been argued that aesthetic appeal should not be only focus on visual aesthetics but encompass aesthetics in movement interaction. Though the quality of movement interaction can be measured through measures of efficiency and ease of use, as interaction design extends into interaction that yields aesthetic experience, measures of quality may also relates to emotional measures.

Graceful interaction is an aesthetic experience of a human in moving beautifully. In HCI, graceful interaction is claimed to be a form of interaction that incorporates quality movement that can invoke the emotional appeal of users engaged with it [3]. However the method of evaluation of quality graceful interaction is still unexplored. The valence–arousal model [4] is a popular model used to capture two dimensions of emotion which are pleasure and arousal. The former describe the pleasantness and the later describes physical activation. In dealing with movement quality, instruments for emotional measures related to arousal can be classified into two major types: verbal and non-verbal. In this work we chose the use context of virtual fitting in the apparel e-commerce environment that tries to enhance the shopping experience of the customer. Here, we seek to evaluate the quality of graceful interaction in virtual fitting tasks using emotional measures through the triangulation of expressed emotional response and emotional response via physiology measures of skin conductance.

2 Related Work

Graceful interaction was first discussed in the design of dialogue systems in [5] in the context of spoken and written man-machine communication where the focus was on user friendliness in command line interaction style. Later graceful interaction was studied in the context of intelligent environment in addressing issues of how the user can deal appropriately with anything a system happens to do so as anyone observing the user perceive the interaction as effective and effortless while at the same time appearing to be rational and elegant [6]. In these early works of graceful interaction the underlying concept is centred on the notion of ease of use rather than the notion of aesthetics where gracefulness should be related to. In [3] the concept of graceful interaction is revisited and argued from the success of invoking the user's emotional stimulation in an artifact in line with concept of flow in [2]. Here, in [3] graceful interaction is viewed as a form of quality engagement through the dynamic property of an interaction in a movement or action that is normally recognized as 'pleasing or attractive' to the users engaged with it. A model of graceful interaction through the use of

Laban Theory of Movement [3] describes four design elements: rhythm, tempo, direction and sequence. The model was further described through the phenomena mapping with dance movement [7] and tested for its ability to evaluate movement quality using the Laban Movement Analysis [8].

Aesthetic concept is generally perceived as a philosophical discipline and scientific effort to make aesthetic judgement is frown upon. Despite that there are arguments that aesthetic perception combines senses, science and the experience of beauty in neural systems that determine pleasure [9]. Based on these arguments attempts of measuring aesthetics were made both from theoretical and empirical perspectives. Theoretical models such as Birkhoff aesthetic measures, Klinger and Salingaros aesthetic measure and informational aesthetic measures informed the influence of harmony, symmetry or order of the aesthetic forms which implies that the complexity and disorder in the forms create an unpleasant response from the viewer [10]. Aesthetic judgment has also been explained through neurological explanatory model where aesthetic is shown to be a function of an evaluation process which implies that habitual aesthetic evaluation may affect the process of aesthetic evaluation [9]. Empirically the value of aesthetic forms is most apparent on the effect of the users as seen through the affective priming paradigm [11] that leads the empirical measurement of emotion to make aesthetic judgement [12]. As affect or emotion is a mind-body phenomenon, it can be defined by different components such as behavioural response, expressed reactions through verbal reaction (e.g. Kansei), non-verbal reaction (e.g. smiling), physiological reaction (heart beat) and [12]. Affect has at least two qualities: valence (pleasantness or hedonic value) and arousal (bodily activation). Emotional granularity can be used in verbal instruments that can be developed based on the valence-arousal model [13]. Non-verbal instruments for measuring expressive reaction may also include measurement from facial and vocal expression analysis. Emotions that manifest into physiological reaction can be detected through various measures such as blood pressure responses and skin conductance responses (SCR). Skin conductance is widely used in research to serve indicators of processes such as attention, habituation and arousal [14].

