Masaaki Kurosu (Ed.)

LNCS 10901

Human-Computer Interaction

Theories, Methods, and Human Issues

20th International Conference, HCI International 2018
Las Vegas, NV, USA, July 15–20, 2018
Proceedings, Part I

1 Part I

HCI 2018
INTERNATIONAL

Springer

Lecture Notes in Computer Science 10901

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, Lancaster, UK
Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA
Josef Kittler
University of Surrey, Guildford, UK
Jon M. Kleinberg
Cornell University, Ithaca, NY, USA
Friedemann Mattern
ETH Zurich, Zurich, Switzerland
John C. Mitchell
Stanford University, Stanford, CA, USA
Moni Naor
Weizmann Institute of Science, Rehovot, Israel
C. Pandu Rangan
Indian Institute of Technology Madras, Chennai, India
Bernhard Steffen
TU Dortmund University, Dortmund, Germany
Demetri Terzopoulos
University of California, Los Angeles, CA, USA
Doug Tygar
University of California, Berkeley, CA, USA
Gerhard Weikum
Max Planck Institute for Informatics, Saarbrücken, Germany

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

Masaaki Kurosu (Ed.)

Human-Computer Interaction

Theories, Methods, and Human Issues

20th International Conference, HCI International 2018
Las Vegas, NV, USA, July 15–20, 2018
Proceedings, Part I

Springer

Editor
Masaaki Kurosu
The Open University of Japan
Chiba
Japan

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-91237-0 ISBN 978-3-319-91238-7 (eBook)
https://doi.org/10.1007/978-3-319-91238-7

Library of Congress Control Number: 2018942175

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

© Springer International Publishing AG, part of Springer Nature 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

The 20th International Conference on Human-Computer Interaction, HCI International 2018, was held in Las Vegas, NV, USA, during July 15–20, 2018. The event incorporated the 14 conferences/thematic areas listed on the following page.

A total of 4,373 individuals from academia, research institutes, industry, and governmental agencies from 76 countries submitted contributions, and 1,170 papers and 195 posters have been included in the proceedings. These contributions address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The contributions 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. The volumes constituting the full set of the conference proceedings are listed in the following pages.

I would like to thank the program board chairs and the members of the program boards of all thematic areas and affiliated conferences for their contribution to the highest scientific quality and the overall success of the HCI International 2018 conference.

This conference would not have been possible without the continuous and unwavering support and advice of the founder, Conference General Chair Emeritus and Conference Scientific Advisor Prof. Gavriel Salvendy. For his outstanding efforts, I would like to express my appreciation to the communications chair and editor of *HCI International News*, Dr. Abbas Moallem.

July 2018 Constantine Stephanidis

HCI International 2018 Thematic Areas and Affiliated Conferences

Thematic areas:

- Human-Computer Interaction (HCI 2018)
- Human Interface and the Management of Information (HIMI 2018)

Affiliated conferences:

- 15th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2018)
- 12th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2018)
- 10th International Conference on Virtual, Augmented, and Mixed Reality (VAMR 2018)
- 10th International Conference on Cross-Cultural Design (CCD 2018)
- 10th International Conference on Social Computing and Social Media (SCSM 2018)
- 12th International Conference on Augmented Cognition (AC 2018)
- 9th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management (DHM 2018)
- 7th International Conference on Design, User Experience, and Usability (DUXU 2018)
- 6th International Conference on Distributed, Ambient, and Pervasive Interactions (DAPI 2018)
- 5th International Conference on HCI in Business, Government, and Organizations (HCIBGO)
- 5th International Conference on Learning and Collaboration Technologies (LCT 2018)
- 4th International Conference on Human Aspects of IT for the Aged Population (ITAP 2018)

Conference Proceedings Volumes Full List

1. LNCS 10901, Human-Computer Interaction: Theories, Methods, and Human Issues (Part I), edited by Masaaki Kurosu
2. LNCS 10902, Human-Computer Interaction: Interaction in Context (Part II), edited by Masaaki Kurosu
3. LNCS 10903, Human-Computer Interaction: Interaction Technologies (Part III), edited by Masaaki Kurosu
4. LNCS 10904, Human Interface and the Management of Information: Interaction, Visualization, and Analytics (Part I), edited by Sakae Yamamoto and Hirohiko Mori
5. LNCS 10905, Human Interface and the Management of Information: Information in Applications and Services (Part II), edited by Sakae Yamamoto and Hirohiko Mori
6. LNAI 10906, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris
7. LNCS 10907, Universal Access in Human-Computer Interaction: Methods, Technologies, and Users (Part I), edited by Margherita Antona and Constantine Stephanidis
8. LNCS 10908, Universal Access in Human-Computer Interaction: Virtual, Augmented, and Intelligent Environments (Part II), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 10909, Virtual, Augmented and Mixed Reality: Interaction, Navigation, Visualization, Embodiment, and Simulation (Part I), edited by Jessie Y. C. Chen and Gino Fragomeni
10. LNCS 10910, Virtual, Augmented and Mixed Reality: Applications in Health, Cultural Heritage, and Industry (Part II), edited by Jessie Y. C. Chen and Gino Fragomeni
11. LNCS 10911, Cross-Cultural Design: Methods, Tools, and Users (Part I), edited by Pei-Luen Patrick Rau
12. LNCS 10912, Cross-Cultural Design: Applications in Cultural Heritage, Creativity, and Social Development (Part II), edited by Pei-Luen Patrick Rau
13. LNCS 10913, Social Computing and Social Media: User Experience and Behavior (Part I), edited by Gabriele Meiselwitz
14. LNCS 10914, Social Computing and Social Media: Technologies and Analytics (Part II), edited by Gabriele Meiselwitz
15. LNAI 10915, Augmented Cognition: Intelligent Technologies (Part I), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
16. LNAI 10916, Augmented Cognition: Users and Contexts (Part II), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
17. LNCS 10917, Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management, edited by Vincent G. Duffy
18. LNCS 10918, Design, User Experience, and Usability: Theory and Practice (Part I), edited by Aaron Marcus and Wentao Wang

19. LNCS 10919, Design, User Experience, and Usability: Designing Interactions (Part II), edited by Aaron Marcus and Wentao Wang
20. LNCS 10920, Design, User Experience, and Usability: Users, Contexts, and Case Studies (Part III), edited by Aaron Marcus and Wentao Wang
21. LNCS 10921, Distributed, Ambient, and Pervasive Interactions: Understanding Humans (Part I), edited by Norbert Streitz and Shin'ichi Konomi
22. LNCS 10922, Distributed, Ambient, and Pervasive Interactions: Technologies and Contexts (Part II), edited by Norbert Streitz and Shin'ichi Konomi
23. LNCS 10923, HCI in Business, Government, and Organizations, edited by Fiona Fui-Hoon Nah and Bo Sophia Xiao
24. LNCS 10924, Learning and Collaboration Technologies: Design, Development and Technological Innovation (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
25. LNCS 10925, Learning and Collaboration Technologies: Learning and Teaching (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
26. LNCS 10926, Human Aspects of IT for the Aged Population: Acceptance, Communication, and Participation (Part I), edited by Jia Zhou and Gavriel Salvendy
27. LNCS 10927, Human Aspects of IT for the Aged Population: Applications in Health, Assistance, and Entertainment (Part II), edited by Jia Zhou and Gavriel Salvendy
28. CCIS 850, HCI International 2018 Posters Extended Abstracts (Part I), edited by Constantine Stephanidis
29. CCIS 851, HCI International 2018 Posters Extended Abstracts (Part II), edited by Constantine Stephanidis
30. CCIS 852, HCI International 2018 Posters Extended Abstracts (Part III), edited by Constantine Stephanidis

http://2018.hci.international/proceedings

Human-Computer Interaction

Program Board Chair: **Masaaki Kurosu, *Japan***

- Jose Abdelnour-Nocera, UK
- Mark Apperley, New Zealand
- Sebastiano Bagnara, Italy
- Kaveh Bazargan, Iran
- Nigel Bevan, UK
- Michael Craven, UK
- Achim Ebert, Germany
- Xiaowen Fang, USA
- Carla Faria Leitão, Brazil
- Stefano Federici, Italy
- Isabela Gasparini, Brazil
- Ayako Hashizume, Japan
- Wonil Hwang, South Korea
- Mitsuhiko Karashima, Japan
- Heidi Krömker, Germany
- Kun-Pyo Lee, South Korea
- Cristiano Maciel, Brazil
- Paulo Melo, Brazil
- Naoko Okuizumi, Japan
- Katsuhiko Onishi, Japan
- Philippe Palanque, France
- Roberto Pereira, Brazil
- Denise Pilar, Brazil
- Alberto Raposo, Brazil
- Guangfeng Song, USA
- Hiroshi Ujita, Japan
- Michiya Yamamoto, Japan
- Fan Zhao, USA

The full list with the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences is available online at:

http://www.hci.international/board-members-2018.php

HCI International 2019

The 21st International Conference on Human-Computer Interaction, HCI International 2019, will be held jointly with the affiliated conferences in Orlando, FL, USA, at Walt Disney World Swan and Dolphin Resort, July 26–31, 2019. It will cover a broad spectrum of themes related to Human-Computer Interaction, including theoretical issues, methods, tools, processes, and case studies in HCI design, as well as novel interaction techniques, interfaces, and applications. The proceedings will be published by Springer. More information will be available on the conference website: http://2019.hci.international/.

General Chair
Prof. Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
E-mail: general_chair@hcii2019.org

http://2019.hci.international/

Contents – Part I

HCI Theories, Methods and Tools

Security, Privacy and Ethics in HCI

Contents – Part II

HCI in Medicine

HCI for Health and Wellbeing

HCI in Cultural Heritage

HCI in Complex Environments

Contents – Part III

Input Techniques and Devices

Speech-Based Interfaces and Chatbots

HCI Theories, Methods and Tools

What Are User Requirements? Developing an ISO Standard

Nigel Bevan[1](✉), Jim Carter[2], Jonathan Earthy[3], Thomas Geis[4], and Susan Harker[5]

[1] Professional UX Services, 12 King Edwards Gardens, London W3 9RG, UK
mail@nigelbevan.com
[2] Computer Science Department, University of Saskatchewan, Saskatoon S7N 5C9, Canada
carter@cs.usask.ca
[3] Lloyd's Register Global Technology Centre, Boldrewood Innovation Campus, Burgess Road, Southampton SO16 7QF, UK
Jonathan.Earthy@lr.org
[4] ProContext Consulting GmbH, Unter Käster 14-16, 50667 Cologne, Germany
Thomas.Geis@procontext.de
[5] Loughborough Design School, Loughborough University, Loughborough LE11 3TU, UK
S.D.Harker@lboro.ac.uk

Abstract. Although much has been published on how to collect user requirements, there is surprisingly little guidance on the specific information that should be included in a user requirements specification, or on the syntax of user requirements statements. An ISO working group that has been developing a series of documents to define good practice for the content of human-centred design deliverables is now working to get consensus on the content of user requirements specifications. Two types of user requirements have been identified: (a) requirements for a user to be able to recognize, select, input or receive physical entities and information, and (b) use-related quality requirements that specify criteria for outcomes such as effectiveness, efficiency, satisfaction, accessibility, user experience and avoidance of harm from use. A user requirements specification should also contain information about constraints, the context of use, goals and tasks to be supported, design guidelines and any recommendations for design solutions emerging from the user requirements. This paper explains some of the challenges in developing the standard.

Keywords: Requirements · User requirements · Standards

1 Introduction

Understanding user requirements and making them available as part of the development process is a key activity in human-centred design. It provides the basis for an appropriate design solution and its evaluation. Without proper statements of user requirements, the development process cannot be informed about what is required from the perspective of the use of the interactive system. Existing published approaches describe a process and the methods that can be used to gather information about users and their tasks, e.g. [1, 7, 23, 24]. For example, commonly used methods are contextual interviews, surveys,

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 3–13, 2018.
https://doi.org/10.1007/978-3-319-91238-7_1

user needs analysis, card sorting, group task analysis, focus groups and field studies. But what are the resulting requirements that provide a basis for design and evaluation of the system from the perspective of its users, and how should these be specified? User requirements are often ignored.

There is extensive literature on what aspects of usability and user experience can be evaluated, but these are rarely expressed in advance as requirements for the design of the interactive system [1]. Cohelo emphasizes the need to specify 'experience' requirements relating to users' expected perceptions and responses about a system or service in addition to task-related requirements and usability requirements [5].

When a new system is being commissioned, a general requirements analysis will take place which will include the business goals and the requirements of the identified stakeholders. One consideration is how user requirements will be differentiated from other types of requirements (e.g. market requirements and organizational requirements). A distinction that is often made is between functional requirements (which specify what the system should do) and non-functional (which specifies how well it does those things) [22, 26]. The non-functional requirements are often considered to include usability and other quality requirements [21]. However, the term "non-functional" is rather disparaging, and not well-defined, as "non-functional" requirements need functions to be specified to implement them. Quality requirements provide a more positive concept that is widely used, for example in ISO/IEC 25010 "System and software quality models".

2 Standards Containing Guidance on User Requirements

The ISO standard for usability, ISO 9241-11 (1998) defined usability as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". The standard contains an example of a usability requirements specification that defines the intended goals and context of use, and specifies measures and criterion levels for effectiveness, efficiency and satisfaction for the product under development. In 2001 an industry working group led by the US National Institute of Standards and Technology (NIST) published a Common Industry Format for documenting the results of a summative usability evaluation of effectiveness, efficiency and satisfaction [2] to enable the usability test results for products to be compared. This was subsequently adopted by ISO as ISO/IEC 25062 [16]. Although the document does not explicitly mention requirements, it implicitly provides a basis for specifying usability requirements. A parallel document, the Common Industry Specification for Usability—Requirements (2007) was subsequently developed by NIST for specifying usability requirements [25], highlighting the value of high-level quantitative requirements for effectiveness, efficiency and satisfaction. Based on this experience, the EU PRUE project successfully trialled the use of usability requirements in several companies [3].

3 What Are User Requirements?

There are now many papers and textbooks that explain how to collect user requirements (e.g. [1, 7, 23, 24]), but very little guidance about what information to include in user requirements, or how they should be specified.

There is also some confusion between the meaning of “user requirements” and “usability requirements”. For example, the Specification for Usability Requirements published by NIST [25] included in addition to the components of usability (effectiveness, efficiency and satisfaction), other types of requirements:

Design guidance

(a) Design principles
(b) Human factors and ergonomics
(c) Style guides
(d) Standards

Usability features

(a) Accessibility
(b) Understandability
(c) Learnability
(d) Operability
(e) Attractiveness

Content and functions

(a) Functionality
(b) Content
(c) Complete, accurate, up to date, style
(d) Effectiveness (for learning)
(e) Trust
(f) Platform independence

These types of requirements would now be commonly regarded as “user requirements”.

4 First Draft of the Standard for User Requirements

An ISO joint working group, with experts from the systems and software engineering and ergonomics committees, has been developing a series of standard Common Industry Formats for the usability-related information that is produced and used during systems development. The documents that have been produced so far are:

- ISO/IEC 25060 (2010) General framework for usability-related information
- ISO/IEC 25063 (2014) Context of use description
- ISO/IEC 25064 (2013) User needs report
- ISO/IEC 25065 (2016) Evaluation report

ISO/IEC 25064 describes what should be included in a user needs report, and notes that "the user needs report is a critical input into specifying user requirements". However, it does not explain how user requirements differ from user needs, and why an additional step of defining user requirements is necessary prior to implementation of the system. Some authors clearly differentiate between user needs and user requirements [7].

As long ago as 2010, the joint working group agreed a definition of user requirements: "requirements for use that provide the basis for design and evaluation of interactive systems to meet identified user needs", and started work on the ISO/IEC 25065 standard for the contents of user requirements specifications. But, despite extensive discussion, development was terminated after two years because no consensus had been reached on the scope of user requirements and how to classify them. The project was restarted in 2014, positioning user requirements in relation to other types of requirements.

The initial draft for ballot (ISO/IEC CD 25065, issued in March 2017) explained that system requirements describe what the system has to do and to what extent it should do it, in order to meet each individual stakeholder requirement.

System requirements for the technical solution "specify, from the supplier's perspective, what characteristics, attributes, and functional and performance requirements the system is to possess, in order to satisfy stakeholder requirements" (ISO/IEC 15288). Stakeholder requirements describe what is required from the viewpoint of each individual stakeholder group. Stakeholder requirements "express the intended interaction the system will have with its operational environment and that are the reference against which each resulting operational capability is validated" (ISO/IEC 15288), as illustrated by Fig. 1 from ISO/IEC 29148:2011.

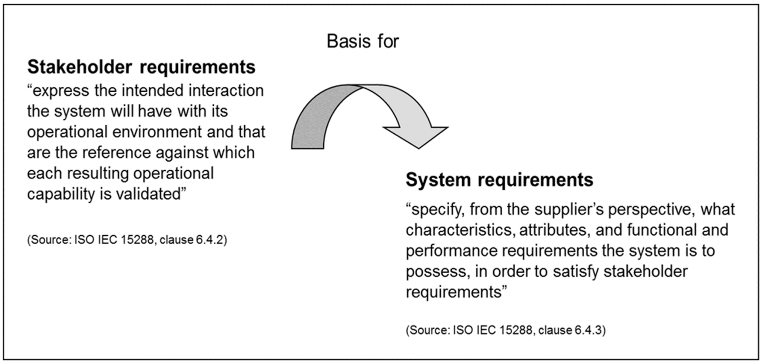


Fig. 1. Stakeholder requirements in relation to system requirements.

User requirements are one type of stakeholder requirement [8]. They provide a basis for system requirements from the viewpoint of the user of the interactive system.

According to the first draft, user requirements specify what a user shall be able to do and/or experience with the system in each particular context of use in order to satisfy one or more user needs. The purpose of user requirements was given as enabling the design of a system with which users could achieve levels of usability, accessibility, user experience and/or avoidance of harm from use (referred to as human-centred quality in ISO 9241-220). Human-centred quality objectives are high-level project objectives that can subsequently form the basis for more specific acceptance criteria for the system. Although they were included in a user requirements specification, they were labelled as objectives rather than requirements.

User requirements themselves were requirements for: a user to be able to recognize specific information in the interactive system (e.g. departure times of trains); or to be able to input a physical entity (e.g. coins) or information (user's age); or to be able to select a physical entity or information (e.g. destination); or to have a specific perception or response (i.e. user experience, e.g. a sense of enjoyment).

Stakeholder requirements other than user requirements can be sources for user requirements as shown in Fig. 2 [8]. All stakeholder requirements are intended to serve as the basis for deriving system requirements.

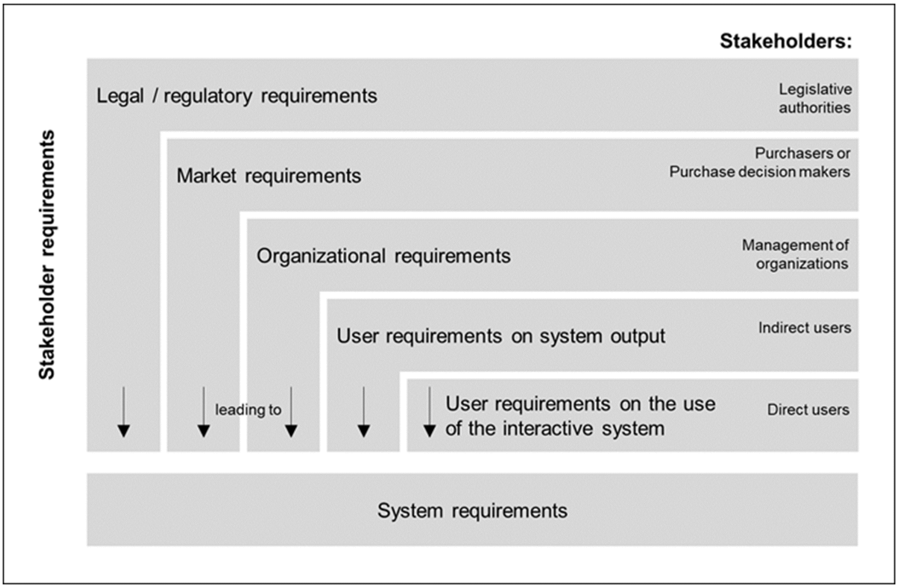


Fig. 2. Types of stakeholder requirements leading to system requirements

Legal requirements (also called regulatory requirements) are constraints set by legal entities (e. g. the Government or a government institution). They are contained in laws and other regulatory documents including harmonized standards. Legal requirements often enforce processes within specific types of organizations. For example: "Section 508 requires that when federal agencies develop, procure, maintain, or use electronic and information technology, federal employees with disabilities have access to and use of information and data that is comparable to the access and use by federal employees who are not individuals with disabilities, unless an undue burden would be imposed on the agency." (Section 508 in the USA).

Market requirements are set by manufacturers and describe which capabilities the system shall possess or meet, so that purchasers, or purchase decision makers, buy the product (or in case of free-of-charge products decide to use the product). For example: “The smart phone shall be cheaper than the equivalent model from a major competitor”.

Organizational requirements are requirements on the behaviour of the organization and on the humans within organizations that describe how people within the organization have to act when performing their tasks. For example: “The physician shall wear gloves during surgery” or “The sales representative shall get quotations higher than 100.000 EUR signed off by the sales director before sending them to the client”.

User requirements on the system output prescribe the required outputs of the interactive system and the attributes of theses outputs (including the accuracy) that these outputs shall have (where applicable). For example: “The invoice produced by the system shall contain the contract number that it relates to”, or “The hard-boiled egg produced by the system shall not contain any liquid egg yolk”).

User requirements are typically included as part of a stakeholder requirements specification document as described in ISO/IEC/IEEE 29148 (Requirements Engineering).

Figure 3 from the draft illustrates the suggested relationship between user requirements and other information items related to human-centred design.

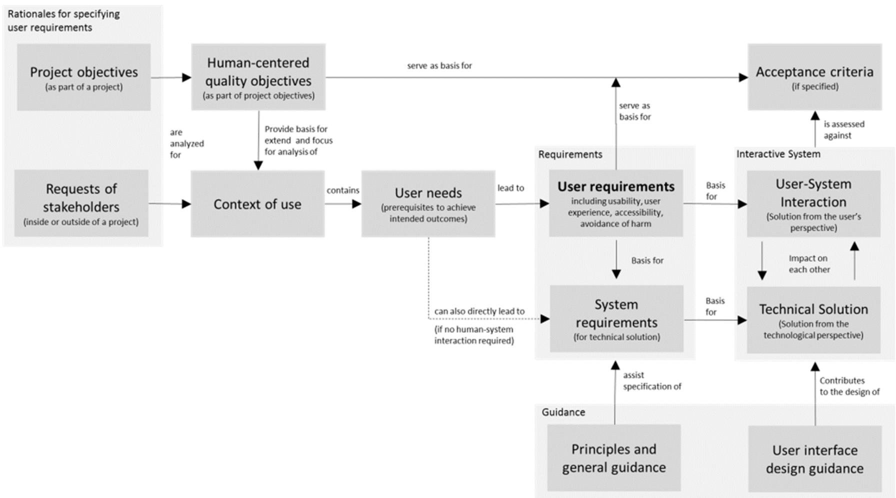


Fig. 3. The interrelationship between user requirements and other information items related to human-centred design.