3 Research Method

To evaluate the quality of graceful interaction we conducted an experiment with six participants to measure their arousal level via expressed verbal reaction and physiological reaction. The participants were asked to perform virtual fitting tasks using three different virtual fitting room (VFR) applications as the artefacts of inquiry. The VFR applications are selected from the retailing websites of fashion stores of H&M (VFR1), brides.com (VFR2) and F&F (VFR3) based on the presence of two identified graceful interaction design elements, which are tempo and sequence following the graceful interaction design model of [3]. The VFR1 represents graceful interaction with design elements of tempo-fast and sequence-order while the VFR2 represent graceful interaction with design element of tempo-slow and sequence-order. VFR3 represents graceful interaction with design elements sequence-disorder. A summary of VFR used are shown in Table 1.

Table 1. Graceful Design Elements in the VFR

VFR Applications		Graceful Interaction Design Element			
		tempo		sequence	
Abbreviation	E-Commerce Site	fast	slow	order	disorder
VFR1	H&M	√		√	
VFR2	brides.com		√	√	
VFR3	F&F				√

For VFR1 and VFR2, the tasks to perform are: selection of an apparel (T1) and virtual fitting of the apparel in the avatar (T2). However for VFR3, both tasks are integrated. During the tasks, the SCR of each participant was recorded as the physiology measure that detects level of arousal using the tool BioGraph Infiniti. The participants wore physiology sensors attached to their two fingers. The SCR graphs are produced by plotting SCR at the y-axis and the response time duration at the x-axis. After performing the tasks the participants were asked to rate a checklist of positive emotion for valance (pleasure) and arousal to capture their emotional response towards the gracefulness of the virtual fitting activities. The participants are then interviewed to determine their preference of the movement quality.

4 Results and Analysis

4.1 Analysis of Verbal Expressed Emotional Response

The high ratings (>3) given by each participants were to each positive emotion of valence and arousal for each VFR is shown in Table 2.

Table 2. Analysis of Arousal and Pleasure Verbally Expressed Emotional Response

VFR	Positive Discrete Emotion of Arousal (rating > 3)				
	aroused	astonished	excited	delighted	happy
VFR1	5	4	5	4	5
VFR2			4	4	4
VFR3	4	4	4	4	4
	Positive Discrete Emotion of Valence (Pleasure) (rating > 3)				
	pleased	glad	content	relaxed	calm
VFR1	4	4	5	5	4
VFR2		4	4		
VFR3	4	4	4	4	4

For the arousal dimension, all participants gave a high score to all discrete emotion of VFR1 and VFR3. However the scores for VFR1 are generally higher compared to VFR3. For VFR2 the participants only gave high score to discrete emotion of excited, delighted and happy. Similarly for the valence dimension, the participants gave high scores to all discrete emotion of VFR1 and VFR3. For VFR2 the participants only gave high score for glad and content. This implies a higher rating of expressed emotional response is given to movement with fast tempo (VFR1) when compared to movement with slow tempo (VFR2). The results also showed that movement with the sequence of order (VFR1) received higher rating of expressed emotional response when compared to sequence of disorder (VFR3). This result is in agreement with theoretical aesthetics measure which states that disorder will cause unpleasant response from reviewers.

4.2 Analysis of Preference

When the participants were asked for VFR applications they preferred, all of them stated VFR1 as the most preferred. When asked for their individual preference of movement quality for graceful interaction all of the participants chose tempo-fast and sequence order. This concurs with the scores for both high and low dimension emotional response for VFR1 as shown in Table 3.

Table 3. Preference for Movement Quality

Participant	Preference for Movement Quality		Preference for the Combination of the Movement Quality
	Tempo	Sequence	
P1	Fast	Order	More than two
P2	Fast	Order	More than two
P3	Fast	Order	More than two
P4	Fast	Order	More than two
P5	Fast	Order	More than two
P6	Fast	Order	More than two

The analysis of movement quality preference also concurs with the results obtain from the analysis of emotional granularity of expressed emotional response.

4.3 Analysis of Expressed Emotional Response via Physiological Measurement of Skin Conductance

Peaks in the SCR graph represent the participants' arousal while using the VFR applications. Observation of the peaks is focused on the time taken to reach the peaks and the frequency of peaks. The high frequency of peaks represents the occurrence of individual peaks in the signal where the density of the peak and peak onset times are associated with sympathetic arousal which give sense on how arousing the activity performed. The SCR graph for each participants showed that each of the participants experienced different types of arousal even though they were doing the same tasks at the same place. The graphs also showed unique and different physiological signal as the body condition of participants are not similar to each other. It means the body condition also influenced the emotional state of participants.