5 Second Draft of the Standard for User Requirements

The second draft, ISO/IEC CD 25065.2, was issued in January 2018. Changes include:

– The interpretation of user requirements has been broadened from what a user shall be able to do and/or experience with the system to include requirements for "use-related qualities" (such as the usability or accessibility) with which intended outcomes are achieved using the interactive system (see Sect. 5.3).
– The standard focusses on the content of user requirements specifications, removing the explanation of the role of user requirements in systems development (as this is beyond the intended purpose of the standard).

5.1 Content of a User Requirements Specification

The second draft states that the following information should be included in a user requirements specification:

– Interactive system for which the user requirements are specified.
– Any constraints in terms of factors known to limit the freedom of design and implementation of solutions to satisfy the user requirements and the interactive system to be developed. These include technical, budget, time, legal, environmental, social and organizational constraints. (This differs from a view commonly held by developers that user requirements are constraints on the freedom of design and implementation of solutions to satisfy the functional requirements.)
– The overall context of use: the users, goals and tasks, resources, and environment for use of the interactive system (this can be in a separate document). It specifies the contexts of use in which the system is required to be usable.
– Goals (intended outcomes) and the tasks to be supported. Goals are the intended outcome(s) to be achieved using the interactive system. Goals are independent of the means used to achieve them. Goals focus on what is to be achieved without necessarily specifying criteria (such as levels of effectiveness, efficiency or satisfaction). Tasks consist of one or more activities undertaken to achieve a goal. Different combinations of activities can provide different ways of achieving the same goal and can lead to different levels of usability.
– User requirements
  • User-system interaction requirements (see Sect. 5.2).
  • Use-related quality requirements (see Sect. 5.3).

Each user requirement should include:

• The user group(s) to which the user requirement applies.
• The goal(s) or task(s) to which the user requirement applies.
• The condition(s) (including relevant aspects of the context of use) under which the user requirement applies.
  – Design guidance to be applied (such as a style guide).
  – If appropriate, recommendations for design solutions emerging from the user requirements.

5.2 User-System Interaction Requirements

User-system interaction requirements specify the outcome of use, in terms of required interaction, e.g. to be able to:

– Recognize specific information in the interactive system (e.g. departure times of trains).
– Input a physical entity (e.g. coins) or information (user's age).
– Select a physical entity or information (e.g. destination).
– Receive (take away) a physical entity (e.g. the printed ticket) or information from the interactive system (an electronic receipt).

The following syntax is suggested to phrase use-related quality requirements: "With the <interactive system> the <user group> shall be able to achieve <outcome in terms of interaction> under <condition(s)> (if applicable)."

5.3 Use-Related Quality Requirements

Use-related quality requirements specify criteria for components of use-related quality:

– Effectiveness (e.g. extent to which the alarm is set correctly).
– Efficiency (e.g. time taken to set the alarm).
– Satisfaction (e.g. user feels secure that he will wake up as intended).
– Other components of use-related quality (e.g. accessibility, user experience, avoidance of harm from use).

These requirements can be at the level of the overall task outcome (e.g. setting an alarm on a clock) or when appropriate be for elements of user-system interaction (e.g. the effectiveness or efficiency of inputting or selecting information).

5.4 Structuring User Requirements Within a User Requirements Specification

Use-related quality requirements can apply to use of the whole system or product, as well as to the achievement of sub-goals and sub-tasks. User-system interaction requirements typically specify interaction at the lowest level of tasks or sub-tasks, which describe required interaction with the user interface. User requirements should be structured by the goals and tasks to be supported by the interactive system rather than by the characteristics of the system.

6 Conclusions

6.1 Benefits

ISO/IEC 25065 helps differentiate two concepts that are often confused: user requirements and user needs (user needs are already described in ISO/IEC 25064).

Reaching an agreement on the meaning and content of user requirements has not been easy. The current draft identifies the information to be included in a specification of requirements for the user interactions with, and the interfaces of, interactive systems. This includes two important types of user requirements at the detailed level of user-system interaction: (a) what requirements does a user have to be able to recognize, select, input or receive information or a physical entity? And (b) are there any quality requirements (effectiveness, efficiency, satisfaction, user experience, accessibility or avoidance of harm from use), for any of these detailed elements of interaction, or as evidence of achievement of higher level goals? Specification of quality requirements is likely to be reserved for aspects of interaction where particular levels of quality (such as efficiency, accessibility or pleasure from use) are important for the success of the system and the specification (and potential evaluation) of these requirements is necessary to ensure that they are achieved.

Uniformity and precision in the definition of user requirements is beneficial in the specification of requirements in both formal and less formal development environments. ISO/IEC 25065 helps to align user requirements with the terminology and processes used in standards such as ISO/IEC 15288 and 29148 which are used to specify requirements, for large, critical and complex systems.

6.2 Types of User Requirements

The focus of the current draft of ISO/IEC 25065 is on two types of user requirements: user-system interaction requirements, and use-related quality requirements for task and sub-task outcomes. There are other types of requirements related to use that are frequently identified within design activities, but appear to be outside the scope of either type described in the standard, such as:

- Requirements for the project that would make it possible for the project to develop systems with human-centred quality (such as the plans, activities, methods, staff competence and resources).
- User requirements for the enabling systems that realise, operate and maintain the system of interest, (e.g. production, installation, training, maintenance, support/supply/inform, management/command/governance etc.).
- User requirements for properties of the system that go beyond a specific user interaction (e.g. "When using the transport system, the user shall be able to use the same ticket when transferring from metro to bus").
- User requirements for functionality (e.g. "The user shall be able to specify a keyboard shortcut for every menu item").
- User requirements specifically relating to the context of use (e.g. "The user shall be able to use the product in the dark").

It would be useful to clarify how these requirements relate to the scope and content of the current standard, in order to either refine this standard and/or identity the need for further standardisation relating to user requirements. This might be within the CIF series of standards for stating usability-related information, or in the ISO 9241-200 series of standards that address the processes, activities and methods used to generate and transform that information.

6.3 Participation in Standards Development

If you would like to contribute to the future development of this or other standards related to usability [4], or to comment on drafts, you can either do this via your national standards body [10], or if you are a member of one of the ISO TC159/SC4 liaison organisations [11] such as UXPA [27] you can participate through the liaison organisation.

Acknowledgements. The authors of this paper are members of the joint working group (JWG28) between ISO TC159/SC4 and ISO/IEC JTC1/SC7 that is developing ISO/IEC 25065.

References

1. Bargas-Avila, J.A., Hornbæk, K.: Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011), pp. 2689–2698. ACM (2011)
2. Bevan, N.: Industry standard usability tests. In: Brewster, S., Cawsey, A., Cockton, G. (eds.) Human-Computer Interaction – INTERACT 1999, vol. II, pp. 107–108. British Computer Society (1999)
3. Bevan, N., Claridge, N., Maguire, M., Athousaki, M.: Specifying and evaluating usability requirements using the common industry format: four case studies. In: Proceedings of the IFIP 17th World Computer Congress - TC13 Stream on Usability: Gaining a Competitive Edge, pp. 149–159 (2002)
4. Bevan, N., Earthy, J.: Benefiting from ISO standards. In: Norman, K.L., Kirakowski, J. (eds.) The Wiley Handbook of Human Computer Interaction, pp. 51–69. Wiley, New York (2018)
5. Coelho, D.A.: Specification of affective user experience in product design. In: Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014, pp. 2216–2224 (2014)
6. Courage, C., Baxter, K.: Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques, 1st edn. Morgan Kaufmann Publishers Inc., San Francisco (2005)
7. Geis, T., Dzida, W., Redtenbacher, W.: Specifying usability requirements and test criteria for interactive systems. Federal Institute for Occupational Safety and Health, Germany (2003)
8. Geis, T., Polkehn, K.: Praxiswissen User Requirements - Nutzungsqualität systematisch, nachhaltig und agil in die Produktentwicklung integrieren, dpunkt.verlag Heidelberg, Germany (2018)
9. International Usability and User Experience Qualification Board UXQB: Certified Professional for Usability and User Experience – Advanced Level "User Requirements Engineering" (2015). www.uxqb.org/en/documents. Accessed Feb 2018
10. ISO: Members. www.iso.org/iso/home/about/iso_members.htm
11. ISO: TC159/SC4 Ergonomics of Human-System Interaction. www.iso.org/committee/53372.html. Accessed Feb 2018
12. ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) —Part 11: Guidance on usability (1998)
13. ISO 9241-220: Ergonomics of human-system interaction—Part 220: Processes for enabling, executing and assessing human-centred design within organizations (2018)
14. ISO/IEC 15288: Systems and software engineering—System life cycle processes (2015)
15. ISO/IEC 25010: Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models (2011)

16. ISO/IEC 25062: Software Engineering—Software product Quality Requirements and Evaluation (SQuaRE)—Common Industry Format (CIF) for Usability Test Reports (2006)
17. ISO/IEC 25063: Systems and software engineering—Systems and software product Quality Requirements and Evaluation (SQuaRE)—Common Industry Format (CIF) for usability: Context of use description (2014)
18. ISO/IEC 25064: Systems and software engineering—Systems and software product Quality Requirements and Evaluation (SQuaRE)—Common Industry Format (CIF) for usability: Context of use description (2014)
19. ISO/IEC CD 25065.2: Common Industry Format (CIF) for Usability: User requirements specification (2018)
20. ISO/IEC 25066: Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—Common industry Format for Usability—Evaluation Report (2016)
21. ISO/IEC/IEEE 29148: Systems and software engineering—Life cycle processes—Requirements engineering (2011)
22. Kotonya, G., Sommerville, I.: Requirements Engineering: Processes and Techniques, 1st edn. Wiley Publishing, New York (1998)
23. Maguire, M.: User-centred requirements handbook. EC Telematics Applications Programme, Project TE 2010 RESPECT (Requirements Engineering and Specification in Telematics), WP4 Deliverable D4.2, version 3.3 (1998)
24. Maguire, M., Bevan, N.: User requirements analysis: a review of supporting methods. In: Hammond, J., Gross, T., Wesson, J. (eds.) Proceedings of the IFIP 17th World Computer Congress - TC13 Stream on Usability: Gaining a Competitive Edge. ITIFIP, vol. 99, pp. 133–148. Springer, Boston (2002). https://doi.org/10.1007/978-0-387-35610-5_9
25. NIST: Common Industry Specification for Usability—Requirements (2007). www.nist.gov/itl/iad/iusr-papers-and-publications. Accessed Feb 2018
26. Robertson, S., Robertson, J.: Mastering the Requirements Process. ACM Press/Addison-Wesley Publ. Co., New York (1999)
27. UXPA: User Experience Professionals Association liaison with ISO. www.uxpa.org. Accessed Feb 2018

UCD: Influenced by an Organizational Culture and Its Maturity

Lara da Costa Brito(✉) and Maria Manuela Rupp Quaresma(✉)

Pontífice Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brazil
larbrito@hotmail.com, mquaresma@puc-rio.br

Abstract. This article proposes to discuss how user-centered design practices (UCD) can be influenced by certain organizational characteristics such as organizational culture and corporate maturity of user-centered design processes. Organizational culture exerts a strong power in all sectors of a company, either in the transmission of values, behaviors or habits and, in particular, the adoption of new practices and methodologies also do not escape from this rule and appear to be under the control of this culture which selects and designates what can and cannot be promoted within the corporation. In addition to corporate culture, the level of corporate maturity is also a determining factor on the implementation of successful processes and without adequate and mature understanding, can affect as a consequence, the results and quality of a products and services, thus compromising the sales, user's satisfaction, and for last, the market positioning.

Keywords: User-centered design · User-centered design process
Organizational culture · Corporate maturity

1 Introduction

Currently, companies have increased their concern with user satisfaction motivated mainly by market competition. In general, the user has gained a meaning related to the profit and as an investment is gradually becoming a key element in the development of projects, becoming a competitive differential.

The use of the term UX has been established in the corporate dictionary as synonymous of success to create attractive products and services, even if there is no consensus or certainty of its meaning. User-focused methodologies are currently one of the most researched practices in both the academic and corporate domains, making it of the utmost importance for project development. Everything looks perfect, but how are companies doing this job? Does organizational culture allow users to effectively participate in the contribution of their own experience? And as for the maturity of the business, is it enough to deal with methods and techniques with the goal of creating products and services that satisfy users?

This paper aims to discuss how user-centered design practices can be directly or indirectly affected by the organizational culture and the corporate maturity of the design process.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 14–23, 2018.
https://doi.org/10.1007/978-3-319-91238-7_2

2 UCD Practices in Organizations

2.1 Understanding UCD

Much is said about how design practices can help in the sale and retention of products and services, capturing and winning the interest of users. According to Courage and Baxter (2005, p. 8), thanks to market pressure and usability awareness, many companies link UCD processes to the life cycles of their products. However, a closer look shows that there are still gaps to be filled in the application of these practices that involve UCD design methodologies within companies.

Initially, it is needs to be a clear understanding of what involves user-centered design in the corporate environment and this includes the proper discernment among the various terms used to designate these activities. The term "user-centered design" is still misused today. Although it is considered a multidisciplinary design approach, some people refer to human-centered design of products and services as user experience, but according to Goodwin (2011, p. 5) this term is presumptuous; it is possible to design all aspects of the environment to encourage a great user experience, but as each one brings their own attitudes, behaviors and perceptions to any situation, no designer can determine exactly what the experience has.

In addition to the mistaken use of user-centered design, the terms usability, human-computer interaction, and user experience are mistaken for synonyms and misunderstandings. Lowdermilk (2013, p. 6) clarifies that differences from definitions: usability, also referred to as human factors, is the study of how humans relate to any product. Man-computer interaction is rooted in usability, but focuses on how humans relate to computing products. User-centered design has emerged from human-computer interaction and is a software design methodology for developers and designers where their practice ensures the application of good usability. With regard to the term user experience, so widely used these days, it is used to summarize the entire experience of a software product or service. It not only covers features, but also how engaging and enjoyable is the use of an application.

2.2 Adoption of UCD Practices

Understanding and implementing UCD should be an important practice in organizations to achieve the goal of designing products and services for their end users. However, Schaffer (2004, p. 16) reminds us the real question that companies often ask themselves: "Will people be converted?" For most organizations, it's an advantage if people love their projects, but making the sale happen will make the company's executives happier. For corporations, of course sales are an expected outcome, but in order to achieve this goal, they need to be aware of a number of factors, including the final quality of what they produce, which should be usable and meet the user's needs. It would seem obvious to think this way, but for some organizations this logic may not exactly match reality.

For Courage and Baxter (2005, p. 8), if the company does not adopt a design process, you have a major problem on your hands and performing some activities with users will not be helpful. You will need to employ a change management strategy to affect the structure, processes, and culture of your organization and these is far from easy job.

Generally, the adoption of methodologies will depend in particular on diverse organizational characteristics such as culture, process maturity, size and hierarchical structure:

- Organizational culture: its role is to determine how design methods and techniques are to be seen and valued in a company and its influence will be directly linked to promoting the evolution of the UCD design process through corporate maturity.
- Corporate maturity of the process: it is fundamental that the company is aware of the current level of maturity with respect to user-centered design processes and thus be able to generate action plans for its evolution in the maturity scale facing the market.
- Size: the size of the company will influence in the introduction of design methodologies and processes, and the larger the company size, the greater the complexity in the implementation of the process.
- Hierarchical structure: like the size of the company, the type of organizational hierarchy will also influence the introduction of new processes. In the case of a hierarchy with a horizontal structure, the methodological changes will be accepted and absorbed with greater ease, as opposed to the vertical structure, which will be subject to complications.

2.3 The Value of Design

For some companies, design methodologies may still have a basic vision within an organization. One of the reasons why organizations have given little importance to design methodologies, according to Mozota (1998), is the lack of understanding of their comprehensiveness. It is often seen as a resource of support rather than as a process. Goodwin (2011, p. 698) complements saying that becoming an organization that truly values design requires changes in structures, processes, and the attitudes and behavioral norms that constitute corporate culture. For Best (2009, p. 54), one of the biggest challenges for design in general is to measure its value. Success measures are easier to measure in qualitative terms (such as improved brand image, increased organizational learning and better communication) than in quantitative terms (profits, units sold, increased market share). In addition, the benefits of design methods and techniques usually reveal themselves over time, not instantly.

In order to clarify to business managers and designers the value that design as a discipline can bring to a company, Mozota (2006) proposes a value model as a differentiating, integrating, transforming and business element for a company:

- Design as differentiator: design as competitive advantage in the market through brand equity, customer loyalty and price or customer orientation.
- Design as an integrator: design as a resource that improves the development of new products, processes (market time, building teams that use data visualization skills); design as a process that favors product architecture, user-oriented innovation models, and project management.

- Design as transformer: design as a resource to create new business opportunities; to improve the company's ability to cope with change.
- Design as good business: design as sales increase; brand value; better Market Share, return on investment (ROI) and design as a resource for society in general.

Through this value model, we can see that design activities can be valued in a variety of ways, but understanding their meaning for a company is a crucial factor in determining these activities, as well as management activities.

3 Corporate Culture

3.1 The Culture Matters

Culture is an important bond between individuals and organizations. It is established through a group with common interests. When it is possible to align individual and organizational behaviors, common visions and values are created and help establish a culture where employees are allowed to aggregate values at all levels. To maintain employees motivation and commitment in fulfilling organizational purposes, Best (2009, p. 86), comments that even if the company has internal structures in place to manage processes and obtain goods and services for the market, companies are significantly dependent on the organization's culture.

For Schein (2009, p. 19) culture matters because it is a powerful, tacit set of forces that determine both our individual and collective behavior, ways of perceiving, thought patterns, and values. Organizational culture in particular matters because cultural elements determine strategy, goals, and modes of operating.

Culture connects and directs all individuals in a company to the same values and therefore has great responsibilities when it comes to spread their visions and goals. But to delivering visions, goals and objectives according to Best (2009, p. 86), requires the coordination and commitment of resources, processes and people, both horizontally and vertically, across the organization.

Culture is capable of controlling and designating practices and processes. Hartson and Pyla (2012, p. 62) comment that sometimes, organizations selects processes it will use based on its own tradition and culture, including how they operated in the past. In addition, certain types of organizations have their culture so deeply built that it determines the types of projects they can undertake.

3.2 Culture and UCD Practices

Organizational culture should play a fundamental role so that the design practices and process are not impaired in the development of its stages. Often investment in UCD practices is not understood by companies. The same happens with design professionals who do not know how to justify the value of their own work. According to Best (2009, p. 12), design and business have distinct cultures: their own beliefs, values, and assumptions about how they measure success and what matters to them. This can sometimes create a "clash of cultures." It becomes a powerful advantage if there is an understanding of the challenges and opportunities pertaining to different corporate cultures. If the UCD practices are devalued by the professionals and the organization, the design process is devalued as a whole. It is necessary to have responsibility and make clear the difference between the design practices and the business practices.

For Lund (2011, p. 238), companies must have the structure to manage their processes and offer products and services to customers and end users. But to meet the demand, companies will depend on the organizational culture to sustain their strategic objectives. Organizational culture should encourage employees to add value at all levels. Through the opinion of the authors, we see how this culture should be responsible for disseminating the value of its internal processes and demonstrate the strategic importance to satisfy the needs of the end user in their daily tasks. Organizational culture, therefore, determines how design should be viewed and valued in a company, helping to promote or directly harm its evolution as a process through corporate maturity in user experience.

In order to build products and services that meet the user's needs it's important to create efforts such as discipline and knowledge. In this sense companies have a fundamental role and the company will always have the choice of being able to support the methods and techniques employed in UCD and design process or will may also hamper the entire process and its development and execution stages by producing a product or service that is not capable of being used.

Just as the company can do a lot to support and assist the design process, it can also hinder the process in its development stages by producing a product or service unsuitable for use.

3.3 Organizational and Culture Change

The challenge of implementing UCD practices in an organization is much more challenging for a professional than working on their own skills or a team project. For Goodwin (2011, p. 697) change is more challenging and more personally frustrating than any other sort of design problem but designers have a very important differential to handle organizational changes: human-centered problem solving. This special ability allows maintain persisting and being focus on practices and approaches that will bring results by medium or long term. Goodwin explains that to integrate a project with design practices and processes can take at least three years, and that's in a small company with strong executive commitment. In larger companies where leaders have mixed feelings, efforts to integrate design could take five or ten years and may still fail along the way if there is insufficient commitment or vigilance regarding cultural change.

While there are companies that do not understand the potential of UCD practices, others state that they consider users, but they don't necessarily get well succeed. While there are professionals struggling for change in the company's vision, it is still need to consider how practices are developed. According to Lund (2011, pp. 227–228), although organizations claim that they are focused on their customers, they make the mistake of providing what they want rather than what they really need. Design based on an understanding of the scenario becomes much more useful and is based on a deep understanding of the user. For this to happen it is necessary to value design and, above all, to change the process and organizational culture.

To have an organizational change it is necessary to:

- Understand the business.
- Understand the organizational culture.
- Understand the company's stakeholders.

Understanding is the basic action. Through these factors, it will be possible to integrate the objectives into the user experience planning. Each organization has its own formal processes (and informal processes) and the user experience needs to incorporate its activities into those processes.

In general, companies must have structure to manage their processes and offer products and services to users. But to meet the demand, companies will depend strongly on the organizational culture to sustain strategic objectives. Organizational culture should encourage employees to add value an all levels and help them not to fear change. If a company is not prepared for change, it will become soon obsolete and closed to innovation besides it will fall behind in market competition.

4 Corporate UCD Maturity

4.1 The Need for Maturity

Associated to the influences of an organizational culture in UCD practices, is the corporate maturity that indicates the maturity level of the company in relation to processes of user-centered design.

As a definition maturity is an orderly scale, which combines practices of organizational attitude, technology and management comprising steps towards a totally human-centered approach (Earthy 1998).

The concern to be "mature" in relation to the user experience stems from a market need for competitive and quality products and services. For Van Tyne (2009), user experience encompasses all aspects of end-user interaction with an organization, its services, and its products. Good user experience is pleasing to customers - it increases adoption, retention, loyalty and, ultimately, revenue. Poor user experience harms customers, drives them to competition, and eventually is no longer a viable source. De Bruin et al. (2005) objectively clarifies that maturity models have been developed to help organizations identify ways to reduce costs, improve quality and reduce process time, thereby providing competitive advantages.

According to Marx et al. (2012), companies evolve on the organizational, strategic and technical aspects. These aspects analyzed together provide:

- Assessment of the current situation.
- Determination of future scenarios.
- Verification of possible and advanced ways that will guide evolution.

The recognition of corporate maturity in relation to the design process is the first step towards an evolution and satisfactory application of the process. Corporate maturity is directly related to how a company treats user-centered design processes.

Design processes in turn define what phases and activities are most appropriate for a project. Each company has a specific maturity level according to the way it understands, develops and applies design methods and techniques. In case there are difficulties or a partial understanding of the design methodologies then it will be an indication that the company is not mature enough to handle all aspects that will involve the future user experience, being less likely to create and develop products and services that meet these needs.