Time to Reach First Peak. An analysis of the time taken to reach the first peak in the SCR graph is shown in Table 4.

Table 4. Time Taken to Reach the First High Frequency Peak

Participants	Time taken to Reach the First High Frequency Peak				
	Task 1		Task 2		Combined Task
	VFR1	VFR2	VFR1	VFR2	VFR3
P1	00:00:12	00:00:20	00:00:15	00:01:00	00:01:30
P2	00:00:26	00:00:20	00:00:16	00:02:30	00:00:45
P3	00:01:05	00:03:00	00:00:15	00:00:30	00:02:05
P4	00:00:11	00:01:00	00:00:19	00:01:00	00:00:20
P5	00:00:14	00:01:30	00:00:21	00:00:30	00:01:40
P6	00:00:21	00:00:10	00:00:12	00:01:16	00:00:20

All participants took a longer time to show an emotional response when performing the Task 1 and Task 2 in VFR2 (except for P2) when compared to VFR1. The result is more difficult to interpret for VFR3 as the tasks were combined (application constraint). However, the time to peak is generally shorter than VFR2 but longer than VFR1. This result concurred with the results obtained for the expressed emotional response in section 4.1 and 4.2.

Number of Peaks Frequency in SCR Graph. The high peak frequency in the SCR graphs is correlated to the high arousal level. The comparison of SCR graphs of each of the participants performing Task 1 and Task 2 using VFR1and VFR2 are shown in Table 5 and Table 6 respectively.

The SCR graphs of each participant using VFR3 is shown in Table 6.

The SCR graphs of every participant show different peaks because some of the participants took a longer time to select a model compared to others. The number of high frequency peaks in the SCR graph for each participant is summarized in Table 7.

It can be seen that in Task 1 the number of high frequency peaks for VFR1 and VFR2 are almost similar. However in Task 2, the number of high frequency peaks for VFR1 and VFR2 are higher than Task 1. This can be interpreted as there is a low arousal for movement activity in Task 1 when compared to Task 2. In addition the number of high frequency peaks Task 2 is higher in VFR1 as compared to VFR2. Similarly there is also an indication of high arousal for movement activity in VFR3. Nevertheless, the number of high frequency peaks for all participants are generally higher in VRF1 compared to VRF3. Again this result indicates that physiological measures also yield results that VFR1 produced a higher arousal when compared to VFR2 and VFR3 and VFR3 produced a higher arousal when compared to VFR2.

Table 5.

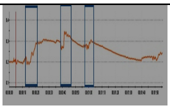
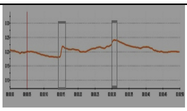
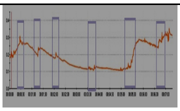
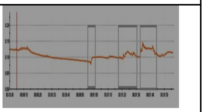
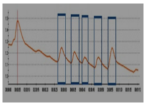
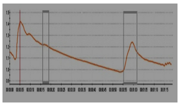
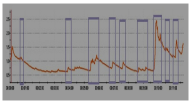
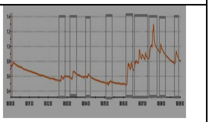
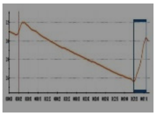
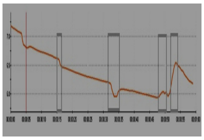
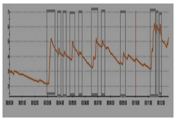
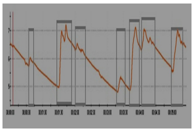
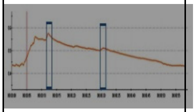
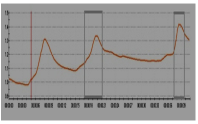
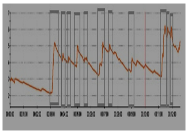
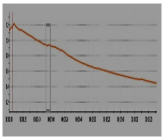
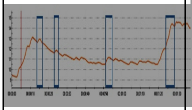
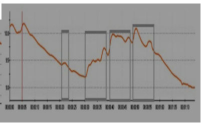
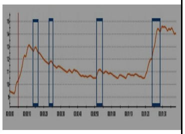
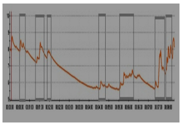
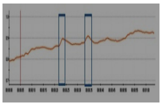
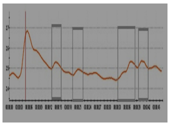
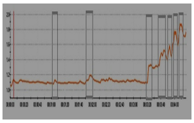
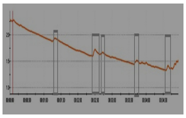
Partici-pants	**SCR Graph for Task 1**		**SCR Graph for task 2**	
	VFR1	VFR2	VFR1	VFR2
P1				
P2				
P3				
P4				
P5				
P6				