4.2 Steps for Maturity

Understanding the level of maturity of the organization is a first step towards improving the products and services offered to users, allowing the organization to advance to better communicate the user-centered design processes across all areas of the corporation, thus ensuring a more strategic of the corporate user experience.

For Best (2009, p. 78), the strategic direction of the organization, the way it moves from where it is to a desired future position is established through three key questions:

- Where are we now?
- Where do we want to go?
- How did we get there?

The essence of strategy is choosing the best path to follow. Through maturity models it is possible to identify at what level a company is and suggest instructions to reach the next level. For Van Tyne (2009), measuring maturity as well as being able to provide a reference for an organization also provides relative comparison to other organizations. Providing references to a company is important in a way to remain competitive in a growing market and may yield advantage if the company known how to use maturity in your favor.

According to Goodwin (2011, p. 698), the integration of design into a company depends strongly on where it stands today. Most companies start out in a less challenging and comfortable position. In an organization where most employees see the design as cosmetic or something that gets embedded in the end, has a steep (but not impossible) hill to climb toward the maturity of the process.

Once the maturity level is recognized through measurement, companies will have the challenge of transcending the current stage to reach the next stages, thus promoting continuous improvement in the user-centered design process and enhancing the quality of products and services. Earthy (1998), states that the most significant aspect for transition between levels is the culture. Moving from the one level to the next level is a major cultural change.

As a reflection of an increasingly competitive market, companies are gradually maturing the way they see the user. This may be the reflection of a shift in focus from technology to a transformation in business culture. Thompson et al. (2017) points out that organization struggle to create a structure that can consistently deliver good experience. As companies continue to mature, we will see a future change for the entire development process and the market.

5 Discussion

While aware of how much the user-centered design practices and processes can bring as a benefits to the organizations, professionals seem to be waging a great battle against a giant: on the one hand we have a company with its established culture for some time in initial maturity level being resistant to changes, requiring justifications for adopting new processes and methodologies. On the other hand there are professionals who wish to establish a change for an evolution of both the company and the products and services that are developed. To win this battle, many authors bet in the work together with teams, sensitizing managers and integrating with processes from other areas, but even these steps can become an arduous and time-consuming task for a single cell within the a business structure.

Besides a change of culture, which as presented by the authors, is not a simple task to be stimulated, there is the maturity barrier that is nothing more than a reflection, the personified form of the values and beliefs of the company.

In the case of companies with intermediate and higher maturity levels, the problems may not be less complex. At every advanced level, charges will become ever larger to prove the importance of the process in the company. And the culture will not change from day to night, but can become more flexible with the positive results, increasing the awareness of the managers.

There needs to be space for change and not just for technological trends that accompany the market. And changes cause fear by nature because they mean changing what is already known and established. In fact, the ideal would be to change the word “change” to “renewal”. Although similar, renewal is more transparent to the real sense of innovation.

In order to remain competitive in the market, companies should be more conscious and open to the desire for renewal and should not impose hieratic barriers that hinder the expression of professionals who develop and take direct responsibility for the products and services that are launched in the marketplace. In fact, companies should spontaneously listen to these professionals who want to plant the seed of innovation, increasing as a consequence the sales capacity and quality of the products allied to the company’s strategy. This does not mean blindly attending to all the desires of the team, but at least hearing and understanding what is wanted as a final result together with the satisfaction of corporate needs as well as end-users who would participate in the user-centered design process naturally.

6 Conclusion

The organizational culture has the role of transmitting the rules and values to be followed and acts as a scenario of corporate maturity which in turn indicates the understanding and the way in which the UCD design process and practices are treated within the company. If this culture recognizes and supports this process, it will increase relatively the chances of developing and launching products that reach the users’ needs, reflecting the sales result.

Many companies believe they are treading the best way to designing products and services based on users' requirements through design practices. But according to Lund (2011, pp. 227–228), while organizations claim they are focused on their clients, they end up making the mistake of providing what they want rather than what they really need. In this case, when the goal of satisfying the user is not achieved, the mistake may be being made by the business culture itself. Add to that the question of maturity in user experience that reflects how the process being employed and also how the company has addressed the strategy to achieve its goals.

Promoting change in organizational culture is also a great mission for design that has a culture very different from the business, provoking a "culture shock". But for this you need to learn to value yourself to change and update the corporate processes and thus ultimately generate changes in the culture to integrate the design processes to the organization.

Companies should be more attentive to use design practices more seriously, valuing their methodologies and processes as a way to guarantee the final quality of a user-centered product or service. In order for this objective to be met, commitment to user satisfaction must be a goal to be achieved not only by the design team, but by the company as a whole.

Acknowledgment. First of all, I would like to thank Denise Pilar for the invitation and opportunity to participate in this important event for the HCI area. I also would like to thank CNPQ for financing the scholarship and the teachers, master and doctoral students of PPG Design - PUC-Rio who have always supported my research.

References

Best, K.: The Fundamentals of Design Management. AVA Academia Publishing, Lausanne (2009)

Courage, C., Baxter, K.: Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques. Gulf Professional Publishing, San Francisco (2005)

De Bruin, T., et al.: Understanding the main phases of developing a maturity assessment model (2005)

Earthy, J.: Usability maturity model: human centredness scale. INUSE Deliverable D5. 1.4 s (1998)

Goodwin, K.: Designing for the Digital Age: How to Create Human-Centered Products and Services. Wiley, Hoboken (2011)

Hartson, R., Pyla, P.: The UX Book: Process and Guidelines for Ensuring a Quality User Experience. Elsevier, Boston (2012)

Lowdermilk, T.: User-Centered Design: A Developer's Guide to Building User-Friendly Applications. O'Reilly Media Inc., Sebastopol (2013)

Lund, A.: User Experience Management: Essential Skills for Leading Effective UX Teams. Elsevier, Burlington (2011)

Marx, F., Wortmann, F., Mayer, J.: A maturity model for management control systems. Bus. Inf. Syst. Eng. **4**(4), 193–207 (2012)

Mozota, B.: Structuring strategic design management. Des. Manag. J. **9**(2), 26–31 (1998). Spring

Mozota, B.: The four powers of design: a value model in design management. Des. Manag. Rev. **17**(2), 44–53 (2006)

Schaffer, E.: Institutionalization of Usability: A Step-By-Step Guide. Addison-Wesley Professional, Upper Saddle River (2004)

Schein, H.: The Corporate Culture Survival Guide, vol. 158. Wiley, Hoboken (2009)

Thompson, E., Ray, A., Callinan, L., Ekholm, J., Poitevin, H., Reva, M.: Future of Experience: a gartner theme insight report. Gartner research: strategic analysis report (2017)

Van Tyne, S.: Corporate user-experience maturity model. In: Kurosu, M. (ed.) HCD 2009. LNCS, vol. 5619, pp. 635–639. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02806-9_74

Check for updates

A Landscape of Design: Interaction, Interpretation and the Development of Experimental Expressive Interfaces

Alan Chamberlain[1](✉), Mads Bødker[2], David De Roure[3], Pip Willcox[3], Iain Emsley[3], and Alessio Malizia[4]

[1] Department of Computer Science, University of Nottingham, Nottingham NG7 2RD, UK
alan.chamberlain@nottingham.ac.uk
[2] Copenhagen Business School, Solbjerg Plads 3, 2000 Frederiksberg, Copenhagen, Denmark
[3] Oxford e-Research Centre, Oxford University, 7 Keble Road, Oxford OX1 3QG, UK
[4] University of Hertfordshire, Hatfield Hertfordshire AL10 9AB, UK

Abstract. This short paper presents the initial research insights of an ongoing research project that focuses upon understanding the role of landscape, its use as a resource for designing interfaces for musical expression, and as a tool for leveraging ethnographic understandings about space, place, design and musical expression. We briefly discuss the emerging research and reasoning behind our approach, the site that we are focusing on, our participatory methodology and conceptual designs. This innovative research is envisaged as something that can engage and interest the conference participants, encourage debate and act as an exploratory platform, which will in turn inform our research, practice and design.

Keywords: Design · HCI · Interaction · Place · Ethnography · Musical instrument Audio · Interface · Physicality · Materiality · Experimental · Expression Musical

1 Introduction

In this short paper we discuss an ongoing research project that focuses upon the use of landscape as an artifact, which supports the design of interfaces for musical expression. We take a somewhat experimental and unorthodox approach, bringing together ethnography (sensory [16] and autoethnography [4]), participatory design, and "research through design". Indeed, a key motivation of the work is to understand the role that designing of such systems can have in terms of ethnographic practice, as we shall later explain.

The research is not only focused in the resulting physical artifacts that emerge as part of the project, but also the processes and participation that enable and support creativity and innovation, feeding into design practice. As designs, theories and ideas come forward that they will not necessarily form any sort of logical iterative design process from an external viewpoint, and this is why it is important to unpack and describe some of the issues that emerge throughout the design process. Key to the success of the project

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 24–34, 2018.
https://doi.org/10.1007/978-3-319-91238-7_3

will be our ability to offer a way for people to express themselves musically, in a way that enables them to define, understand and control.

As Pinch and Trocco [15] write, "…it is not possible to design a musical instrument by beginning with an objective set of performance specifications. Rather, a musical Instrument design usually begins with a designer's intuition… Where does that intuition come from?"

Our focus on the landscape as a catalyst for developing an interface from musical expression has emerged from a series of our earlier publications across domains and was influenced by Ingold's work [9] on landscape. We wanted to imagine expressive systems that could engage a performer directly with the landscape whilst performing. As Ingold suggests, landscapes are 'practice-scapes' rather than 'vistas' or flat visual depictions. This, he argues, is "…an alternative mode of understanding landscapes, based on the premise of our engagement with the world, rather than our detachment from it" [9]. We shall come back to this notion of 'engagement' later. In many ways engaging with our landscape leads us to reflect upon ourselves in and as part of the ongoing construction that is, us *being* part of the landscape. "…this life process is also the process of formation of the landscapes in which people have lived" [9]. Indeed, in etymological terms, the suffix "–scape" is derived from the Old Norse 'scyppan' – to create, to shape. Tracing the etymological root of the term also suggests our interpretation of what a soundscape can be: a creative engagement, a shaping, an engagement with the world, or a trace of other people's engagement with a their culturally embedded worlds of sound. Importantly, both –scapes do something to each other: being in the landscape does something to the soundscape and the soundscape does something to the landscape. Using the landscape is a way of shaping musical expression and enabling people to use their experience and understanding as a way to mediate and perform.

One of our aims is to draw on the way that we engage with the landscape, mapping these into forms that enable and create mechanisms and opportunities for expressive musical interaction. This may sound like a radical proposition, but in many respects we are surfacing the mundane [17] aspects of our lives in different ways. This is expressed through landscape-based musical artifacts that relate back to lived experience, which like any other 'documented' expression of experience is an abstraction. Our approach understands this, making evident the understanding that the expressive nature of the instrument is musical, itself in turn an expression, an abstract manifestation. However, this is not to say that it is not possible to invoke a shared experience or feeling through music, and in turn we would hope to fold such expressions into a more sensory-ethnography [16] inspired understanding of musical instrument/interface design. Later in the paper we provide some conceptual designs that form part of our initial ideation phase.

1.1 Ethnography: A Note

In order to appreciate the methods and tools that we have used to study this we first offer a brief insight into the world of ethnography (within the confines of this paper), before moving into the context of the site and the design of the system.

In this particular instance we focus upon the use of design ethnography, that is when ethnographic methods are used as part of the design process, with a special focus on the experiences of people in relation to the location. In order to explicate our approach, we use Goffman [7] as a reference point who, when discussing fieldwork in his book Asylums, states that, "any group of persons-prisoners, primitives, pilots or patients - develop a life of their own that becomes meaningful, reasonable and normal once you get close to it … a good way to learn about any of these worlds is to submit oneself in the company of the members to the daily round of petty contingencies to which they are subject." With this in mind we must turn to ethnographies of the site specific, self and the sensory as others [16, 18] have. "Sounds… are not merely abstract acoustic events" [19] their interpretation and use is social, and expressive.

As LaBelle [10] quite aptly argues "sound is intrinsically and unignorably relational: it emanates, propagates, communicates, vibrates and agitates; it leaves a body and enters others; it binds and unhinges, harmonizes and traumatizes; it sends the body moving, the mind dreaming, the air oscillating. It seemingly eludes definition, while having profound effect". This move to the expressive 'instrumental' nature of landscapes to support ethnographic practice is where our interdisciplinary research resides, as a way to understand the more humanistic influences on musical instrument design (intuition) [5].

2 The Site of Expression

Choosing a site and community to work with is difficult, and carrying out research "in the wild" is resource intensive, with the real world often impinging upon the research. Our approach, based on previous research projects is to use an embedded researcher, who is part of the community, understands the site and has experience of using and designing audio performance-based systems. The community that we are currently working with is based in West Wales (UK) and has a hill fort in the area that they live. More specifically there is a history project based around the hill fort and we have been able to attend sessions and take part in the activities relating to the site.

There are also non-governmental organisations that maintain the site and other organisations with an interest in it: as a tourist location, a historical site, and a managed wildlife site. There are data available relating to the site that might be used as a mechanism to impact upon the interface. This multiplicity of stakeholders and data provides a rich set of resources that can be pulled into the design. It is key that we understand some of the issues and data that are created and used by such groups as this data may be called on as a performative tool by which to express one's self.

The images (Fig. 1) show an aerial photograph of the site and an image of a 3D model created from LiDAR data, which is openly available. Having access to the LiDAR data means that models can be quickly made, printed as physical 3D models, or virtual models imported for example into Max/MSP. The LiDAR data can also be used to alter the pitch, volume, and chosen sound and could also be used as a triggering mechanism. At this stage of the project we are using the shapes contained within the site as a way through which we can start the initial ideation or design phases of the project, enabling us to ask people to engage in the design process and develop early prototypes.

Fig. 1. The two images above show an aerial photograph (left) and a virtual 3D model created with LiDAR data.

It can be difficult for community groups to engage with concepts such as sound control, gesture and in this case the landscape as something that can form an interface for physical expression. Our work enables people to understand and engage in concepts, while we acknowledge that we may be guiding the design by providing a framework and some sort of motivation for their explorations. This makes it vital that an experienced researcher is present. In our previous studies taking a probe, or provotype [1] has been key to enabling users to engage with the technology in order for researchers and designers to understand both the actual and potential use of the instruments and software being designed [11] and the role that such groups can play in supporting and developing tools to support existing communities of practice [12]. In many respects this move from the societal, to the institutional and to the self is mirrored in the way that Human-Computer Interaction and Computer Supported Cooperative Work has moved from the lab, to understanding industrial organizationally embedded systems, out into the real world and ultimately into the realm of the maker and a world where one can design and develop systems for oneself.

A recent workshop [2] found that using tangible artifacts as a way to both use and interact with non-tangible media meant that we were able to collect sounds, to explore their nature in respect to their interaction with other people and machines, and to study their integration with a physical format. These findings led us towards issues relating to physicality, performance and the expressive nature of audio and landscape in a physical format. We used Percussa AudioCubes[1] as a way to enable people to bring sounds together easily and to explore contrast and similarity of sounds, pleasurable and unpleasant sounds, affectively engaging and uninteresting sounds, sounds that evoked memories, sounds that were unrecognizable, and many more. We see this as a move to the expressive 'instrumental' possibilities of landscape to support ethnographic practice in regard to developing an understanding of such sensory resources for expressive systems design. The purpose of fusing ethnography and musical expression is to explore the intersection of the arts, sensory ethnography [16], phenomenological aspects of place experience (Genius Loci, the spirit of place), and the discipline of human-computer

[1] *See* - https://www.percussa.com/what-are-audiocubes/.

interaction as a means of conceptualizing, building, and evaluating a new musical interface. What might be gained from rendering knowledge about culture in a more generative form, through the collaborative and creative process of 'playing a landscape' as an instrument?

From our perspective our position is further expounded by the performative nature of Goffman's ethnographic approaches, that in many ways leads us to reflect upon ourselves in and as part of the ongoing construction that is us 'doing' being (to borrow from ethnographic terminology) part of the landscape. Taking part in, being and expressing ourselves in a dramaturgical way.

The discipline of anthropology also maintains a long-standing engagement with sound and how acoustic cultures or worlds are performed [6]. Sound matters in the composition of both cultures and bodies. We do not propose that sound should be sequestered from an overall multi-sensory perspective on what it means to experience. Yet emphasizing our concern with sounds and acoustic environments can be a useful counterpoint to the overwhelmingly visual-centric epistemologies of western science.

3 Sonification

Sonification, the "transformation of data relations into perceived relations in an acoustic signal" [29] provides an alternative way of exploring the landscape but raises challenges for design. Where soundscapes, such as Virtual St Paul's Cathedral[2], simulate a scenario using data, sonification may be used to display more abstract concepts, such as the environment, archaeology and the land.

The technique is a way of interacting landscape through seismology [25] but might be a useful way of interacting with archaeological context. Using the soil contexts, we can build a time based sonification demonstrating how the layers have developed and whether there are anomalies.

Environmental data that may be sonified includes the weather or soil conditions. The complexity of these variables poses issues with the complexity and volume of the information as this will include humidity, temperature and wind speed amongst others [30].

Networks, such as power and telephone lines, traverse the landscape; some lines visible, others not. Home appliance electrical has been sonified [26, 27] but not at the national level. The national academic network [23] is sonified in various methods to show how part of academia's fabric using white nose and identifying the sections. The sonification might show how modern and ancient landscapes co-exist in time and space as layers, waiting to either be discovered or made invisible. This might also be extended to maps of land use and boundaries to show the changes over time.

In previous work, reported performances of David Garrick, the Eighteenth-Century actor [22], were sonified. This allows us to build a model to test how assumptions about a voice might act within a space, such as a theatre where the building specifications can be represented. Using a hill fort poses challenges through the way that a voice or set of voices might work within not only the physical factors but also the environmental ones.

2 https://vpcp.chass.ncsu.edu/.

As the fortifications are open air, weather and direction changes how sounds may be heard or interpreted.

This raises the question of how to sonify the potential movements of a particular person or people and the audio calculations between the listener and the sonified object. It would suggest the sounds should reflect not only the actual space but also the historical aspects of the location.

A consideration is whether the sonification has interactive experience or not and how this may be provided. Two choices are immediately apparent: mobile phones or wearable devices. Using an application, mobiles could work to provide a sonification unique to each phone and may work in concert with others [24]. Wearable devices, such as the MozziDuino[3] could be used to allow humans to interact the environment without relying on being able to download and application. In both cases, the machinery might be programmed to understand the landscape and signals. There is also the question of leakage - is each listener hearing something for someone else and does it intrude? – or creating a social instrument.

This provokes questions about how the sonification is controlled. Work with museum visitors with visual impairments [23] concluded that non-linear, interaction with similarities to phone interaction helped discover how the system worked. It also allowed the testers to navigate the audio in their own manner, discovering the narrative that is relevant to them as well as navigating between introduction and description. It suggests that the ethnography and participatory design are useful tools to develop interaction models but that a non-linear interaction may be more meaningful. They might be more individually meaningful if interacting with both human and landscapes.

Limitations on the interactions limit the facets and, potentially the complexity, that can be shown. This also imposes the sonification onto the user, directing the attention. Conversely, too many would be too difficult to understand.

These various aspects present challenges for design in terms of the number, and complexity of each, aspect. This has a symbiotic relationship with designing the interaction and the number of layers that we may consider, both from a technical and a human perspective. This may create technical issues with designing the types of interaction, such as buttons or beacons, and the person understanding the amount of audio information as well as potentially visual information. Consideration of these factors will affect the sound design.

4 Ideation, Data and Design

Shifting towards an understanding of the soundscape as an engaged encounter (a shaping, a practice rather than a depiction or a –scope) opens up a number of interesting possibilities. The main challenge is to set the scene for a participatory and an engaged process of ideation with the relevant stakeholders. A secondary challenge is to start to think about and envision how systems such as this might inform design and the development of tools that can support the development of cyber-physical systems.

[3] https://sensorium.github.io/Mozzi/.

First, the intention is not to create a static soundscape. Rather than considering the soundscape as an artifact or a 'depiction' of an existing structure, how can land- and soundscapes become part of a musically interesting expression? How do we create the possibility of interacting with the landscape as an instrument?

Some of the indicative questions to explore include:

- How can we collaboratively explore the landscape with the purpose of finding out what is significant and relevant to the performer?
- What features of the landscape are interesting and meaningful to manipulate for an audience?
- What are the potential affective connections an engaged land/sound-scape and how can they be augmented or made interactive in interesting ways?
- Physical (direct embodied manipulation or physical movement) or 'ephemeral' (abstract data, "invisible" sensor data,) interfaces, what is the role of the body in interaction?
- Distance of interaction (what is a meaningful "shape" of an interactive landscape, when does it become meaningless or un-engaging for performer and audience?
- Discrete experience or event. Or both?
- Self-design as way of cultivating relations to other (and vice versa?)
- A musical instrument as an artifact of relations vs. an artifact of meditation or personal growth?
- How do we sense time/progression of a (temporal long-form) musical piece, how does the time/rhythm of a piece relate to local times/rhythms?
- What are sources in the landscape of fascination, mystery, hauntings, longing, loss, love, life, death, decay, activity, rest, congregation, dispersal and so on?
- What does history do to a landscape – local vs. non-local vs. global narratives about place.
- The experiencing body or the topographic abstraction as parameter in interaction.

4.1 Interfacing with the Past as Resource for Expression in the Present

Landscapes, the land that is formed by human activity, are rich with historical markers. These range from the incidental (an old path) to the monumental (a monument or any other built artifact that aims to prompt a distinct historical awareness). The land is haunted by the ghosts of personal and wider cultural memories [4]. Both personal and cultural histories disappear or are buried under layers of other histories. Some are distorted (glorified, demonized or otherwise mangled) by narratives that somehow displace their significance or dispute their truthfulness.

In an experiment under the banner of "Numbers into Notes" [3], we have installed musical devices within an interactive space. Each device generates music according to algorithms, which are parameterized by sensor readings and physical interactions within the space. This challenges the traditional boundaries of composer, performer and audience; rather, each interaction is an intervention, which influences the experience of the next audience member. It regards the landscape as an interface to the co-created composition and experience of music.

Furthermore, a historical dimension to our work bases algorithms on the mathematics of the 1800s. Essentially we are conducting an experiment, using the methodology of "experimental humanities", in which we explore the idea that mechanical music boxes from the time of Ada Lovelace, have evolved into mutually responsive digital devices today. It is the making manifest and the bringing into being that has thus far inspired compositions, performances (opera) and the design of (physical) digital music objects. It is this move from the intangible to the tangible, and the possibilities that this affords in terms of expression and experience that bear upon our design understanding.

The image above (Fig. 2) starts to map out, and physically illustrate, how the forms that exist in the landscape and archival imagery relating to the place can be pulled together and imagined as catalyst for conversation and ideation. Using tools such as these, which bring together archives, ideas, a range of technique and personal reflections, all from disparate sources both allow people to imagine what can be and envisage what is possible. In the next section we discuss some of the technical possibilities: although not the key motivation of this research, it is important as a feature of our endeavors. It also illustrates the multitude of technical possibilities that are available to contemporary developers and designers.

Fig. 2. Features of the hill-fort outlined as control zones, historical images added to support ideation.

4.2 Future Forms and Plans

As we have seen, we are currently engaging with a virtual 3D model of the site: we have images, maps, drawings and recordings. In our next phase of development we will further explore the landscape with people in order to deep map [Pearson Arch] the space. Our research through design [20] builds upon existing work, by taking a fully multidisciplinary participatory approach. Understanding the many layers of meaning when designing such systems plays an important part in the player being able to express themselves and the audience being able to understand the nature of the expression. Our discussions based in what is technically possible have ranged far and wide, opening up a range of opportunities and possible challenges, when related back to the landscape.

For example using sensors that might feed back to the interface, perhaps based on the weather, as Ingold [8] maintains, the weather connects the sky and the earth, and is a significant part of the identity of a landscape and the people who live in it. However being able to manage an ever-changing interface could be difficult, equally an interface that was based on such weather data could equally prove less than expressive, it the weather was calm. How might we design for change, for transience, for things that sometimes are a key part of the landscape and at other times disappear? In terms of building physical interfaces this is a problem, but easier to accomplish in software-based systems. Equally thermal interfaces could be used to convey the seasonality, which could mean that a physical interface could be hot, or cold to the touch, having a dramatic impact upon the way that player/s would perform. Hardware and software could be combined: as we have already noted, it is possible to pull in the virtual 3D model into Max/MSP and experiment with implementing designs for musical expression; it is also possible to 3D print the landscape in different materials and add textures or images, where the images need to be permanent, but could be mapped over to another controller or existing system in the first place in order to experiment with different layers and the player's responses to different layers, techniques and sounds. The layers could consist of anything from wildlife data, to personal reflections and even geophysical data that could enhance different ways of thinking about what the performer could express. Our research has also led us to start to look at other platforms such as opera [30] as a mechanism to use and stage interpretive pieces and the use of intelligent agents [31] as a way to support and assist people as part of the process of musical creation

Currently we have a large range of possibilities ranging from using ferrofluids for emergent interfaces, to software systems and physical interfaces and there have even been some discussions about the use of Hertzian space *see*[4]. As such even discussing and imagining what it might be possible to realize is creating a rich resource for design.

5 Conclusion

Using the landscape as an artifact that can be used, reflected upon and used to imbue and develop new forms of expressive musical interaction is both interesting and challenging. It is clear that understandings of landscape are multi-faceted and as such we hope that this short paper has started to raise some important issues that relate to this truly multidisciplinary space, that needs to be understood in terms of developing social, physical and technical understandings of what can be accomplished. We discuss the initial stages of the project that brings together understandings of landscape and people in order to further the ways in which expressive interfaces can be designed and developed. Even in these initial stages of the project important issues and possibilities are highlighted that we propose other researchers and designers will find of interest. As such we would hope that this short paper serves as a platform for reflection and discussion.

Acknowledgments. This research was supported through the following EPSRC projects: Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1) and Living with Digital Ubiquity (EP/M000877/1).

[4] *See* the LOM Elektrosluch (https://lom.audio/product/elektrosluch-3-plus/).

References

1. Boer, L., Donovan, J.: Provotypes for participatory innovation. In: Proceedings of the 8th ACM Conference on Designing Interactive Systems. Association for Computing Machinery. ACM Press (2012)
2. Chamberlain, A., Bødker, M., Hazzard, A., Benford, S.: Audio in place: media, mobility and HCI - creating meaning in space. In: Proceedings of Mobile HCI 2016. ACM Press (2016)
3. De Roure, D., Willcox, P., Chamberlain, A.: Experimental digital humanities: creative interventions in algorithmic composition on a hypothetical mechanical computer. In: DMRN +11: Digital Music Research Network, London, December 2016 (2016)
4. Ellis, C., Bochner, A.P.: .ethnography, personal narrative, reflexivity: researcher as subject. In: Denzin, N.K., Lincoln, Y.S. (eds.) Handbook of Qualitative Research, pp. 733–68, 2nd edn. Sage Publications, Thousand Oaks (2000)
5. Eslambolchilar, P., Bødker, M., Chamberlain, A.: Ways of walking: understanding walking's implications for the design of handheld technology via a humanistic ethnographic approach. Hum. Technol. **12**(1), 5–30 (2016). An Interdisciplinary Journal on Humans in ICT Environments - Human-Technology Choreographies: Body, Movement and Space – Special Issue
6. Feld, S.: Wept thoughts: the voicing of kaluli memories. Oral Tradit. **5**(2–3), 241–266 (1990)
7. Goffman, E.: Asylums. Penguin Books, London (1994)
8. Ingold, T.: The temporality of the landscape. In: World Archaeology, vol. 25, no. 2 (1993)
9. Ingold, T.: The Perception of the Environment: Essays in Livelihood, Dwelling and Skill. Routledge, London and New York (2000)
10. LaBelle, B.: Background Noise: Perspectives on Sound Art. Continuum Books, New York (2006)
11. McPherson, A., Chamberlain, A., Hazzard, A., McGrath, S., Benford, S.: Designing for exploratory play with a hackable digital musical instrument. In: Proceedings of Designing Interactive Systems, DIS 2016, Brisbane, Australia, 4–8 June 2016. ACM Press (2016)
12. Morreale, F., Moro, G., Chamberlain, A., Benford, S., McPherson, A.: Building a maker community around an open hardware platform. In: Proceedings of CHI 2017, Denver, USA (2017)
13. Pearson, M., Shanks, M.: Theatre/Archaeology. Routledge, New York (2001)
14. Pearson, M.: In Comes I: Performance, Memory and Landscape. University of Exeter Press, Exeter (2006)
15. Pinch, T., Trocco, F.: Analog Days: The Invention and Impact of the Moog Synthesizer. Harvard University Press, Cambridge (2002)
16. Pink, S.: Doing Sensory Ethnography. Sage, London (2015)
17. Pollner, M.: Mundane Reason: Reality in Everyday and Sociological Discourse. Cambridge University Press, Cambridge (2010)
18. Shafer, R.M.: The Soundscape: Our Sonic Environment and the Tuning of the World. Destiny Books, Vermont (1994)
19. Rodaway, P.: Sensuous Geographies. Taylor and Francis, Oxford (2002)
20. Zimmerman, J., Forlizzi, J., Evenson, S.: Research through design as a method for interaction design research in HCI. In: Proceedings of CHI 2007 SIGCHI Conference on Human Factors in Computing system. ACM (2007)
21. Emsley, I., Chamberlain, A., De Roure, D.: Hearing the Humanities: Sonifying Steele's Shakespeare, Centre for Digital Music, Queen Mary University London, UK (2017). https://doi.org/10.26494/DMRN.2017.30583

22. Emsley, I., Graven, T., Bird, N., Griffiths, S., Suess, J.: Please touch the art: experiences in developing for people who are visually impaired in museums. In: Research Software Engineer's Conference, Manchester, UK (2017). http://rse.ac.uk/conf2017/talk-abstracts/#please-touch-the-art–experiences-in-developing-for-people-who-are-visually-impaired-in-museums
23. Emsley, I., De Roure, D., Chamberlain, A.: A network of noise: designing with a decade of data to sonify JANET. In: Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences (AM 2017). ACM, New York (2017). https://doi.org/10.1145/3123514.3123567
24. Schiemer, G.: Satellite Gamelan: microtonal sonification using a large consort of mobile phones. In: 22nd International Community on Auditory Display (ICAD 2016), Sydney, Australia. International Community on Auditory Display (2016)
25. Boschi, L., Paté, A., Holtzman, B., Le Carrou, J.L.: Can auditory display help us categorize seismic signals? In: 21st International Community on Auditory Display, Graz, Austria. International Community on Auditory Display (2015)
26. Lockton, D., Bowden, F., Brass, C., Gheerawo, R.: Powerchord: towards ambient appliance-level electricity use feedback through real-time sonification. In: Hervás, R., Lee, S., Nugent, C., Bravo, J. (eds.) UCAmI 2014. LNCS, vol. 8867, pp. 48–51. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13102-3_10
27. Lockton, D., Bowden, F., Brass, C., Gheerawo, R.: Bird-Watching: exploring sonification of home electricity use with birdsong. In: Sonification of Health and Environmental Data (SoniHED), York, UK. Sonification of Health and Environmental Data (2014)
28. Schuett, J.H., Winton, R.J., Batterman, J.M., Walker, B.N.: Auditory weather reports: demonstrating listener comprehension of five concurrent variables. In: Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound (AM 2014). ACM, New York (2014). http://dx.doi.org/10.1145/2636879.263689
29. Kramer, G., Walker, B.N., Bonebright, T., Cook, P., Flowers, J., Miner, N., Neuhoff, J., Bargar, R., Barrass, S., Berger, J., Evreinov, G.: The sonification report: status of the field and research agenda. Report prepared for the National Science Foundation by members of the International Community for Auditory Display. International Community for Auditory Display (ICAD), Santa Fe, NM (1999)
30. Chamberlain, A., Kallionpää, M., Benford, S.: The art and 'science' of opera: composing, staging & designing new forms of interactive theatrical performance. In: DMRN+12, Digital Music Research Network, London, UK, 19th December (2017)
31. Chamberlain, A., Malizia, A., De Roure, D.: An agent on my shoulder: AI, privacy and the application of human-like computing technologies to music creation. In: DMRN+12, Digital Music Research Network, London, UK, 19th December (2017)

InterArt: Learning Human-Computer Interaction Through the Making of Interactive Art

Emanuel Felipe Duarte(✉) and M. Cecília C. Baranauskas(✉)

Institute of Computing, University of Campinas,
Av. Albert Einstein, 1251, Campinas, Brazil
{emanuel.duarte,cecilia}@ic.unicamp.br

Abstract. Technological advances are substantially changing what we understand as a computer. For Human-Computer Interaction (HCI) educators, it is a challenge to stay updated and prepare students to cope with the state of the art technology, interaction possibilities, and ever-growing sociotechnical challenges. We advocate that, besides traditional HCI topics, students should be encouraged to creatively engage and tinker with novel tools and technologies and design for open-ended scenarios. In this paper, we report on results of the InterArt project: HCI students explored tools and technologies from the maker culture to design and implement an interactive artwork. We detail our teaching methodology and the conducted activities; the products of these activities and feedback from students compose our results. Our approach made possible for students to creatively express themselves, be a part of a participatory design and collective sensemaking process, discover and practice socially aware design, and ultimately expand what they understand as HCI.

Keywords: HCI education · Interactive art · Evaluation
Socially aware design · Maker culture

1 Introduction

Recent technological advances such as personal mobile devices, artificial intelligence, and Internet of Things (IoT) devices are substantially changing what we understand as a computer, as well as how we interact with it. Nevertheless, each future technological advance has the potential to bring even more dramatic changes. When it comes to Human-Computer Interaction (HCI), technological advances also bring to researchers and practitioners alike the inevitable challenge of staying updated. Inevitably, the field of HCI has no option but to evolve in response to changes in the technological landscape, infrastructure and the expanding capacities and contexts of technology use [6,11,12].

According to True, Peeters and Fallman [27], education has an important role in molding and shaping people who will practice design, and designers, in turn, have the ability to shape the world. For HCI education in specific, it is a

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 35–54, 2018.
https://doi.org/10.1007/978-3-319-91238-7_4

great challenge for an educator to constantly stay updated and properly prepare students to cope with the state of the art technology, a vast amount of interaction possibilities, and an ever-growing collection of sociotechnical challenges. Furthermore, it is an even greater challenge to prepare a future generation of HCI researchers and practitioners for possibilities and implications that yet do not exist, cannot be precisely predicted, but will certainly be made available at some point in the future. These educational challenges point us towards our general research question in this paper:

General research question: Can we prepare HCI students not only for current design problems but also for those that will be faced in the next 5, 10 or 20 years from now?

We believe that to better prepare HCI students to cope with current technologies, while also trying to prepare them for forthcoming technological advances, (1) students should be encouraged to creatively engage and tinker with novel tools and technologies, such as Arduino and similar electronics gadgets, artificial intelligence APIs, and IoT devices. Furthermore, following the well-accepted notion that HCI research and practice is no longer limited only to well-defined and/or workplace problems [3,4,13,25], we also believe that (2) students should be encouraged to design solutions for elusive, open-ended scenarios. Building upon (1) and (2), we devise our specific research question:

Specific research question: Can we teach HCI by employing novel technologies, such as Arduino and other devices, in the design of open-ended scenarios, such as the creation of interactive art?

In this paper, we report on how we promoted an articulation between art and science in a HCI education context. Besides studying and practicing what can be considered traditional topics within the field, undergraduate students also explored tools and technologies from the maker culture to approach an open-ended design problem presented to them: to design, implement and evaluate an interactive art project. Our primary objective was to engage Computer Science and Computer Engineering students undertaking a HCI course with challenging (but open-ended) concepts and new technologies at the intersection of art and science. We believe that this approach may lead respectively to more open-minded HCI practitioners and higher awareness to novel forms of interaction.

The paper is structured as follows: in Sect. 2 we present our theoretical and contextual background; in Sect. 3 we present the InterArt project and related activities; in Sect. 4 we report our results from data collected throughout the course, including descriptions of the created interactive artworks; in Sect. 5 we discuss our results and main findings; and finally, in Sect. 6 we present conclusions by revisiting our research questions.

2 Background

In the literature, there seems to be no consensus on what topics precisely constitutes the field of HCI. In 1992, when attempting to establish a curriculum, Hewett *et al.* stated “There is currently no agreed upon definition of the range of topics which form the area of human-computer interaction” [14]. Even though the field has considerably evolved and matured in the following decades, in Churchill, Bowser and Preece’s survey [6], the authors conclude that Hewett *et al.*’s observation was still valid in 2013. We understand this lack of consensus not as a flaw in HCI education, but as a sign of the diversity of thought in our field, as well as a response to the need of adapting HCI education to different contexts. Thereby, it is not our aim to communicate our approach as a universal solution, but instead as a possibility that made sense in our geographic, economic, technological and social context.

Literature also suggests a relative lack of detailed reports and formal discussion regarding practical approaches to HCI education. True, Peeters and Fallman [27] argue that there are numerous approaches to HCI education across existing programs worldwide. However, the inner workings of these programs are rarely discussed outside their own institutions. Moreover, Grandhi [11] argues that most discussions regarding HCI pedagogy are informal and done through social networks or in conferences. Thus, it is our intention to report our approach with as much detail as possible, including the rationale behind our choices.

Our study was conducted in a HCI undergraduate course during the first semester of 2017, ministered by a professor and an intern teacher (second and first authors, respectively). A total of 55 Computer Engineering and Computer Science students from the University of Campinas (UNICAMP), Brazil, participated. For group activities, students self-organized themselves at the beginning of the semester in 9 teams of 6 or 7 members. Regarding our pedagogical approach, we adopted a sociotechnical perspective to the design of interactive systems, making use of inclusive and participatory approaches [2,18,26]. We also adopted a just-in-time teaching methodology [22], which means that students were proposed with pre-class “warm-up” activities, designed to prepare them for the topic to be discussed in class. With Preece, Sharp and Rogers [16] and Rocha and Baranauskas [24] as the main bibliography, the course program included:

1. History and evolution of the field of HCI;
2. Human factors in HCI (*e.g.*, perception, memory);
3. HCI paradigms and respective design and evaluation methods;
4. Introduction to Semiotics and Organisational Semiotics;
5. Accessibility and Universal Design concepts and methods;
6. (UI) design tools and environments; and
7. Selected subjects (*e.g.*, IoT interaction design, cultural aspects).

These topics were taught in a traditional classroom context, and students were evaluated with two tests and two hands-on projects they conducted along the semester. The first project was a more traditional approach to a HCI problem: students were asked to redesign, in the form of a mobile app, their university’s current web app for managing classes, grades, and other academic matters. This project included classic techniques and methods, such as the creation of personas, heuristic evaluation and paper and digital prototyping.

The second hands-on project, however, is the one we will focus on this paper. The project was entitled InterArt, and students had to design and build an interactive artwork.

In our study, we choose to work with art for two main reasons. First, the articulation of art and science has been an important source of innovation and ground-breaking contributions in many fields throughout history [29]. Considering John Lasseter's quote that "[...] art challenges technology, and technology inspires art" [17], we take the perspective that there is a two-way path of influences between art and science, and an articulation can be beneficial for both sides. As a second reason, in HCI art and science may be articulated through the overlapping concept of interactive art [9,10,21]. We opted to not adhere to any precise or definitive definition of what is interactive art, as it would require equally precise definitions of what is both art and interaction, and every attempt to do so will always be subject to debate [1,15]. However, for the practical purposes of this paper, we consider interactive art to be broadly any form of art enhanced with any kind of computer-based interactivity.

The creation of interactive artworks, in turn, brings a fortunate consequence: it also represents an opportunity to explore novel technologies and interaction possibilities that could otherwise be overlooked in more traditional design problems. Concerning the importance of exploring novel technologies and materials, Löwgren [19] emphasizes the importance of a maker culture in interaction design research. According to the author, it can support the exploratory design of what he refers to as non-idiomatic interaction, a kind of interaction that is not yet inside the "established idiom", *i.e.*, not yet broadly known or understood. Posch [23], in turn, discusses how our tools, that we often take for granted, have the potential to shape our interaction with any kind of technology in a making process, and how we may appropriate our tools in new ways by reflecting and experimenting with them. Therefore, in this study, we opted to employ technologies and tools often associated with the maker culture phenomenon.

3 The Path Towards the InterArt Project

Even before formally presenting the project to the students, whenever possible we articulated the course's content with the InterArt design problem. Some of the proposed warm-ups and in-class activities conducted allowed students to discuss aspects that might intersect with art, such as perception, and to form their own definitions of interactive art. In the following subsection, we list all the activities that were somehow connected to the InterArt project.

3.1 Relevant Activities

Introductory Questionnaire. The first warm-up was a questionnaire aimed at better knowing the students and their motivations. We asked open questions about their motivations for choosing Computer Science or Computer Engineering, and things they like and dislike in it. The last question, however, was our first approach towards the subject of art: we asked them to "Indicate an artist or artwork that you admire." In the following week, students named their teams after an artist or artwork of their liking.

Perception Warm-Up. As a warm-up for a class on the topic of human perception, students were asked to openly state what they "perceive" in a picture of Lygia Pape's Divisor (1968). The class, however, was randomly divided into two groups for this assignment: the first group answered without having access to the responses of colleagues, while the second group had access to previous responses from same group colleagues before responding. The design of the warm-up, which is illustrated in Fig. 1, was intended to discuss with students that perception is not only physiological but also socially informed.

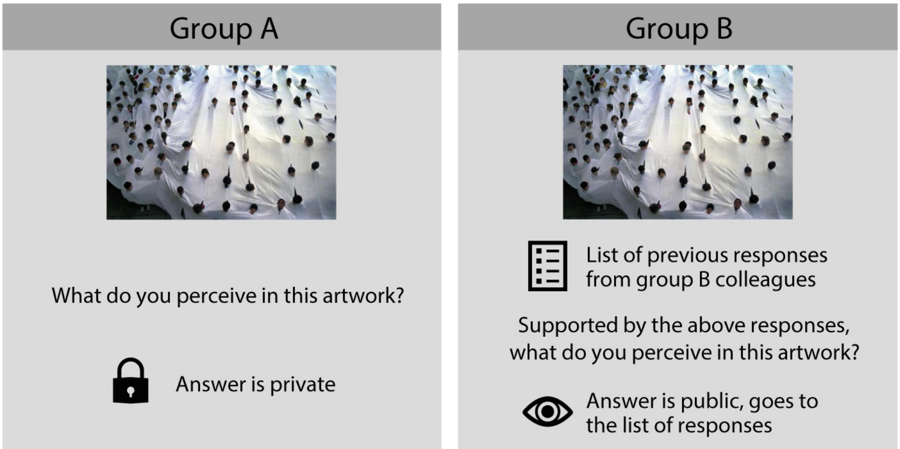


Fig. 1. Illustration of the design of the perception warm-up.

Forming a Concept of Interactive Art. After a brief introduction about the theme of the InterArt project, we tasked students with researching the subject and coming up with an initial concept of what they understood as interactive art. Each team presented their initial concept in 10-min seminars, including at least one illustrative example, and describing their research process. After the class, we asked in an online form: "Did watch your colleague's presentations somehow aggregate or modified your initial concept of interactive art? Please justify your answer." Instead of imposing some arbitrary definition, our objective was to have students form their own understanding of the subject and to expand their understanding by sharing it with each other.

Ideation Techniques for InterArt. We proposed three ideation techniques to help teams come up with ideas for an interactive artwork. The first technique, "challenging existing assumptions", was adapted from Michalko [20, pp. 43–52]. Students listed preconceived ideas related to the project and then proceeded to challenge these ideas, promoting unconventional thinking patterns. The second technique was a brainwriting session [28], in which the team selected their favorite ideas from the previous activity and collaboratively wrote a proposal on how to make the reversal of that idea come true. A third technique, "translating sensory experiences", is based on the reported influence of sensory experiences on creativity [30], and consists of listing non-visual sensory elements and trying to give them a visual representation. The third technique could not be conducted in class due to time constraints but was still presented to the students.

InterArt Requirements. As a next step towards creating their interactive artwork, students were asked to write formal requirements for their projects. To support this activity, students made use of the Semiotic Framework [18, pp. 26–35]. The Semiotic Framework can be described as a "ladder" with six "steps": (1) physical world; (2) empirics; (3) syntactic; (4) semantics; (5) pragmatics; and (6) social world. Each step reveals different levels of requirements that are necessary for any system to be made and used, from the very physical components that make up a computational system in the physical world to aspects of human relations in the social world. To help clarify the more technical levels of requirements, we had already discussed with students the inner workings of some interactive art examples, exploring topics such as emotion recognition APIs, sensors, actuators, microcontrollers, and connectivity.

After the Semiotic Ladder, students also created a diagram that we named "Communicational Map". In this diagram, they represent both physical and virtual components from the Semiotic Ladder, as well as involved human actors, and illustrate how these components communicate with each other. The diagram should represent what are the paths in which the information travels, and what kind of information travels through each path. For instance, a vibration sensor may be triggered by shaking the sensor; this information reaches a microcontroller through wires; the microcontroller uses Wi-Fi to broadcast the information, and a web page is notified of the new value for the vibration variable.

Hands-On Workshop with Electronics Kit. To support the project and encourage novel forms of interaction, we provided each team with a custom-made Arduino-compatible electronics kit (we kept in mind the growing interest, low cost and easy to learn curve of these devices [8]). The kit, properly presented with a set of slides to which we will refer to as "kit slides", was composed of:

– Controller: NodeMCU (Arduino-compatible, with built-in Wi-Fi);
– Sensors: temperature & humidity, light dependent resistor, sound, reflexive obstacle, vibration and push buttons with colorful covers;
– Actuators: assorted single-color LEDs, RGB LEDs and buzzer; and
– Other components: organizer box, breadboard, jumper wires and resistors.