Table 6.

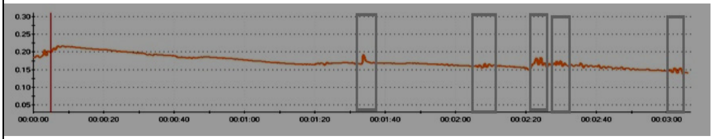
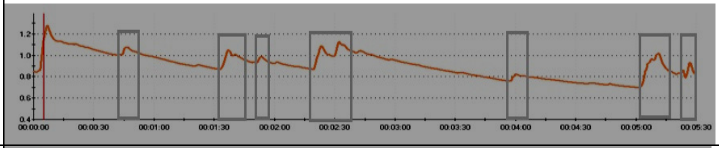
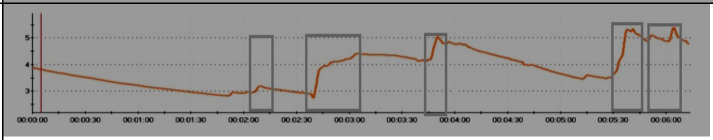
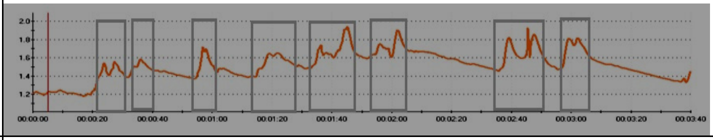
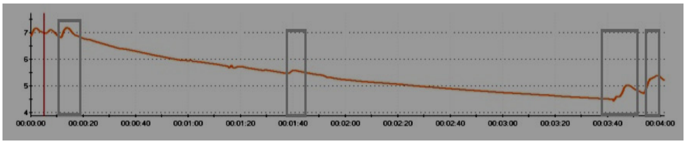
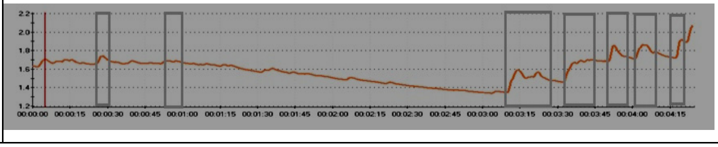
Partici-pants	**VFR3 (tempo – fast; sequence – disorder)**
P1	
P2	
P3	
P4	
P5	
P6	

Table 7.

Participants	Number of High Frequency Peak				
	Task 1		Task 2		Combined Task
	VFR1	VFR2	VFR1	VFR2	VFR3
P1	3	2	6	3	5
P2	5	2	9	7	7
P3	1	4	11	5	5
P4	2	2	10	8	8
P5	4	4	16	4	4
P6	2	4	7	7	7

4.4 Discussion

The results of obtained in this work shows a similarity in the measures of emotional response towards the movement quality exhibited in the activities of the VFR application. The results consistently indicate that movement quality with graceful element of tempo-fast produced a higher arousal as compared to tempo-slow. As for graceful element of sequence, sequence-order produced a higher arousal as compared to sequence-disorder. This finding is support theoretical models aesthetics by Birkhoff' and Klinger and Salingaros. Though these models were used to describe aesthetics of static form, this work shows the same can be applied to dynamic forms of aesthetics such as graceful interaction. The findings showed that the non-verbal measure of expressed emotional appeal of graceful interaction is in congruent with the measure of verbal expressed emotional appeal. Nevertheless, more analysis can performed based on the data of the physiology measures of skin conductance. For instance an analysis of emotional response for each task performed by the participants can be analyzed. This is more is more difficult to do when using self report assessment where participants may be unsure on the ratings to be given as they find it difficult to differentiate each tasks.