The kit was accompanied by custom documentation on every component, to which we will refer to as "Examples of Circuits and Codes", and by an original illustrative tutorial on how to send information from the microcontroller to a remote HTML page and *vice versa*, called "Interactive Mona Lisa". In the example from the tutorial, a LED could be turned on or off from a web page, and a virtual representation of the *Mona Lisa* shook when the vibration sensor was shaken. Students had an entire 2-h class dedicated to exploring the kit, following the tutorial and asking any questions that might arise in the process. Furthermore, students could keep the kit in their possession until the end of the semester. After the hands-on workshop, there were two more 2-h classes dedicated to using the electronics kit, but this time for prototyping the interactive artwork the teams had been designing up to that moment.

InterArt Peer Review. After the teams finished a functional prototype, we conducted a session of peer review. Students set up their prototypes inside the classroom, and, in a circular manner, each team experimented with and evaluated the prototype of the next team in the alphabetical order. To support the evaluation, students used Costello and Edmond's Pleasure Framework [7]—we opted for this framework instead of more classical HCI evaluation methods because of the nature of the project: in our context, to analyze aspects of playful interaction could yield more meaningful and helpful results than a usability test. However, we complemented the Pleasure Framework with a part of the (SAM) [5]: for each of the thirteen categories of the Pleasure Framework, we inserted a 5-point Likert scale with the pleasure dimension from the SAM, and evaluators had to justify their answers in writing. Afterwards, the teams gathered the results of the peer review and used it to further improve their interactive art prototypes. Besides helping students to finish their projects on time and improving it for the final presentation, this activity was also designed to encourage interaction between groups.

InterArt Final Presentation and Feedback Questionnaire. At the end of the semester, each team presented their project in a 10-min, free format presentation, followed by a live demonstration of their interactive artwork. Later, after the course was over and all the grades were assigned, we sent students an online form containing some questions about their experience during the semester. Among other questions, we asked: "Considering the experiences and activities conducted in this course, did your perception of what can be art, technology and HCI change during the semester? Please justify your answer." We also asked students to individually evaluate the tools and methods employed in the project, including the electronics kit. Because the semester was already over, the filling of this questionnaire was entirely optional. Lastly, students had the option to fill it anonymously if they wanted to.

4 Results

Our main results are organized into three categories. First, we present the 9 teams and their interactive artworks; then, we present results obtained from the perception and interactive art concept warm-up activities; finally, we present responses from both the introductory questionnaire and the course feedback questionnaire, highlighting student's reported experience and acceptance of the employed methods and artifacts. Quotes from participants are numbered for reference and are a free translation from Brazilian Portuguese.

4.1 Teams and Interactive Artworks

The teams and their projects are listed in alphabetical order in Table 1.

4.2 Perception and Interactive Art Concept Warm-Ups

For the perception warm-up, there were 28 responses from group A and 24 from group B. We gathered the responses, translated them from Brazilian Portuguese to English, and conducted a brief analysis of the data by looking at the most used words by students in their responses (stop words were not considered, and we grouped words such as "children" and "child" into the singular form). As a result, Fig. 3 illustrates the top 15 most frequent words in both groups A and B. It is noticeable how the first group's responses had a greater emphasis on literal aspects of the artwork (*e.g.*, child, cloth, head), while second group's responses tended more towards finding conceptual meaning (*e.g.*, individual, people, collective). This different emphasis becomes evident when looking at some of the responses, for instance:

> *"Children in the middle of a large sheet with holes, through which they pass their heads. Most of them are looking forward, but some look at each other." (Response #12 from Group A)* (Q1)

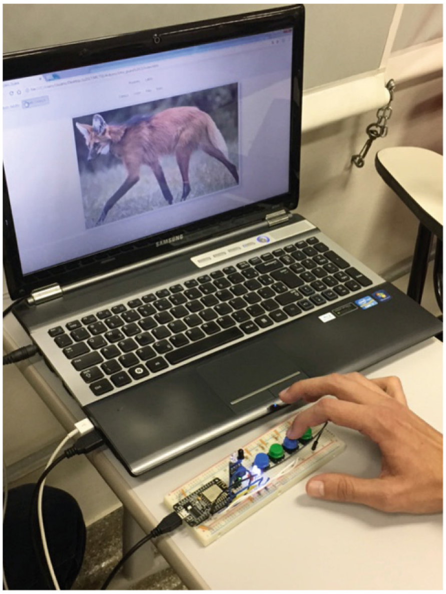
> *"The work portrays the participation of the individual as part of something greater. In this case, it is a work of art of a neo-creative artist. However, the analogy can expand to the participation of being in Society." (Response #11 from Group B)* (Q2)

For the task of forming a concept of interactive art, the teams conducted a free research on the subject and collected a wide range of what they considered interactive art examples. After the presentations, 46 students answered the follow-up question. A total of 43 students agreed that watching their colleague's presentations aggregated to or modified their initial concept of interactive art. When asked to inform what has been aggregated to or modified in their initial concept, most students reported different perspectives they had not thought about before, as can be seen in this sample of three answers:

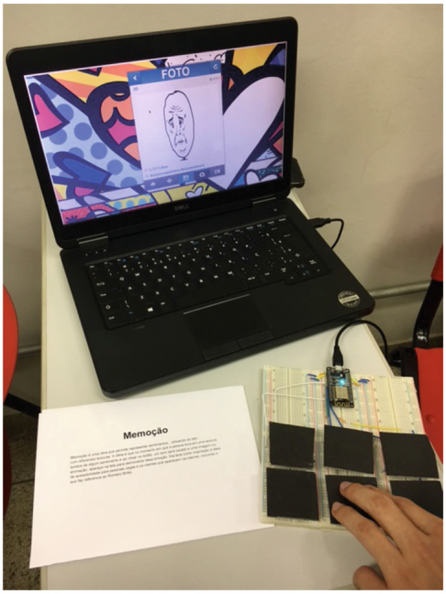
> *"Acquiring other perspectives and examples of Interactive Art. For example, taking accessibility into account in art"* (Q3)

Table 1. Teams and their respective interactive art projects.

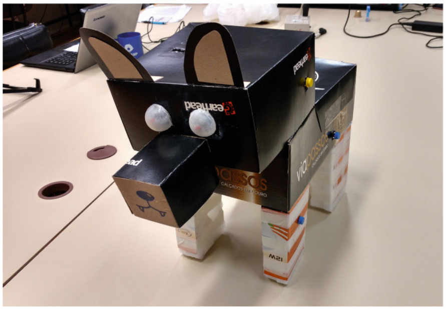
Team	Interactive Artwork
500cc (*cinquecento*), after a moment of the Renaissance, in specific the 16th century.	A sensory dancing platform with a visual drawing of the dancing. The audience can freely dance to a song on a wooden platform with speakers that vibrate according to the music. The drawing happens as the dancing is captured by a Microsoft Kinect connected to a computer. The computer then projects an abstract painting being generated by capturing the motion of anyone dancing over the platform.
Autorretrato (self-portrait), after the category of artworks in which the artists portray himself.	A dynamic display of self-portraits based on physiological measurements. There are 16 self-portraits, categorized according to brightness, color temperature, and movement. Vibration, heart rate (not included in the kit, but borrowed) and temperature sensors collect data from the audience, which is used to modify a self-portrait (*e.g.*, making it warmer and brighter), or to bring up another self-portrait more aligned with the inferred emotional state of the audience.
Gabe Newell, after the BAFTA Fellowship Award-winning game developer known for *Half-Life*.	A "non-game" exploring the concept of loneliness. Based on the *Loneliness* non-game by Jordan Magnuson, the audience uses gestures to control a virtual character that tries to approach new friends, but inevitably repels everyone, leading to a loneliness feeling. As the feeling of loneliness increases, it is accompanied by the *Dark Was the Night, Cold Was the Ground* song by Blind Willie Johnson.
Guns, after the hard rock band Guns N' Roses.	A digital musical instrument for composing songs without the need of knowing how to play musical instruments. With a curated set of rhythms from various instruments, the audience can shake the artifact and press buttons on it to change instruments and their rhythm. There is also an algorithm that automatically keeps the instruments synchronized to ensure a harmonious melody.
Kubrick, after the Academy Award-winning film director Stanley Kubrick, known for *2001: A Space Odyssey*.	A miniature monolith to interact with scenes from Kubrick's *2001: A Space Odyssey*. While a psychedelic part of the movie is projected in a loop sequence, the audience can interact with the monolith by picking it up and moving it in the air. An accelerometer and gyroscope (not included in the kit, but borrowed) capture the movement, used to control the projection accordingly (*e.g.*, speeding or slowing down the playback rate and adding a red filter proportional to the speed).
Lobisomem Atacando o Galinheiro (Werewolf Attacking the Chicken Coop), after the painting by Brazilian artist Felipe Abranches.	A farm mock-up with sensors to control an interactive storytelling involving a chicken coop being attacked by a werewolf. There is a proximity sensor in a chicken, a sound sensor in a tree, and a luminosity sensor. The sensors' activation order determines the sequence of the story being projected, which, based on the original painting, addresses the contrast between urban and rural settings and way of life.
Lobos-Guará (Maned Wolves), after the paintings *Lobo-guará I* and *Lobo-guará II* also by Felipe Abranches.	An interactive cardboard maned wolf designed for educational museums. The maned wolf artifact, which can be seen in Figure 2, has buttons in important parts (head, body, leg, and tail) that, when pressed, presents relevant information about the wolf both in text and speech. There is also a proximity sensor in its head to detect an attempt to pet him. When petted, his eyes become red and he barks, a behavior that is explained by the wolf being a wild and dangerous animal.
Nychos, after the urban/graffiti artist and illustrator known for his "dissections" of animals and other famous pop culture characters.	An interactive interpretation of famous Nychos' "dissections" of Darth Vader and Yoda from Star Wars. When the audience reaches out to a proximity sensor, the characters act as if you are trying to use "the force" against them (*e.g.*, Darth Vader shows disdain in your futile attempt), and when a vibration sensor is shaken, an animation of the dissection is played back and forth.
Romero Britto, after the Brazilian artist known for his use of vibrant colors and bold patterns.	A sensory black box, which can be seen in Figure 2, with textures inside, that evoke emotions associated with Internet memes. Inside there are six buttons covered by different textures (*e.g.*, rough, soft, gooey), and when a texture is pressed, a related meme is projected, along with a corresponding sound. For instance, pressing the gooey texture evokes a disgust meme and sound. To keep the experience non-repetitive, the meme and sound are selected randomly from a curated collection of 10 memes and 2 sounds for each emotion.



(a) Lobos-Guará's prototype.



(b) Romero Britto's prototype.



(c) Lobos-Guará's final artifact.

(d) Romero Britto's final artifact.

Fig. 2. Prototypes used for peer review and presented final versions of the artifacts from the teams Romero Britto and Lobos-Guará.

"The idea that interactive art is metaphorically a symbiosis between the artwork and the viewer because it exists with his participation." (Q4)

"I did not consider some types of art as capable of interactivity." (Q5)

4.3 Questionnaire Responses

There were 51 responses to the introductory questionnaire. We will focus our analysis on the last question, regarding an admired artist or artwork.

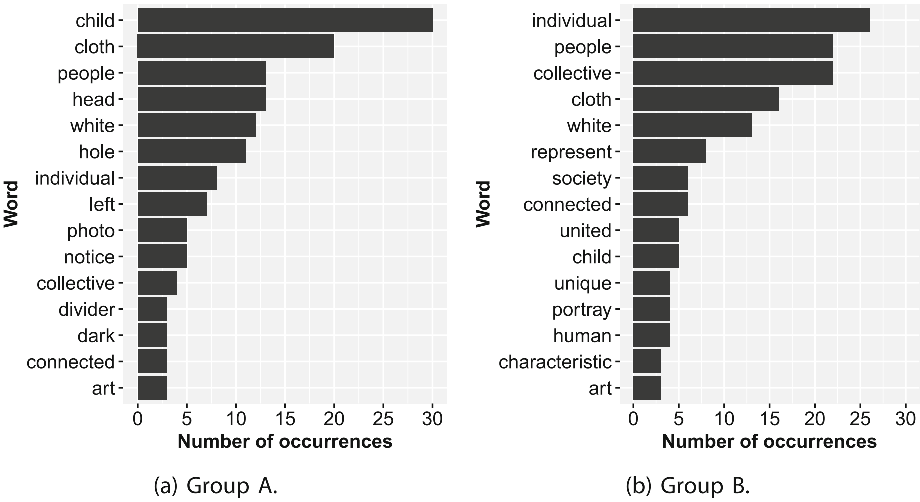


Fig. 3. Top 15 most frequent words from the perception warm-up, by group.

The answers were considerably varied among respondents. Some students pointed out famous painters, graphic artists or sculptors (*e.g.*, Vincent Van Gogh, Leonardo da Vinci, Claude Monet, Pablo Picasso, M. C. Escher, Michelangelo). Other responses contemplated music (*e.g.*, Ed Sheeran, Raul Seixas, Johann Sebastian Bach, System of a Down), literature (*e.g.*, George R. R. Margin, Carlos Drummond de Andrade) or cinema (*e.g.*, Dennis Villeneuve, Bill Murray), but also other trades, such as industrial design (Jonathan Ive) or video game development (Falco Girgis). Three responses, however, stood out and caught our attention (it is worth noting that the student from (Q8) had shown interest in digital games in a previous question):

"none" (Q6)

"I did not think of any" (Q7)

"I'm not very attached to art" (Q8)

For the course feedback questionnaire, in turn, there were only 12 responses. This drop in the number of responses can be attributed to the questionnaire being optional and being sent only after final grades were attributed. Although the responses may not be representative of the entire class, they still provide relevant feedback towards the overall experience of the course and the usefulness of methods and artifacts employed during the InterArt project. For instance, for the question "Considering the experiences and activities during the semester, did your perception of what can be Art, Technology and Human-Computer Interaction change?", 9 out of 12 respondents answered "yes". To further explore the answers, quotations (Q9) and (Q10) are two justification examples from students who responded "yes", while quotation (Q11), in contrast, is from a student who responded "no":

> *"The projects of the other teams showed me the various interpretations of art and made me reflect that there really are several kinds of interaction between art, technology and human-computer interaction that I had never stopped to think about."* (Q9)

> *"I believe that my perception of what can be considered art has improved greatly. At the beginning of the course, the students were asked their favorite artist and I answered none because at that moment I did not see that innumerable ways of expressing yourself are great examples of art."* (Q10)

> *"I think that the way the topics were covered did not help to create the design notion needed for a computer engineer."* (Q11)

Regarding the methods and artifacts used to support the InterArt project along the semester, we provided 14 statements that students had to answer to by using a 5-point Likert scale. The scale ranged from completely agree to completely disagree, and students also had to justify their answer. As can be seen in Fig. 4, there were 2 questions referring to ideation, 2 to requirements, 5 to evaluation and 5 to the electronics kit. In general, the feedback was mostly positive, and the responses are accompanied by qualitative feedback.

Feedback for Ideation Techniques. For the Challenging Existing Assumptions and BrainWriting techniques (questions A and B in Fig. 4, respectively) there were mostly positive responses. Quotation (Q12), for instance, highlights how the participatory nature of these activities facilitated engagement within the team and allowed the emergence of several different options for the project. The student from quotation (Q13), in turn, expresses how the Challenging Existing Assumptions technique had a major role for his team in determining the basic idea of the project. Lastly, quotation (Q14) contrasts with the previous two examples with the argument from the student that, in his opinion, the BrainWriting was not a suitable technique and led to a polluted design.

> *"In my opinion and personal experience, one of the biggest barriers to team brainstorming is the self-censorship that most participants do. However, in the way the ideation techniques were conducted, for example, a paper for each participant, it was much easier to express the ideas that emerged, allowing the team to generate several different options and choose the best ones to implement/execute."* (Q12)

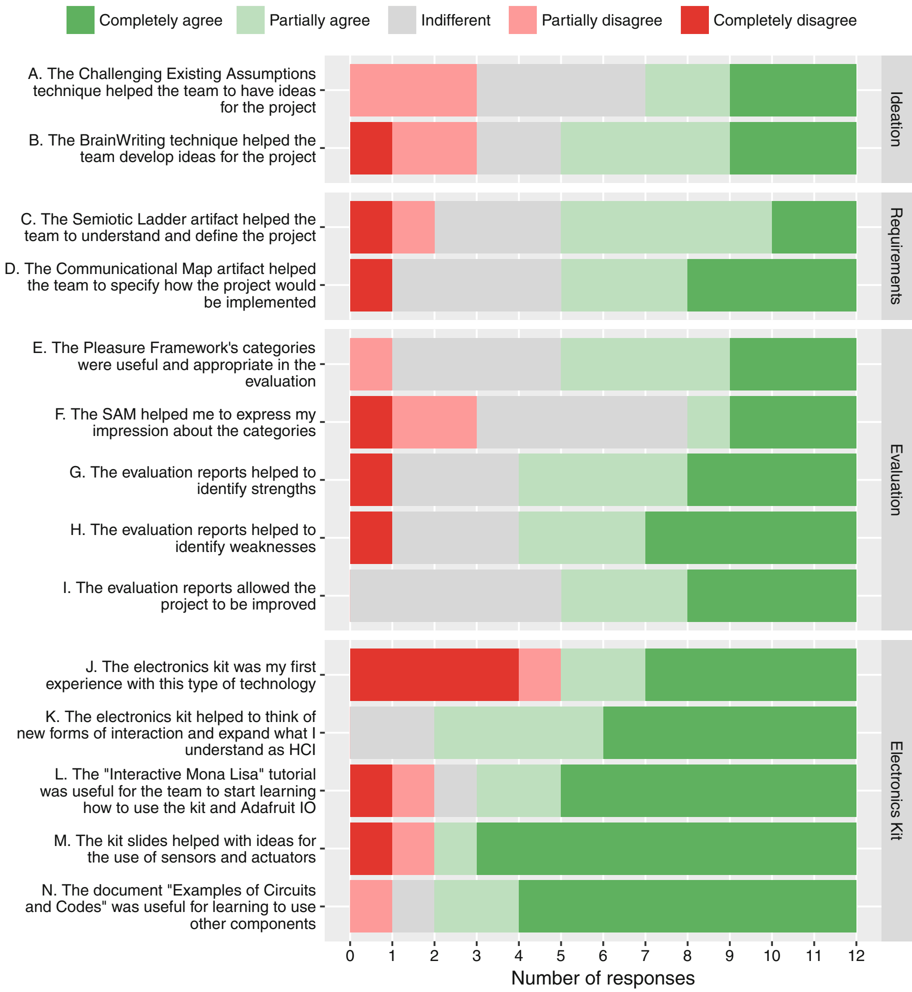


Fig. 4. Student feedback on the used methods and artifacts.

"The Challenging Existing Assumptions basically guided our work, which had as its frame the creation of accessible and creative zoos (contrary to the idea that a person with visual impairment, for example, would not have much to enjoy in such an environment). The BrainWriting, in turn, only had the function of guiding the challenge of assumptions." (Q13)

"The BrainWriting has greatly hindered the creation of the application interface. I do not think the technique works for this purpose, the creation of an interface is much more linked to the system requirements and what the client wants, while the technique led to several different styles of good design being superimposed, creating a polluted design. In the end, my team had to redo everything." (Q14)

Feedback for Requirements Artifacts. Regarding the Semiotic Ladder and Communicational Map artifacts (questions C and D in Fig. 4, respectively), most of the responses are positive for both artifacts. In quotation (Q15), the student argues how the Semiotic Ladder helped him to obtain a more holistic view of the project components, while in quotation (Q16), another student reported the Communicational Map's importance in planning the parts of the project. The student from quotation (Q17), in contrast, argue that these artifacts may not be helpful if there are already well-defined ideas for the project.

"I do not know if my understanding is in accordance with the objectives of the use of the Semiotic Ladder, but the 'climb' from the physical world to the 'meaning' of the presented concepts, organized by the Semiotic Ladder, helped a great deal to understand the role of each component of our project and to communicate the desired message." (Q15)

"The map helped plan the parts of the project, improving the way we organized ourselves." (Q16)

"I believe that, with clear ideas, these artifacts provided little help." (Q17)

Feedback for Evaluation Artifacts. For the Pleasure Framework and SAM artifacts (questions E and F in Fig. 4, respectively), the former had mostly positive responses, while the latter had mixed feedback. Furthermore, the majority agreed that the evaluation reports helped in identifying strengths, weaknesses and allowed the project to be improved. Quotation (Q18) highlights how the Pleasure Framework helped in understanding hedonic qualities, and how the SAM can be useful to identify and categorize feelings. The student from quotation (Q19) emphasizes on how the Pleasure Framework can bring forth aspects overlooked by the team but reports that the SAM was overshadowed by written justifications. The student from quotation (Q20) argues how evaluation is essential in his understanding of interactive art, and that the peer review process was not only useful but necessary. In contrast, the student from quotation (Q21) argues that the SAM can be confusing and was not helpful during the evaluation process.

"The Pleasure Framework's categories were useful because they helped categorize hedonic quality in a way that was not very subjective. The SAM also helped to interpret my impression, because it is not always easy to identify and categorize what we are feeling or what the object makes us feel." (Q18)

"The Pleasure Framework introduced concepts and aspects not previously considered by the team, so it was very useful. The SAM had less attention, considering that the focus of the feedback was on the justifications." (Q19)

"The concept of Interactive Art is precisely the art in symbiosis with its appreciators, so it does not make sense only the creators to evaluate if it is good. The peer review was necessary to identify the positive and negative aspects of the work." (Q20)

"The SAM is confusing if you are not used to the framework, and through it, I was not able to explain what could be improved in the work of the team I evaluated, nor did it help to critically evaluate the project." (Q21)

Feedback for the Electronics Kit. Overall, 7 students reported that this was their first experience with this type of technology, while the remaining 5 students had previous experience (question J in Fig. 4). Ten students agreed that the kit helped them to think of new forms of interaction and expand what they understood as HCI (question K), and no student disagreed. Lastly, most students agreed that the "Interactive Mona Lisa" tutorial, presentation slides and examples of circuits and code were useful (questions L, M, and N). The student from quotation (Q22) considered the kit to be the best experience in the course, and the student from quotation (Q23) praised the quality of the provided material. The student from quotation (Q24), in turn, reported how his concept of what can be an IoT system has been expanded during the course. Lastly, in contrast, the student from quotation (Q24) reported that a lack of previous technical experience could increase the project difficulty.

"It was definitely the best experience I've had in the course." (Q22)

"All the material was very well explained and was a great base for the creation of our project." (Q23)

"It helped a lot to think about new forms of interaction because my concept of IoT was just to automate some actions and to collect data on everyday objects. Even having seen some examples of art using technology, implementing a system whose goal was to get a message across and be interactive has caused a change of perspective." (Q24)

"Anyone with little or no experience with HTML, CSS, and JS had difficulty implementing the project, and the tutorial did not help in this regard." (Q25)

5 Discussion

Even though there is no agreed consensus on what are the topics that compose the field of HCI, Hewett *et al.*'s [14] broadly accepted definition of HCI as "[...] a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" can provide some insights on what, in general, should be taught to HCI students. It seems clear that design alone is not enough, and it needs both evaluation and implementation to form a full cycle. Interactive systems and major phenomena surrounding them, however, are concepts that seem to be constantly challenged by innovation and new contexts of technology use. Therefore, these ever-evolving concepts should often be revisited inside the classroom, preferably with the involvement of the students themselves.

It is important to emphasize that we are not advocating against the design, evaluation, and implementation of more "traditional" interactive systems, designed for well-defined and/or work-related problems. In fact, as presented in Sect. 2, one of the projects of the HCI course addressed in this paper (the redesign one) involved what we consider a traditional system. We are, however, advocating that HCI educators could and should explore new ways of expanding student's perceptions of what can be an interactive system (which is likely to also expand our own). Our approach of inserting art in the course with an interactive art project, for instance, yielded the rich results we presented in Sect. 4, and that we will briefly discuss here. We can highlight the following main aspects:

- **Creative Freedom:** the high degree of freedom we provided students within the InterArt project had several positive effects. As can be seen in Table 1, by choosing their team name and freely designing the interactive artwork, students had the opportunity to express themselves, and represent and pay tribute to artists and artworks of their admiration (which also brought an unexpected cultural value that came from, for instance, appreciating local and relatively unknown artists). This freedom, in turn, also led to a high degree of engagement from students. It was noticeable how many of them cared for the project and felt proud after the final presentation.
- **Participation and Collective Sensemaking:** during the whole course we aimed towards a collective construction of knowledge. Instead of providing authoritative definitions that students would probably listen to in a passive manner, we designed the course's activities to encourage active participation. Students could initially research a subject on their own, but afterward, they would openly discuss it with the class. The perception and interactive art concept warm-ups are two examples if this approach and both led to meaningful

in-class discussions. Both activities showed that encouraging a collective construction of knowledge can be fruitful for most students, as their colleagues are likely to bring in different aspects and points of view that can enrich an initial, individual concept. This effect can be seen in quotations (Q3), (Q4), (Q12), and (Q5).

- **Social Awareness:** during the course, we also encouraged students to think about the concept of socially aware design [2]. In the InterArt project, students could question themselves: if their interactive artwork was placed in a museum, or other public space, who could affect or be affected by it? Could a blind person also appreciate their work? What about other disabilities or conditions? In the Challenging Existing Assumptions technique, for instance, some teams challenged the idea that a blind person could not properly visit a museum. In the end, most of the teams created interactive artworks that do not depend on vision alone, and three of them, 500cc, Lobos-Guará, and Romero Britto, had accessibility as a central aspect in their project. Even though social awareness goes beyond accessibility, this was the students' first experience with the subject and resulted in an expanded view of the impact of working with and designing new technology. This effect can also be seen in quotations (Q3) and (Q13).
- **Expanded View of HCI:** the idea of bringing elements from maker culture to the classroom was intended to provoke and expand what the students previously understood as HCI. The answers to question K in Fig. 4, complemented by quotation (Q24), indicate that this objective was achieved. Quotations (Q22) and (Q23), in turn, indicate that our substantial effort of preparing an electronics kit with hand-picked components and writing meaningful documentation for each one of them was important for the success of the InterArt project. In the end, the projects tacitly explored important vanguard concepts in HCI, such as pervasive and ubiquitous computing, IoT, and enactive systems and embodied interaction.
- **Appropriation of Methods and Artifacts:** when we expand our view of what can be HCI, inevitably we must also expand our view of what can be HCI evaluation. In this study, our appropriation of the Pleasure Framework also seemed important in expanding students' views of qualities to be considered in interactive interfaces, as can be seen in (Q18) and (Q19). We, however, must emphasize that we opted for the Pleasure Framework due to the nature of the InterArt project. HCI researchers and practitioners should be able to critically assess the usefulness of different evaluation methods in different domains and contexts. This is especially important considering the wide range of specific domains and contexts that may arise when we start designing for open-ended scenarios that go beyond traditional, work-related problems.

5.1 Limitations

Besides the positive results, our approach does not come without some limitations. Working with interactive art, for instance, may not resonate well with every student, and quotations (Q6), (Q7), and (Q8) indicate this possibility.

Participatory methods, in turn, may also not be well received by some students, as indicated by quotations (Q14) and (Q17). No approach, of course, will be universally accepted by all the students. The student from quotation (Q11), for instance, already had a preconceived view of what a computer engineer should know about design, which we conjecture to be a more objective, market-oriented perspective. However, our approach did work with some students that first seemed prone to dislike it: the student from (Q10) did not show interest in art at the beginning of the course but in the end reported an overall positive experience.

Another limitation is affordability and know-how of the required electronic components. In our study, we relied on relatively low-cost hardware. The electronics kit that students used, for instance, cost approximately US$60.00 per kit already considering local availability and taxes. Hypothetically, this value can be lowered to around US$20.00 per kit if it is possible to directly import components from China without additional taxes. However, even though this value is relatively affordable, not every institution and program may have a budget available for providing students with this kind of material, especially in economically disadvantaged countries. Furthermore, some of these components are eventually going to need replacement, and there is a constant release of newer, better and/or cheaper alternatives. Besides the actual price, a substantial level of know-how is needed to both initially pick the components to compose a kit, as well as to keep the kits supplied and updated.

Lastly, considering the necessary physical components and the idea of active participation and collective sensemaking, it is possible that our approach may not be suitable for the expanding modality of virtual classes, such as in massive open online courses. Our close contact with the students, particularly during practical activities, seemed to be essential to the success of the InterArt project. Therefore, when necessary, a virtual classroom would need to emulate and/or find proper alternatives to our activities and methods.

6 Conclusion

Considering our specific research question, our study shows that it is possible to teach HCI by employing novel technologies and proposing the design of open-ended scenarios. As an example of an open-ended scenario, or study also shows that interactive art can be used to articulate art and science in a HCI classroom context. The technologies and tools we employed can be useful in expanding students' view on what can be an interactive system. Open-ended scenarios, in turn, may foster creativity and participation. Even though we cannot expect our approach to resonate well with every student (can any approach?), our articulation allowed a significant number of students to rethink their perception of art, technology, and HCI. Furthermore, our approach also provided a relatively open teaching environment for students to explore new technologies (*e.g.*, Arduino, electronics and the Internet of Things), express themselves and their tastes creatively, and ultimately play an active role in enacting the course's main project.

Regarding our general research question, we cannot predict the design problems our students will face in the next 5, 10 or 20 years from now. Nevertheless, we do believe that by exploring new tools and technologies to create novel, unconventional forms of interaction, and by practicing design in a socially aware manner, our students will be somehow prepared for the unforeseen challenges they may face in the future as researchers and practitioners. Nevertheless, by encouraging participation and critical thinking, and being able to work with state of the art technology and interaction techniques, these students may not only be prepared to cope with future design challenges but may indeed play an active role in defining the technology and respective design problems of the future.

Acknowledgements. We would like to thank all the students who participated in this study for embracing the InterArt project and doing your best. This study was financially supported by the National Council for Scientific and Technological Development (CNPq) through grant #308618/2014-9 and by São Paulo Research Foundation (FAPESP) through grants #2015/16528-0 and #2017/06762-0.

References

1. Adajian, T.: The definition of art. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Stanford, CA, USA, summer 2016 edn. (2016)
2. Baranauskas, M.C.C.: Social awareness in HCI. Interactions **21**(4), 66–69 (2014). http://doi.acm.org/10.1145/2621933
3. Bødker, S.: When second wave HCI meets third wave challenges. In: Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles, NordiCHI 2006, pp. 1–8. ACM, New York (2006). http://doi.acm.org/10.1145/1182475.1182476
4. Bødker, S.: Third-wave HCI, 10 years later-participation and sharing. Interactions **22**(5), 24–31 (2015). http://doi.acm.org/10.1145/2804405
5. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)
6. Churchill, E.F., Bowser, A., Preece, J.: Teaching and learning human-computer interaction: past, present, and future. Interactions **20**(2), 44–53 (2013). http://doi.acm.org/10.1145/2427076.2427086
7. Costello, B., Edmonds, E.: A study in play, pleasure and interaction design. In: Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces, DPPI 2007, pp. 76–91. ACM, New York (2007). http://doi.acm.org/10.1145/1314161.1314168
8. Cressey, D.: Age of the arduino. Nature **544**(7648), 125–126 (2017)
9. Edmonds, E.A.: Human computer interaction, art and experience. In: Candy, L., Ferguson, S. (eds.) Interactive Experience in the Digital Age. SSCC, pp. 11–23. Springer, Cham (2014). http://dx.doi.org/10.1007/978-3-319-04510-8_2
10. England, D.: Art.CHI: curating the digital. In: England, D., Schiphorst, T., Bryan-Kinns, N. (eds.) Curating the Digital. SSCC, pp. 1–7. Springer, Cham (2016). http://dx.doi.org/10.1007/978-3-319-28722-5_1

11. Grandhi, S.: Educating ourselves on HCI education. Interactions **22**(6), 69–71 (2015). http://doi.acm.org/10.1145/2834811
12. Grudin, J.: A moving target: the evolution of human-computer interaction. In: Jacko, J.A. (ed.) Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications. Human Factors and Ergonomics, 3rd edn., pp. xxvii-lxi. CRC Press, Boca Raton (2012)
13. Harrison, S., Tatar, D., Sengers, P.: The three paradigms of HCI. In: Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems, pp. 1–18. ACM, San Jose (2007)
14. Hewett, T.T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., Verplank, W.: ACM SIGCHI curricula for human-computer interaction. Technical report, The Association for Computing Machinery, New York, NY, USA (1992)
15. Hornbæk, K., Oulasvirta, A.: What is interaction? In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI 2017, pp. 5040–5052. ACM, New York (2017). http://doi.acm.org/10.1145/3025453.3025765
16. Preece, J., Helen Sharp, Y.R.: Interaction Design: Beyond Human-Computer Interaction, 4th edn. Wiley, Chichester (2015)
17. Lasseter, J.: John Lasseter on the Future of Animation: Future of Animation Panel at The Academy's Development of the Digital Animator, May 2012. https://youtu.be/uFbOOjAC_Fg
18. Liu, K.: Semiotics in Information Systems Engineering. Cambridge University Press, Cambridge (2000)
19. Löwgren, J.: On the significance of making in interaction design research. Interactions **23**(3), 26–33 (2016). http://doi.acm.org/10.1145/2904376
20. Michalko, M.: Thinkertoys: A Handbook of Creative-Thinking Techniques. Potter/TenSpeed/Harmony, Berkeley (2010)
21. Muller, L., Edmonds, E., Connell, M.: Living laboratories for interactive art. CoDesign **2**(4), 195–207 (2006). http://dx.doi.org/10.1080/15710880601008109
22. Novak, G.M.: Just-in-time teaching. New Directions Teach. Learn. **2011**(128), 63–73 (2011). http://dx.doi.org/10.1002/tl.469
23. Posch, I.: Crafting tools. Interactions **24**(2), 78–81 (2017). http://doi.acm.org/10.1145/3038227
24. Rocha, H.V., Baranauskas, M.C.C.: Design e avaliação de interfaces humano-computador. Unicamp (2003)
25. Rogers, Y.: HCI theory: classical, modern, and contemporary. Syn. Lect. Hum. Centered Inform. **5**(2), 1–129 (2012). https://doi.org/10.2200/S00418ED1V01Y201205HCI014
26. Schuler, D., Namioka, A.: Participatory Design: Principles and Practices. Taylor & Francis, London (1993)
27. True, N., Peeters, J., Fallman, D.: Confabulation in the time of transdisciplinarity: reflection on HCI education and a call for conversation. In: Kurosu, M. (ed.) HCI 2013. LNCS, vol. 8004, pp. 128–136. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39232-0_15
28. VanGundy, A.B.: Brain writing for new product ideas: an alternative to brainstorming. J. Consum. Market. **1**(2), 67–74 (1984). https://doi.org/10.1108/eb008097
29. Wilson, S.: Science and Art - Looking Backward/looking Forward (2003). http://www.artcenter.edu/williamson/neuro/exhibit/f_body_wilson.html
30. Zhu, R.J., Mehta, R.: Sensory experiences and consumer creativity. J. Assoc. Consum. Res. **2**(4), 472–484 (2017). https://doi.org/10.1086/693161

UX Evaluation Design of UTAssistant: A New Usability Testing Support Tool for Italian Public Administrations

Stefano Federici[1(✉)], Maria Laura Mele[1], Rosa Lanzilotti[2], Giuseppe Desolda[2], Marco Bracalenti[1], Fabio Meloni[1], Giancarlo Gaudino[3], Antonello Cocco[3], and Massimo Amendola[3]

[1] Department of Philosophy, Social and Human Sciences and Education, University of Perugia, Perugia, Italy
stefano.federici@unipg.it, marialaura.mele@gmail.com, marco.bracalenti90@gmail.com, fa.meloni@gmail.com

[2] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
{rosa.lanzilotti,giuseppe.desolda}@uniba.it

[3] ISCOM – Superior Institute of Communication and Information Technologies, Ministry of Economic Development, Rome, Italy
{giancarlo.gaudino,antonello.cocco, massimo.amendola}@mise.gov.it

Abstract. Since 2012, usability testing in Italian public administration (PA) has been guided by the eGLU 2.1 technical protocols, which provide a set of principles and procedures to support specialized usability assessments in a controlled and predictable way. This paper describes a new support tool for usability testing that aims to facilitate the application of eGLU 2.1 and the design of its User eXperience (UX) evaluation methodology. The usability evaluation tool described in this paper is called UTAssistant (Usability Tool Assistant). UTAssistant has been entirely developed as a Web platform, supporting evaluators in designing usability tests, analyzing the data gathered during the test and aiding Web users step-by-step to complete the tasks required by an evaluator. It also provides a library of questionnaires to be administered to Web users at the end of the usability test. The UX evaluation methodology adopted to assess the UTAssistant platform uses both standard and new bio-behavioral evaluation methods. From a technological point of view, UTAssistant is an important step forward in the assessment of Web services in PA, fostering a standardized procedure for usability testing without requiring dedicated devices, unlike existing software and platforms for usability testing.

Keywords: Experimental UX evaluation methodology · Usability evaluation tool · Public administration · UX Semi-Automatic assessment · International usability standards

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 55–67, 2018.
https://doi.org/10.1007/978-3-319-91238-7_5

1 Introduction

This paper describes the design of an experimental methodology that aims to evaluate the User eXperience (UX) of a new Web platform, called UTAssistant (Usability Tool Assistant) [1]. This is a semi-automatic usability evaluation tool that supports practitioners in usability evaluations of Web systems and services provided by a public administration (PA), according to the eGLU 2.1 technical protocol [2]. This protocol provides a set of principles and procedures to support specialized usability assessments in a controlled and predictable way.

The UX evaluation design of UTAssistant described in this paper is an experimental methodology for assessing the UTAssistant platform with end-users and Web managers of PA Web sites, both in a laboratory setting and using a Web-based recruitment platform. The methodology proposed here includes several types of end-users, with the aim of assessing (i) the UTAssistant method through bio-behavioral measurements; (ii) the usability evaluation process of UTAssistant with Web managers in Italian PA; (iii) a heuristic evaluation of UTAssistant conducted by experts in UX; and (iv) a usability evaluation of UTAssistant with a highly representative number of end-users using a Web-based recruitment platform.

2 Usability Testing of Italian Public Administration Web Services

In October 2012, the Department of Public Function of the Italian Ministry for Simplification and Public Administration formed a working group called GLU (Working Group on Usability). The GLU team was composed of Italian universities, central and local Italian PAs, and other independent information and communication companies. The purpose of GLU is to support PA practitioners involved in Web content management, website development, or e-government systems development in performing usability evaluations, and particularly those who are not usability experts. The primary goal of GLU is to collect and identify golden rules for developing and evaluating systems that are easy to use and appropriate for this purpose. To this end, GLU developed a set of guiding protocols that are able to operatively support both the analysis and evaluation of graphical user interfaces for the Web. GLU can guide Web masters, and its protocols are explorative tools that can investigate how good or satisfactory the experience is for a user when using a PA Web service, e.g. searching for certain information, consulting or downloading a digital document, or completing an online form. GLU protocols guide PA practitioners in exploratory analyses to better understand the problems (or strengths) of their Web services, in order to collect use cases for future development. Since 2013, GLU has developed four different usability evaluation protocols called eGLU 1.0, eGLU 2.0, eGLU 2.1 [2], and eGLU-M [3]. Three protocols (1.0 and 2.0) are designed for desktop solutions (2.1), while the other (eGLU-M) is designed for mobile platforms [3].

The eGLU 1.0 protocol was developed in May 2013 [4]. The protocol involves two levels of analysis, basic and advanced, which can be used independently of each other according to the testing period and the practitioner's skill. The basic level is specifically recommended for performing quick analyses to check the main problems affecting the usability of a short number of Web pages. It is a macroscopic analysis that asks users to freely navigate the content of the main pages of a given Web service, and then to complete a questionnaire to investigate the quality of the interaction. In a basic level analysis, practitioners primarily collect information on how many navigation tasks users achieved or failed, how difficult it was for users to perceive or understand Web interface elements, and user satisfaction. The advanced level analysis is recommended for practitioners who need a more detailed analysis of interaction problems. At this level, participants are required to report their actions and thoughts during their interaction with the system. Compared to the basic level, an advanced analysis allows practitioners a greater level of detail and information on user interactions, both in terms of the users' navigation paths and the difficulties they encountered in perceiving or understanding information during the tasks. Both basic and advanced levels describe how to create and describe tasks for users, how to set parameters, the apparatus involved, and the selection of participants. The eGLU 1.0 protocol provides practitioners with practical advice on how to properly conduct the test, including how to verbally describe both the goals of the test and the instructions to participants. Both the basic and advanced levels follow five phases, which describe: (i) how to prepare testing documents; (ii) how to prepare tools and materials; (ii) how to conduct the test; (iv) how to handle the collected data; and (v) how to draw up the evaluation report. eGLU 1.0 recommends the use of at least one of two usability assessment questionnaires: (i) the System Usability Scale (SUS) [5, 6] or the Usability Evaluation (Us.E. 2.0) questionnaire [7].

The eGLU 2.0 protocol was released in 2014. Compared to eGLU 1.0, eGLU 2.0 provides practitioners with an easier and simpler methodology for conducting evaluation tests, together with a wide range of design and evaluation approaches and methods from which practitioners can freely choose according to their needs. eGLU 2.0 consists of two parts: the first gives recommendations and instructions to practitioners on how to design and conduct tests, while the second focuses on advanced design methods and evaluation techniques, and describes which alternative and/or complementary usability methods can be used.

In the same way as eGLU 1.0, eGLU 2.0 offers a first-level usability test methodology that is suitable for both expert and non-expert usability evaluation practitioners. eGLU 2.0 involves three phases, which describe how to (i) prepare, (ii) execute, and (iii) analyze the results. The protocol recommends using at least one of three usability assessment questionnaires: (i) the SUS [5, 6], the Us.E. 2.0 questionnaire [7], and the Usability Metric for User Experience, lite version (UMUX-LITE) [8–10]. The second part of eGLU 2.0 involves several in-depth analyses of and extensions to the basic procedure. These schedules can be useful in planning, conducting or analyzing the interaction, and increase the possibility of intervention via Web site redesign by providing elements from a broader and more complex range of methodological approaches compared to the basic protocol procedure. The advanced techniques described in eGLU 2.0 are the kanban board, scenarios and personas, evaluation strategies using the think-aloud verbal protocol, the methodology of the ASPHI non-profit organization foundation (http://www.asphi.it/), and the usability cards method (http://www.usabilitycards.com/).

An updated version of the methods and techniques proposed in eGLU 2.0 was developed in 2015 [2] with the eGLU 2.1 protocol. eGLU 2.1 is distributed together with the eGLU-M (eGLU-mobile) protocol [3], which is specifically designed for usability evaluations using mobile devices. Although the evaluation of mobile websites and Web services has some aspects that are operationally different from evaluations using desktop devices, the approach, methodology and phases of the exploratory analysis procedure remain substantially unchanged. The development of a new version of the protocol is currently in progress, and its release is expected in 2018.

3 UTAssistant: A New Usability Testing Support Web Platform for Italian Public Administration

UTAssistant is a Web platform, designed and developed within the PA++ Project. The goal of this platform is to provide Italian PA with a lightweight and simple tool for conducting user studies based the eGLU 2.1 protocol, without requiring installation on user devices.

One of the most important requirements driving the development of this platform was the need to perform remote usability tests with the aim of stimulating users to participate in a simpler and more comfortable way. To accomplish this, UTAssistant was developed as a Web platform so that the stakeholders involved, namely the evaluator (Web manager of a PA site) and users (typically of PA Web sites), can interact using their PCs, wherever and whenever they prefer. This is possible due to the recent evolutions of the HTML5 and JavaScript standards, which allow Web browsers to gather data from PC devices such as the webcam, microphone, mouse, and keyboard. This represents an important contribution to state-of-the-art of usability test tools, since remote participation fosters wider adoption of these tools and consequently of the usability testing technique. Indeed, the existing tools for usability testing require software installation on a PC with specific requirements (e.g. Morae® https://www.techsmith.com/morae.html [11]).

The following sub-sections describe how UTAssistant supports evaluators in designing a usability test and analyzing the results, and how users are supported by UTAssistant in completing the evaluation tasks.

3.1 Usability Test Design

A usability test starts from the test design, which mainly consists of: (i) creating a script to introduce the users to the test; (ii) defining a set of tasks; (iii) identifying the data to be gathered (e.g. the number of clicks and the time required by the user to accomplish a task, audio/video/desktop recording, logs, etc.); and (iv) deciding which questionnaire(s) to administer to users.

UTAssistant facilitates evaluators in performing these activities by means of three wizard procedures. The first guides evaluators in specifying: (a) general information (e.g. a title, the script); (b) the data to gather during execution of the user task (e.g. mouse/keyboard data logs, webcam/microphone/desktop recordings); and (c) the post-test questionnaire(s) to administer. The second procedure assists evaluators in creating the task lists; for each task, start/end URLs, the goal and the duration have to be specified. Finally, the third procedure requires evaluators to select the users, either from a list of users already registered to the platform, or by typing their email addresses. The invited users receive an email including the instructions for participating in the usability test. The following sub-section illustrates how UTAssistant aids users in performing the test.

3.2 Usability Test Execution

Following the creation of the usability test design, users receive an email with information about the evaluation they are asked to complete, and a link to access UTAssistant. After clicking on this link, users can carry out the evaluation test, which starts by giving general information about the platform use (e.g. a short description of the toolbar with useful commands), the script for the evaluation and, finally, privacy policies indicating which data will be captured, such as mouse/keyboard logs and webcam/microphone/desktop recordings.

Following this, UTAssistant administers each task, one at a time. The execution of each task is closely guided by the platform, which shows the task description in a pop-up window, and then opens the Web page at which users are asked to start the task (Fig. 1). To keep the platform as minimally invasive as possible during execution of the evaluation test, we grouped all the functions and indications in a toolbar placed at the top of the Web page. This toolbar indicates the title of the current task, its goal, the duration of the task, the task number, and a button to move to the next task, which shows the message "Complete Questionnaire" when the user finishes the last task and is asked to complete the questionnaire(s). During execution of the task, the platform collects all data identified by the evaluator at the design stage, in a transparent and non-invasive way.