Although this work shows some promising results to determine suitable evaluation methods for graceful interaction design, the study has its limitation. This is because the VFR applications used are readily available applications which do not incorporate all four design elements of graceful interaction. The study is limited to the design element of temp and sequence only.

5 Conclusion

The usual method of evaluating user experience is based on performance metrics such task completion time seems cold and unfeeling and is not suitable for evaluating aesthetics experience. Other methods suggested in the literature include the arousal measure either through the use of verbal and non-verbal expressed emotional appeal. For the non-verbal expressed emotional appeal, biophysical data can reflect the arousal that takes place during the interaction. In this work we have shown that the measures from biophysical data are in congruent with the data obtained from the verbal expressed emotional appeal which is the more method of measuring arousal. As non-verbal expressed measures cannot be faked it may be a more reliable measure for the movement quality of graceful interaction. This work is an early effort to determine methods of evaluation of graceful interaction. More work is needed to explore on measures for the other graceful design elements which are rhythm and direction.

Acknowledgement. This work benefitted from the funding of the Ministry of Education, Malaysia under the grant 600-RMI/ERGS 5/3 (5/2011). We are indebted to the Research Management Institute of Universiti Teknologi MARA for their support in our work.

References

1. Löwgren, J.: From HCI to Interaction Design in Human Computer Interaction. Issues and Challenges, pp. 29–43. IGI Global Publishing (2001)
2. Löwgren, J., Stoltment, E.: Thoughtful Interaction Design: A Design Perspective on Information Technology. MIT Press (2004)
3. Hashim, W.N.W., Noor, N.L.M., Adnan, W.A.W.: The Design of Aesthetic Interaction: Towards a Graceful Interaction Framework. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human 200, Seoul, Korea (2009)
4. Schlosberg, H.: 3-Dimensions of Emotions. Psychological Review 61(2), 81–88 (1954)
5. Hayes, P., Reddy, R.: Graceful Interaction in Man-Machine Communication. In: Proceedings of the UCAI (1979)
6. Wiberg, P.: Graceful Interaction in Intelligent Environments. In: Proceedings of the International Symposium on Intelligent Environments, Cambridge (2006)
7. Noor, N.L.M., Hashim, W.N.W., Adnan, W.A.W., Saman, F.M.: Mapping Graceful Interaction Design from Dance Performance. In: Kurosu, M. (ed.) Human-Computer Interaction, Part III, HCII 2014. LNCS, vol. 8512, Springer, Heidelberg (2014)
8. Hashim, W.N.W., Noor, N.L.M., Adnan, W.A.W., Saman, M.F.: Graceful Interaction Design: Measuring Emotional Response towards Movement Quality. In: Proceedings of the International Conference on User Science and Engineering (i-USEr 2011), Subang Jaya, Malaysia (2011)
9. Xenakis, I., Arnellos, A., Darzentas, J.: The Functional Role of Role of Emotions in Aesthetic Judgment. New Ideas in Psychology 30, 212–226 (2012)
10. Filonik, D., Baur, D.: Measuring Aesthetics for Information Visualization. In: Proceedings of the 2009 13th International Conference on Information Visualization. IEEE Computer Society, Washington, DC (2009)
11. Fazio, R.H.: On the Automatic Activation of Association Evaluation. Cognition and Emotion 15(2), 115–141 (2001)
12. Desmet, P.: Measuring Emotion: Development and Application of an Instrument to measure emotional response to products. In: Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C. (eds.) Funology: From Usability to Enjoyment. Springer (2003)
13. Latinjak, A.: The Underlying Structure Of Emotions: A Tri-Dimensional Model Of Core Affect and Emotion Concepts For Sports. Revista Iberoamericana De Psicología Del Ejercicio Y El Deporte 7(1), 71–87 (2013)
14. Figner, B., Murphy, R.O.: Using Skin Conductance in Judgment and Decision Making Research. In: Schulte-Mecklenbeck, M., Kuehberger, A., Ranyard, R. (eds.) A Handbook of Process Tracing Methods for Decision Research, pp. 163–184. Psychology Press, New York (2010)

Author Index