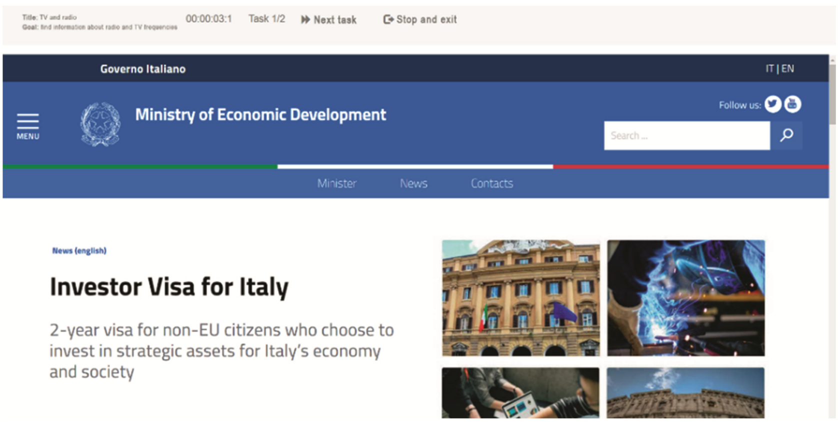


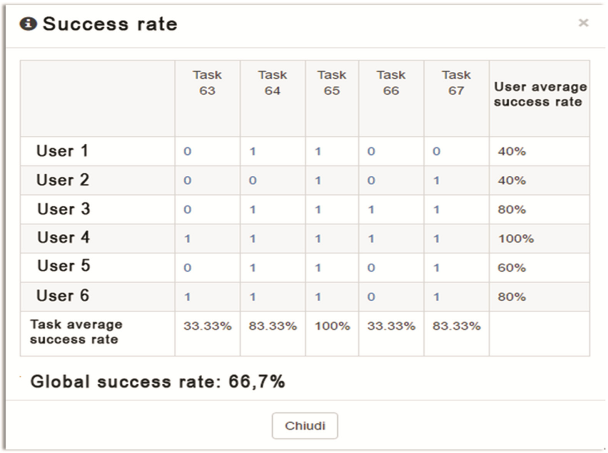
Fig. 1. An example of execution of a task. The UTAssistant toolbar is shown at the top of the evaluated website page.

3.3 Evaluation of Test Data Analysis

One of the most time-consuming phases of a usability test is the data analysis, since evaluators are required to manually collect, store, merge and analyze a huge amount of data such as mouse logs, video/audio recordings and questionnaire results. Due to the effort required, this phase becomes a deterrent towards the adoption of usability testing techniques. UTAssistant automates all of these activities, thus removing the barriers to the analyses of usability test data. The evaluators have access to the data analysis results via the control panel and can exploit several functionalities that provide useful support in discovering usability issues. The next sub-sections present an overview of some of these tools.

3.4 Task Success Rate (Effectiveness)

Analysis of the results of the usability test often starts by investigating the task success rate, an essential indicator of the effectiveness of the website in supporting the execution of a set of tasks. This metric is calculated as the percentage of tasks correctly completed by users. It can be also calculated for each task, thereby estimating the percentage of users who completed that task. UTAssistant calculates these frequencies and displays them in a table (Fig. 2).



Success rate

	Task 63	Task 64	Task 65	Task 66	Task 67	User average success rate
User 1	0	1	1	0	0	40%
User 2	0	0	1	0	1	40%
User 3	0	1	1	1	1	80%
User 4	1	1	1	1	1	100%
User 5	0	1	1	0	1	60%
User 6	1	1	1	0	1	80%
Task average success rate	33.33%	83.33%	100%	33.33%	83.33%	

Global success rate: 66,7%

Chiudi

Fig. 2. Example of a table reporting the success rates of a study. The columns display the tasks, while the rows show a list of users. The last row reports the success rate for each task, while the last column depicts the success rate for each user. The overall success rate is reported below the table.

3.5 Questionnaire Results

Another phase requiring a great deal of effort by evaluators is the analysis of the results of the questionnaire. Using UTAssistant, evaluators can administer one or more questionnaires at the end of each usability evaluation. The platform automatically stores the user's answers and produces results in the form of statistics and graphs. For example, if the SUS [5, 6] questionnaire is used, UTAssistant calculates the global SUS score (a unidimensional measure of perceived usability [12]), the usability score and the learnability score. In addition, different visualizations can display these results from different perspectives, e.g. a histogram of each user's SUS scores, a box-plot of SUS score/learnability/usability (Fig. 3), and a score compared with the SUS evaluation scales (Fig. 3).

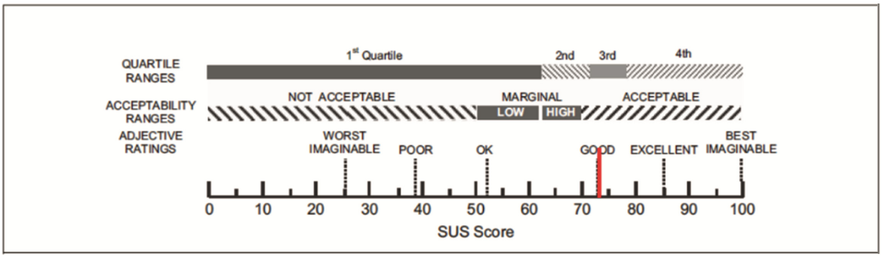


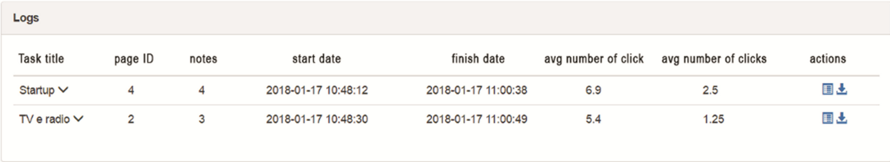
Fig. 3. Example of SUS score plotted on the three SUS scales.

3.6 Audio/Video Analysis

While users are executing the tasks, UTAssistant can record the user's voice using the microphone, their facial expressions using the webcam, and the desktop display using a browser plugin. This recorded content can be analyzed by evaluators in order to understand, for example, low performance in executing a particular task or the reasons for a low success rate. To support a more effective audio/video analysis, UTAssistant provides annotation tools, so that when evaluators detect the existence of difficulties, as indicated by means of verbal comments or facial expressions, they can annotate the recorded audio/video tracks. If the evaluators decide to record both camera and desktop videos, the video tracks are merged and displayed together.

3.7 Mouse/Keyboard Logs Analysis (Efficiency)

Important information about the efficiency of performing tasks is given by metrics such as the time and number of clicks required to complete each task. UTAssistant tracks the user's behavior by collecting mouse and keyboard logs. Based on the collected data, the platform shows performance statistics for each task, such as the number of pages visited, the average number of clicks and the time that each user needed to complete the task (Fig. 4).



Logs

Task title	page ID	notes	start date	finish date	avg number of click	avg number of clicks	actions
Startup	4	4	2018-01-17 10:48:12	2018-01-17 11:00:38	6.9	2.5	
TV e radio	2	3	2018-01-17 10:48:30	2018-01-17 11:00:49	5.4	1.25	

Fig. 4. Summary of metrics measuring performance related to three tasks.

4 UX Evaluation Methodology for Assessing UTAssistant

4.1 Methodology

The UX evaluation design proposed here is an experimental methodology, consisting of four phases:

- Phase 1. Heuristic evaluation of the UTAssistant platform;
- Phase 2. Usability evaluation with PA practitioners, under workplace conditions;
- Phase 3. Usability evaluation with Web end-users, under experimental laboratory conditions;
- Phase 4. Usability evaluation with Web end-users, under remote online conditions.

4.2 Objective

This experimental methodology aims to provide a new approach to assessment of the UTAssistant semi-automatic usability evaluation tool. This methodology combines expert assessment methods with usability evaluation models, under workplace, laboratory, and remote online conditions. The implementation of the UX evaluation methodology for the UTAssistant platform is planned for future work.

4.3 Methods and Techniques

The experimental methodology proposed here involves both usability assessment and psychophysiological measurement methods. Different methods and techniques are used in each phase, as described below.

Phase 1. Heuristic Evaluation. This is an inspection method [13–16], which consists of experts assessing the usability of a product. In general, the experts involved in heuristic evaluation use a list of principles, also called heuristics, to compare the product with a baseline representing how the product should meet the main usability requirements. Heuristics are based on sets of features for the ideal matching of a user model. At the end of each evaluation, the expert carrying out a heuristic evaluation provides a list of problems and related suggestions.

Phase 2. Usability Evaluation Under Workplace Conditions. During Phase 2, users follow the evaluation methodology provided in the eGLU 2.1 protocol, as explained above in Sect. 2. A tailored procedure that applies the eGLU 2.1 protocol for usability evaluation tests is provided to the PAs involved in the UTAssistant experimental project. This protocol differs from eGLU 2.1 in that it uses the think-aloud (TA) technique rather than the partial concurrent think-aloud (PCTA) technique (see Sect. 4.3, Phase 3). The TA technique is used in traditional usability testing methods [17–24] and is especially useful in indoor conditions such as laboratories or workplaces. The TA technique asks users to verbalize ("think aloud") each action and the problems they encounter during their interaction with the system. Evaluators are asked to transcribe and analyze each user action in order to identify interaction problems.

Phase 3. Usability Evaluation Under Laboratory Conditions. Usability testing of the UTAssistant platform is also conducted under laboratory conditions. In this phase, evaluators use the PCTA technique, created by some of the current authors [25–28] to provide a technique for easily comparing collected data with blind, cognitively disabled, and non-disabled users. The PCTA technique asks users to interact silently with the interface and to ring a bell on the desk whenever they identify a problem. In the PCTA technique, all user interactions are registered. As soon as the test is complete, the user is invited to identify and verbalize any problems experienced during the interaction [29].

Any psychophysiological reactions of the users that may occur during this interaction are measured using two bio-behavioral measurement techniques: (i) facial expression recognition; and (ii) electroencephalography (EEG). The EEG method allows practitioners to record the electrical activity generated by the brain using electrodes placed on the user's scalp. Due to its high temporal resolution, the EEG is able to analyze which areas of the brain are active at any given moment. The scientific community also recognizes a limited number of facial expressions (about 45) as universally able to express hundreds of emotions resulting from the combination of seven basic emotions [30]: joy, anger, surprise, fear, contempt, sadness, and disgust. In human beings, the user is mostly unaware of the ways in which facial muscles express basic emotions [31]. An analysis of involuntary facial expressions returns information about the emotional impact on the users of an interaction with a given interface.

Phase 4. Usability Evaluation Under Remote Online Conditions. In this phase, users are recruited through a Web recruitment platform and redirected to the UTAssistant Web platform. This methodology has previously been validated for psychological studies [32, 33].

4.4 Material and Equipment

Phases 1, 3, and 4 use the UTAssistant platform to evaluate the Ministry of Economic Development (MiSE) website (http://www.sviluppoeconomico.gov.it), while Phase 2 uses the platform to evaluate the websites of each PA involved under workplace conditions. All phases are conducted using either desktop or laptop computer with a screen size of between 13″ and 15″, and a minimum resolution of 1024×640. Computers should be equipped with a Google Chrome browser (http://www.google.com/intl/en/chrome). Computers should be plugged into a power source, and the brightness of the display should be set to the maximum level. In each phase, different materials and equipment are used, as described below.

Phase 1. Heuristic Evaluation. Many heuristic lists are proposed in the literature [29]. In this work, we use 10 heuristics for Web interface analysis created by Nielsen and Molich [16]; these take into account many aspects of the user interaction such as safety, flexibility, and efficiency of use. The Nielsen heuristics are based on 10 principles derived from a factorial analysis carried out on a list of 249 problems detected by many usability evaluations.

Phase 2. Usability Evaluation Under Workplace Conditions. This phase uses a tailored protocol asking managers of PA websites to evaluate them in conjunction with users. This evaluation should be done using the UTAssistant platform with a desktop or laptop computer.

Phase 3. Usability Evaluation Under Laboratory Conditions. In this phase, UX experts are asked to measure user interaction by means of two bio-behavioral measurement devices: a facial expression recognition system, and an EEG. Both devices return data that can be synchronized using a biometric synchronization platform called iMotions (http://imotions.com).

Phase 4. Usability Evaluation Under Remote Online Conditions. Tests are administered using an online recruitment procedure involving a crowdsourcing platform for psychological research called Prolific Academic (http://www.prolific.ac).

4.5 Subjects

Phase 1. Heuristic Evaluation. A heuristic evaluation requires a small set of between three and five expert evaluators.

Phase 2. Usability Evaluation Under Workplace Conditions. PA Web managers are asked to conduct their tests with a minimum of five participants.

Phase 3. Usability evaluation under laboratory conditions. Ten participants are involved, equally divided by gender.

Phase 4. Usability Evaluation Under Remote Online Conditions. One hundred users should be recruited. Participants should be equally divided by gender and language (50 native English speakers, and 50 native Italian speakers).

4.6 Procedure

Phase 1. Heuristic Evaluation. Experts are asked to evaluate the main actions required by the UTAssistant platform to assess a website. In particular, experts are asked to evaluate the user experience of an evaluator using UTAssistant during the following actions:

- Create a new usability test with UTAssistant in order to evaluate the MiSE website.
- Define four user tasks.
- Determine which user questionnaires will be administered to users at the end of the test.
- Define which export data the system should record during the interaction.
- Export navigation, questionnaire and log data.
- Use the help function.

Phase 2. Usability Evaluation Under Workplace Conditions. The Web managers involved in this phase are asked to evaluate the usability of their PA website. Web managers are asked to perform the same actions as required in Phase 1, and then to evaluate their websites with users recruited from within their workplace. Users should be asked to navigate the administration website to carry out four tasks, presented in the form of usage scenarios. A help service embedded into the platform is provided to users, which activates an error message and automatically sends a request to a remote help service.

Phase 3. Usability Evaluation Under Laboratory Conditions. In this phase, users are required to perform the test in a quiet and sufficiently bright environment, using a comfortable chair placed at least 50 cm from the screen of a desktop or laptop computer. Users are asked to navigate the MiSE website to carry out the four tasks previously created by the UX expert conducting the sessions. Tasks should be presented to the users in form of usage scenarios.

Phase 4. Usability Evaluation Under Remote Online Conditions. Online participants should be redirected to the UTAssistant platform to evaluate the MiSE website. In this phase, participants are asked to set up their devices as required in Phases 1, 2, and 3, and to perform the same tasks as defined in Phase 3, presented in the form of scenarios.

4.7 Data Collection

At the end of each phase, evaluators are asked to store their collected data in a database hosted by the Superior Institute of Communication and Information Technologies (ISCOM), Italy. Stored data will be analyzed, phase by phase, in aggregate form. Statistical analyses and comparisons will be carried out using the IBM SPSS platform, and then discussed and disseminated through reports and conference papers.

5 Conclusion

The paper describes UTAssistant, a semi-automatic usability assessment web platform for Italian PA, and proposes a new experimental evaluation methodology for assessing the UX of the proposed platform. Both UTAssistant and the experimental assessment methodology were developed as part of a multidisciplinary project involving design engineers, UX experts and PA Web managers. UTAssistant is a new tool aimed at the international scientific community; its goal is to provide a standardized model to guide non-experts in the usability of PA websites, in a quick and straightforward way, to meet international usability protocols and standards. Unlike the most commonly usability evaluation methods, the assessment methodology proposed for evaluating the UTAssistant platform uses bio-behavioral measures in addition to the standard validated usability assessment methodologies. The methodology proposed here provides an evaluation strategy that avoids the involvement of social desirability factors (often related to explicit satisfaction questionnaires), since bio-behavioral measures are hidden from users. This work is part of a two-year project (2017–2018) involving Italian PAs, the University of Bari and the University of Perugia. In future work, the proposed experimental methodology will be implemented to assess the UX of the UTAssistant platform.

References

1. Desolda, G., Gaudino, G., Lanzilotti, R., Federici, S., Cocco, A.: Utassistant: a web platform supporting usability testing in italian public administrations. In: CHItaly 12th edn. CHItaly 2017, pp. 138–142 (2017)
2. Dipartimento della Funzione Pubblica: Il Protocollo Eglu 2.1: Come Realizzare Test Di Usabilità Semplificati Per I Siti Web E I Servizi Online Delle Pa. Formez PA, Rome, IT (2015)
3. Dipartimento della Funzione Pubblica: Il Protocollo Eglu-M: Come Realizzare Test Di Usabilità Semplificati Per I Siti Web E I Servizi Online Delle Pa. Formez PA, Rome, IT (2015)
4. Borsci, S., Federici, S., Mele, M.L.: Eglu 1.0: Un Protocollo Per Valutare La Comunicazione Web Delle Pa. Diritto e Pratica Amministrativa, vol. 1, pp. 9–10. Il Sole 24 Ore, Milan, IT (2014)
5. Brooke, J.: SUS: a “Quick and Dirty” usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London (1996)
6. Borsci, S., Federici, S., Lauriola, M.: On the dimensionality of the System Usability Scale (SUS): a test of alternative measurement models. Cogn. Process. **10**, 193–197 (2009)
7. Di Nocera, F.: Usability Evaluation 2.0: Una Descrizione (S)Oggettiva Dell’usabilità. In: Ergoproject edn., Rome, IT (2013)
8. Borsci, S., Federici, S., Mele, M.L., Conti, M.: Short scales of satisfaction assessment: a proxy to involve disabled users in the usability testing of websites. In: Kurosu, M. (ed.) HCI 2015. LNCS, vol. 9171, pp. 35–42. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21006-3_4
9. Borsci, S., Federici, S., Bacci, S., Gnaldi, M., Bartolucci, F.: Assessing user satisfaction in the era of user experience: comparison of the SUS, UMUX and UMUX-LITE as a function of product experience. Int. J. Hum. Comput. Interact. **31**, 484–495 (2015)
10. Lewis, J.R., Utesch, B.S., Maher, D.E.: UMUX-LITE: when there’s no time for the SUS. In: Conference on Human Factors in Computing Systems: CHI 2013, pp. 2099–2102 (2013)
11. Tomlin, C.: 14 Usability Testing Tools Matrix and Comprehensive Reviews. Useful Usability, vol. 2018, USA (2014)
12. Lewis, J.R., Sauro, J.: Revisiting the factor structure of the system usability scale. J. Usability Stud. **12**, 183–192 (2017)
13. Molich, R., Nielsen, J.: Improving a human-computer dialogue. Commun. ACM **33**, 338–348 (1990)
14. Nielsen, J.: Heuristic Evaluation. In: Nielsen, J., Mack, R.L. (eds.) Usability Inspection Methods. Wiley, New York (1994)
15. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: SIGCHI Conference on Human Factors in Computing Systems: CHI 1994, pp. 152–158 (1994)

16. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: SIGCHI Conference on Human Factors in Computing Systems: CHI 1990, pp. 249–256 (1990)
17. Bettman, J.R.: An Information Processing Theory of Consumer Choice. Addison-Wesley, Cambridge (1979)
18. Bettman, J.R., Park, C.W.: Implications of a constructive view of choice for analysis of protocol data: a coding scheme for elements of choice processes. Adv. Consum. Res. **7**, 148–153 (1980)
19. Biehal, G., Chakravarti, D.: Experiences with the Bettman-Park verbal-protocol coding scheme. J. Consum. Res. **8**, 442–448 (1982)
20. Biehal, G., Chakravarti, D.: Information-presentation format and learning goals as determinants of consumers' memory retrieval and choice processes. J. Consum. Res. **8**, 431–441 (1982)
21. Biehal, G., Chakravarti, D.: Consumers' use of memory and external information in choice: macro and micro perspectives. J. Consum. Res. **12**, 382–405 (1986)
22. Biehal, G., Chakravarti, D.: The effects of concurrent verbalization on choice processing. J. Mark. Res. (JMR) **26**, 84–96 (1989)
23. Green, A.: Verbal protocol analysis. Psychologist **8**, 126–129 (1995)
24. Kuusela, H., Spence, M.T., Kanto, A.J.: Expertise effects on prechoice decision processes and final outcomes: a protocol analysis. Eur. J. Mark. **32**, 559 (1998)
25. Federici, S., Borsci, S., Stamerra, G.: Web usability evaluation with screen reader users: implementation of the partial concurrent thinking aloud technique. Cogn. Process. **11**, 263–272 (2010)
26. Federici, S., Borsci, S., Mele, M.L.: usability evaluation with screen reader users: a video presentation of the Pcta's Experimental Setting and Rules. Cogn. Process. **11**, 285–288 (2010)
27. Federici, S., Borsci, S.: Usability evaluation: models, methods, and applications. In: Stone, J., Blouin, M. (eds.) International Encyclopedia of Rehabilitation, pp. 1–17. Center for International Rehabilitation Research Information and Exchange (CIRRIE), Buffalo (2010)
28. Borsci, S., Federici, S.: The partial concurrent thinking aloud: a new usability evaluation technique for blind users. In: Emiliani, P.L., Burzagli, L., Como, A., Gabbanini, F., Salminen, A.-L. (eds.) Assistive Technology from Adapted Equipment to Inclusive Environments: AAATE 2009, vol. 25, pp. 421–425. IOS Press, Amsterdam (2009)
29. Borsci, S., Kurosu, M., Federici, S., Mele, M.L.: Computer Systems Experiences of Users with and without Disabilities: An Evaluation Guide for Professionals. CRC Press, Boca Raton (2013)
30. Ekman, P.: Darwin and Facial Expression: A Century of Research in Review. Academic Press, New York (1973)
31. Qu, F., Yan, W.-J., Chen, Y.-H., Li, K., Zhang, H., Fu, X.: "You Should Have Seen the Look on Your Face…": self-awareness of facial expressions. Front. Psychol. **8**, 832 (2017)
32. Mele, M.L., Millar, D., Rijnders, C.E.: Validating a quality perception model for image compression: the subjective evaluation of the Cogisen's image compression plug-in. In: Kurosu, M. (ed.) HCI 2016. LNCS, vol. 9731, pp. 350–359. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39510-4_33
33. Mele, M.L., Millar, D., Rijnders, C.E.: Using spatio-temporal saliency to predict subjective video quality: a new high-speed objective assessment metric. In: Kurosu, M. (ed.) HCI 2017. LNCS, vol. 10271, pp. 353–368. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58071-5_27

The Development of Individuals' Competencies as a Meaningful Process of the Audiovisual Design Methodology

Daniel Gambaro[1], Valdecir Becker[2(✉)], Thais Saraiva Ramos[3], and Rafael Toscano[2]

[1] School of Communications and Arts (PPGMPA), University of São Paulo, São Paulo, Brazil
[2] Informatics Center, Federal University of Paraiba, João Pessoa, Brazil
valdecir@ci.ufpb.br
[3] Audiovisual Design Research Group, CNPq, São Paulo, Brazil
audiovisualdesign@lavid.ufpb.br

Abstract. This article presents and classifies the competencies inherent to Audience, Synthesizer, Modifier, Producer and Player, roles that individuals can assume in the Audiovisual Design model. It presents a descriptive table of core competencies, technical as well as technological skills and behaviors, knowledge and capabilities related to each role. In Audiovisual Design model, individuals have certain abilities to act within each role. Audiovisual Design is a communication model and a method to analyze and to develop content combining audiovisual elements and software, or digital interfaces, in interactive systems. It is based on integration of Human-Computer Interaction elements with Audience Studies, essential for the development of modern audiovisual content.

Keywords: Audiovisual Design · Human Computer Interaction
Audience Studies · Competencies · Methodology

1 Introduction

The creation and production processes of audiovisual content, including video in mobile applications and on demand systems, are rapidly changing with technological convergence, bringing together previously distinct areas, such as Human Computer Interaction (HCI) and Audience (or Media) Studies [1–3]. This integration demands a theoretical relation that is not currently contemplated by these fields individually [4, 5].

The communicational and methodological model of Audiovisual Design (AD) brings together relevant methods and concepts from those two normally separate disciplines, but with similar profiles. In short, AD has two practical purposes to attain. First, the model provides instruments to analyze audiovisual productions developed for interactive systems, using tools afforded by both HCI and Media Studies fields. Second, the theory provides methods to predict interactions and to propose innovative audiovisual systems, suitable to be used by different profiles described in the model. The main characteristic of this model is planning the interaction, or interactivity, in the process of audio, video and

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 68–81, 2018.
https://doi.org/10.1007/978-3-319-91238-7_6

software production. AD describes four roles to be assumed by individuals and four Lines of Design, to guide and to model the creative development process [4, 5].

The design of interactive computer systems had its initial focus on problems, tasks and functions, but it gradually changed its scope and also incorporated perspective of new possibilities, meanings and emotions [6–8]. Nowadays, individual's behavior is a central part in design projects. Moments of passive fruition, for example, gain relevance when the final object of an interactive system contains audiovisual content. For example, we can mention video on demand systems, accessed both in smart TV sets and mobile devices, with search resources and evaluation tools [5].

A similar phenomenon can be observed from the point of view of audiovisual consumption. The experience using software has become as relevant as the quality of movies, series, songs or online videos. Different digital systems to access video content have one feature in common: interaction using software-based interfaces. Electronic programming guides of digital TV, search tools for specific titles in movie applications, or recommendation tools of videos on social media, are examples in which the viewer's experience mixes an active posture (browsing or searching for information) with moments of passive fruition or enjoyment (video watching). The simple act of choosing and watching an audiovisual show may require the individual to perform different roles, with degrees of greater or lesser interaction or participation. Consequently, a revision of theories and methods supporting the development of these contents becomes relevant, especially in the field of HCI, in which the general notion of user has a limited role. This concept does not contemplate total immersion in the different media, especially those based on audiovisual content. It is necessary, therefore, to understand that this mediatized process incorporates content, technique, aesthetics and meaning, aspects relevant to the complete processes of communication and fruition by individuals [2, 9, 10].

To adequately act in response to this new scenario relative to production and fruition of audiovisual content, where audios, videos and software are integrated into a single product, AD defines four roles: the Audience, the Synthesizer, the Modifier, and the Producer [4, 5]. Each role has elements that permit individuals to extrapolate common behaviors, then becoming the enhanced role 'Player'. In this article, we present and classify the competencies related to each of these roles that an individual (sometimes identified as user, interactor or spectator) can assume. In the model, individuals need to have certain competencies, represented as a set of knowledge, skills, capabilities and behaviors, to be able to act within each role.

Mastering new competencies is a key element to modifying behaviors in certain roles [11–13]. It is a theoretical and conceptual proposal that contributes to the scaling of AD's reach and predictability, a discussion that allows both researchers to evaluate media interactions and producers to design environments and meanings of interaction.

2 Communication Models

Aiming to describe communicative acts and the flow of information between people and technologies involved in communication, the process can be organized in theoretical and conceptual models. In the communication field, different models have been suggested since the beginning of Media Studies, including both transmission and reception technologies, as well as content and meaning of the message [14]. Recently, Jenkins, Ford and Green [2] proposed three simple models to describe the different scenarios of current communication ecosystem: one sender for multiple receivers, as Broadcast; online communication, in which individuals take initiative to search content, as Stickiness; and Spreadable, when content reaches the audience through the action of individuals, usually using digital tools. Although different, in practice there is some complementarity among those models. Often content shared within the propagation logic was produced by television stations (Broadcast), released in a controlled way (Stickiness) and is shared in an uncontrolled way by action of individuals (Spreadable). Digital interfaces, of which Stickiness and Spreadable models depend, have been occupying the space of traditional media and are becoming, themselves, new media elements.

In this way, audiovisual content producers need to incorporate interaction tools as essential elements inherent to production process, which changes the experience of audiovisual fruition. While producers maintain some control over the product, they encourage, at the same time, some audience engagement.

Of course, since many people will only enjoy content at a basic level, it would be a mistake if engagement tools decreased the quality of Broadcast content to passive audiences. Moreover, the interface design should consider not anonymous users, but individuals who, at some point, will use the content as a reference to compose their identity in social networks.

Despite the relevance and appropriateness of those models in Media Studies field, they do not adequately describe the interaction of the individual, nor do they provide tools to analyze, or to create, audiovisual products that depend on software to be watched. Jenkins, Ford and Green [2] describe the communicational processes in information exchange and different forms of media consumption. However, the authors do not contemplate technical-creative process of audiovisual production, where subjectivity of audiovisual narrative interrelates with objectivity of interaction demands. Interface problems, or poor user experience, can compromise the entire audiovisual product.

One example can be represented by *affordances* of interaction interfaces. Affordance refers to possibilities of interaction and use that the system offers or allows for users [8]. This idea, although longstanding in HCI, does not exist in audiovisual production process, including theoretical, conceptual or methodological descriptions in Media Studies. Unlike audiovisual fruition, traditionally considered as a simple process, with no need for any technological domain [14], the development of software depends on potential actions intrinsic to the perception that individuals have about an object or technology, when interacting with the system [5].

2.1 The Audiovisual Design Model

From this scenario, previous studies have identified the need for a theoretical and methodological model that integrates both software development and audiovisual creation processes. As an outcome of the intersections between disciplines of Media Studies and HCI, Audiovisual Design is represented by a graphic model that allows understanding the dynamic flow of audiovisual production, including software and considering different scenarios and roles played by individuals [4].

The horizontal arrows (Fig. 1) indicate the starting points for understanding interaction scenarios. While Media Studies allow inferring behaviors from people in front of audiovisual devices, HCI offers consistent theories and methods for the producer, or software developer, to design interactions and uses. In this way, mechanisms such as user surveys, formative evaluations of interfaces and usability tests help to identify the requirements, demands and needs of the product. In this model, the same individual can play different roles at different times: Audience, Synthesizer, Modifier and Producer. Every role has levels of fruition, with minimum and maximum activities supported in each one. When the individual extrapolates characteristics of each role, but without changing the position, an enhanced role emerges, characterized as Player-Audience, Player-Synthesizer, Player-Modifier and Player-Producer. For analytic purposes, in this article the Player is analyzed separately, to facilitate identification of competencies necessary to act in this enhanced role.

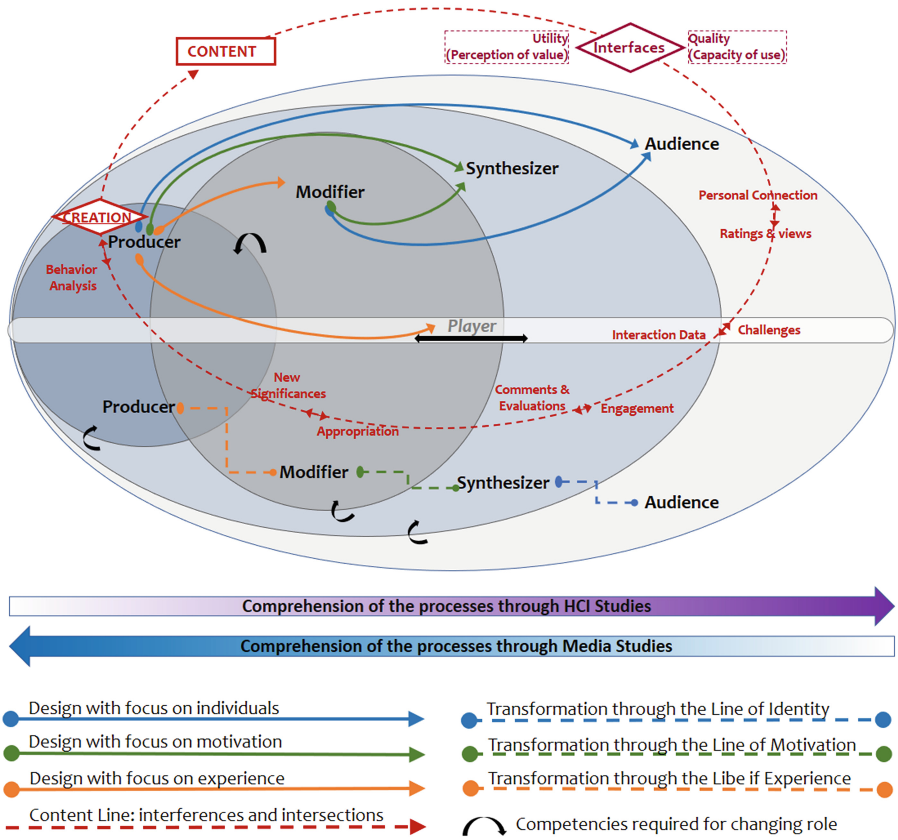


Fig. 1. Flow diagram of the process in the Audiovisual Design (AD) (Source: Audiovisual Design Research Group).

Audience: this is the basic role an individual can assume, denoting low level or absence of interactions in media consumption. This is the passive behavior associated with the Broadcast model [2, 14], whose relationship with digital interfaces are channel tuning, video search and playback, channel subscription, etc. In this way, these people are understood together through audience ratings, which give only a collective view of preferences. The relationship of individuals with content is through personal choices, and remains relatively private.

Synthesizer: the concept was brought in by Jenkins, Ford and Green from an idea developed by Bradley Horowitz [2]. Individuals have skills to compile, to classify, to comment, to recommend and to share audiovisual products they like, often creating an "identity", that is, a profile staged in a network with other people. The role of the Synthesizer involves notion of engagement, or a deep emotional connection that allows individuals to express something about they from the content they share.

Modifier: this role is a subset of Synthesizers, with other competencies and skills: they know how to use software to manipulate and recreate contents with which they identify, going beyond the idea of engagement towards the idea of appropriation. Some examples of Modifier activities are: to improve content, when individuals appropriate content to enhance something they do not like; to remix, when individuals create new productions from the original content; to participate, when individuals send messages and interact with television shows, mainly live, providing information useful for directors and producers. In many cases, participants' messages are screened during the shows.

Producer: one person or a group of people who creates original product, even if inspired by content already existing in other media, autonomously and independently or as part of large media corporations. Although every Producer is a Synthesizer (the distribution of content is part of his competencies), the Modifier's competencies are applied only in situations where an industrially planned workpiece is an adaptation of an existing one, so the overlap between these roles is only partial.

Player: Refers to the individual who fully uses the interaction tools available within each level, being an 'enhancement' within each role. They seek content that will bring challenges and make them, even in isolation, think and do something. This role is identified with "early adopters" or "early users", i.e. people who take the risks of using a new technology and, thus, contribute to its development. As examples, we mention a Producer creating new distribution channels (Player-Producer); a Modifier creating a new product based on different media, such as a cartoon from images other people have taken (Player-Modifier); a Synthesizer acting as a fan and searching posts in many blogs, sites or social media to share them with their own social media contacts (Player-Synthesizer); an individual who looks for easter eggs or other hidden content in off-line media, such as DVD or Blu-ray discs, and engages in its content (Player-Audience).

The relationships inherent in these roles occur through four Lines of Design: (a) Content, with relationships that vary (and add up) according to the role played by the individual: subjective relations at Audience level; engagement in Synthesizer level; appropriation in Modifier level; creation for Producer; and challenges for Player; (b) Identity, which bases the personal relationship with content; (c) Motivation, which leads to engagement; and (d) Experience, which links fruition to advanced use of interactive features, which should provide practical experience with content and interfaces.

3 Competencies Related to the Roles

As we can see in the AD workflow description, the way designers or producers direct their production to different user levels, and how they correlate to each other, depends on articulation of contents, techniques and interfaces that accompany goals defined by the individuals. These objectives correspond directly to the way individuals engage with media (privately or publicly, passively or actively, as spectator or co-producer). To do this, individuals need to have a "competency set" appropriate to technological affordances available for audiovisual product and its interfaces.

In a very general first approach, the main affordances of contemporary media ecosystem are: navigation, search and retrieval possibilities; interactivity; adaptability of content; possibility of being appropriable (both tools and content); openness to new forms of distribution (such as sharing); to be a facilitator for data mining about usage. This list can be expanded inexhaustibly with specific items, since new affordances can emerge at each technological and/or social convergence process.

The competencies, as we understand them here, is a term borrowed from the field of Business Administration and People Management. It refers mainly to behavioral repertoire and intentions demonstrated by an individual to efficiently perform a given task [11, 12]. Nevertheless, this notion is not limited to the workplace, since people must dedicate intentions, actions and behaviors in all interactions in which they have intention to engage. A competency, or the ability to perform a task competently, involves knowledge, skills and capabilities of individuals, ordered from their behaviors and their intentions [11, 15].

We identify the subset of individual's characteristics that allow them to perform an intentional operation as skills. These skills are acquired and developed through the individual's intentions to act or react to a situation, thus shaping their behavior during performance in an engaged manner. Capabilities, on the other hand, are potential skills, present at an unconscious level and not yet developed. Capabilities are inherent in the individual's learning process, intermediate stage between acquisition of knowledge and putting it into operation.

Considering new communication affordances created within contemporary media ecosystem, it is natural for individuals to demonstrate new competencies, or new sets of skills, knowledge, capabilities and behaviors, or even modify those already ingrained both psychologically and culturally. We present in this paper a panoramic description of competencies related to each user role within AD, however, some observations are necessary.

Firstly, the elements that belong to specific competencies may have socioeconomic, cultural, technical or technological characteristics. Within AD, we consider them related to: physical-financial conditions to access the content; fruition, appropriation and cultural use of messages transmitted in a communicative process; the way this fruition, appropriation, and use occur; and knowledge of technologies necessary to apply them to a particular use. We do not offer a description of socioeconomic or cultural characteristics in this article; we believe that this requires a greater degree of specificity, which must necessarily be related to each context (geographic, political and economic).

Secondly, a competency only becomes effective when skills, knowledge, and behaviors activate the affordances from available communication technology. Thus, it is not just a theoretical appropriation or notion of the operation that ensures a person has actually acquired the skills necessary to be classified within an AD role. To do so, the individual must apply the respective competencies in the process of media fruition. Only then it is possible to state, in fact, the individual has moved in between roles. This proviso is necessary because we reinforce that there is an opening for progression or regression of each individual in between roles, which is always confirmed by the Line of Identity: the same person can use all tools at his disposal to engage more fully with a product or a workpiece, which touches him deeply, and can be a passive spectator of another, with which it relates only for the satisfaction of entertainment.

A third observation refers to actions the Producer can take to enable acquisition of knowledge and skills by individuals, which may sustain behaviors that apply the necessary skills to change from one role to another. It is an intrinsic part of designing complex interactive audiovisual pieces. Affordances need to be adaptable and learnable by target audience, in an intuitive and harmonious way with general narrative universe. In other words, if the Producer wants (controllable) groups of Synthesizers and Modifiers in a participatory and collaborative logic, it is very likely that they will have to enable people to learn how to develop those roles.

3.1 Competencies Framework

In order to facilitate the understanding of how competencies are applied to different roles, we present a list divided by role individuals can act in the AD model. We emphasize that not always will all skills be necessary for a person to be identified acting within a role. It depends on production characteristics and how individuals interact. However, at least some of the characteristics should be present. In addition, since the logic of Audiovisual Design workflow is based on the notion of "sets", where one group is contained in another, competencies accumulate from one level to another.

This description is not exhaustive of the competencies of each role, since they may vary according to proposed work. The items described initially come from profiles and skills traditionally identified by HCI and Media Studies from authors such as: [1, 6, 10, 16–31]. A detailed review was conducted looking for competencies necessary to act minimally in each role, and how to extrapolate this actuation, both to change among roles and to become an enhanced role, as analyzed in Player's competencies description. With the combination of these elements proposed by AD, common attributes were established to individuals and explained according to their level of interaction. What we intend, with this framework, is to open a new discussion about elements currently involved in design of complex audiovisual products. We recognize that these observations are, at least, fairly generic to describe all possible processes.

As summarized in Table 1, the first level of roles of individual clearly requires the simplest and most easily appropriable competencies. For the viewer who only watch audiovisual content in Broadcast model, cultural and linguistic competencies for understanding and interpretation of the message are basically enough. However, when we think about the increasing number of individuals using alternative fruition forms to the Broadcast, but still for passive consumption - such as legal or illegal download, on demand, fragments, and so on - skills accumulate.

Knowing the functioning and limitations of internet itself is still a limiting factor in many countries, which is being rapidly overcome as we observe the growing number of population with access to computer and to network connections. After surpassed this stage, the competencies listed here make possible two central phenomena: individual's identification with the produced content (theme, characters, narrative, ambiences, etc.), and feedback to the Producer, especially in quantitative formats (qualitative feedback is possible when the correct tools are available).

In the second role, (Synthesizer), the feedback becomes more qualitative. The Producer should observe more closely the ways in which people have engaged with their product. This analysis can direct choices for workpiece in production. Among these competencies, we emphasized those related to social media. In an attempt to create their network identities, Synthesizers are part of communities of mutual interests and act as representations of themselves, especially designed for particular situations. Thus, they become poles within social networks articulated by digital media. These competencies also make evident one's condition as a fan, whose discussions revolve around their active participation in available content activities.

Regarding the role of the Modifier, it should be noted that appropriation of technology is fundamental, especially when Producers do not make available tools for modification. Despite that, the appropriation of content that occurs at this level is a cultural appropriation of some parts or of the whole workpiece or technology. This appropriation includes generating some meaning, sense or a different discourse from that originally manifested by the Producer. This process can be observed, for example, when an element in an audiovisual piece (a song, an emblematic scene in a film, a character in a fictional narrative) becomes both the link of a dispersed group of people and a symbol of a cause or a fight.

The Modifier gets elements from original context, modifies them, or recreates parts to transform them in a representative idea, or even in an ideology. Such means of expression represent a fundamental source of information for the Producer who, knowing how to read these signs, can correct future works or adapt ongoing demands. From a sociological point of view, it can represent a very aggressive capacity for expression, especially when the group appropriating content is a minority that uses a popular resource to make itself heard. Anyway, the Modifier is often able to create his own social network, becoming an important node of content diffusion.

The activities of the Players should be widely explored by Producers interested in knowing fruition possibilities of audiovisual productions. Players represent, however, a small part of total audience. They are best placed in the projects elaborated within the logic of achievements and rewards.

Finally, the Producers' competencies are more complex, divided mainly between the field of technology and the ability to interpret the users' demands. It remains to be pointed out that, except for a few authorial or amateur projects, an audiovisual production usually involves a group of people because even the director usually is not able to control all steps. In this way, when we refer to the Producer we are talking about each person who works at a project, with their own competencies within the area of expertise. This is common in large commercial products (e.g. writers, system

Table 1. Competencies per individual role in Audiovisual Design

Role	Competencies and Skills
Audience	**Core Competencies:** • Consumption of audiovisual content for personal satisfaction **Technical/Technological Skills** • Master basic uses of telecommunications services (Internet access, web browsing, etc.) **Behaviors, Knowledge and Capabilities** • Be able to organize an audience logic as a personal schedule to enjoy content even when fragments of the same story are spread across different media • Know how to control audiovisual consumption, by simple interactions such as forward, save to see later, etc. • Know how to create content lists, such as bookmarks and playlists; • Feel free to respond to simple inquiries from producers, or use the tools available ("liking" and "commenting" on interactions that can change the narrative flow)
Synthesizer	**Core Competencies:** • Engagement in social groups • Influence people within social circle • High communicability through interfaces • Motivation to circulate content **Technical/Technological Skills** • Understand and use social media extensively, articulating narrow or broad networks • Use smartphones for conversations, to exchange information, and content sharing **Behaviors, Knowledge and Capabilities** • Use interactive toolsets provided by Producers to distribute promotional content on social media, whether customized (such as photos edited from a website and linked to the product) or not (such as music playlists on Spotify or a teaser on YouTube) • Know how to create and to participate in discussion groups, forums and other forms of socialization, based essentially on digital networks, to exchange information about products, comments and evaluations • Feel free to select parts of works made available by the producers, and share them • Know how to create his own network to share playlists and comments while receiving and processing the message. • Subscribe to networks related to enjoyable audiovisual products, official or otherwise, to promote the product (or related brands) through spontaneous disclosure • Participate or organize campaigns to promote the continuity of the workpiece, either through pressure on producers or collective funding (crowdfunding)
Modifier	**Core Competencies:** • Creativity to recreate/modify content • Stand their own opinion/point of view • Influence people inside and outside their own social circle

(*continued*)

Table 1. (*continued*)

Role	Competencies and Skills
	Technical/Technological Skills • Master image and sound editing software • Edit using publishing tools, with or without editing knowledge, such as XML or HTML • Use advanced digital communication tools, such as blogs and personal websites • Use distribution services like YouTube and Sound Cloud, customized with one's own identity **Behaviors, Knowledge and Capabilities** • Use advanced tools and material for content editing, when available from Producers, or demand these tools from them • Be able to copy and edit excerpts of content, in a way previously authorized or not by the producer • Interact in fan groups to exchange derivative content. • Use digital content dissemination techniques to reach a recurring audience and target his personal channel in different media (blogs, social media, websites, etc.) • Know how to use (or create) classification systems - such as "tags" - that allow propagation of their content through a wider network than originally linked to the Modifier individually • Be able to explain with derived productions, criticisms, suggestions and praise to the Producers • Know other elements of the culture that can be related, to create derivative products
Producer	**Core Competencies:** • Creativity to produce original content • High communicability through interfaces and content • Empathy to know audience needs and demands **Technical/Technological Skills** • Know the different tools used throughout the interactive content production process • As part of a team, master the technology that applies to their own function in production process • Know logics and processes of different design methods, especially user centered approaches • Have access to digital data reading tools capable of feeding data mining procedures **Behaviors, Knowledge and Capabilities** • Know how to interpret data from different sources (audience ratings, view and visit metrics, group searches, focus groups, algorithms, direct feedback from receivers, etc.), and performing data mining • Know how to create personas and scenarios that allow application of collected data in product development • Create interactive, user-friendly interfaces for each role, incorporating needs and expectations of individuals

(*continued*)

Table 1. (*continued*)

Role	Competencies and Skills
	• Anticipate a scope of uses to be given to the content, so the set of individuals in a role finds the tools necessary to perform the role • Deep knowledge of interactive interfaces that are being offered to incorporate instructions or other tutorial tools for users when necessary • Master content distribution networks to create indexable fragments that facilitate their propagation • Be open to incorporation of innovations to the product suggested by different roles of audience, especially those evinced by the Modifiers
Player	**Core Competencies:** • Accept challenges • Self-learning and exploration of new objects and technologies **Technical/Technological Skills** • Master technologies through which the producer distributed the workpiece (from connected TV to videogame consoles) • Know digital system codes, when necessary, to interpret certain functions of the product **Behaviors, Knowledge and Capabilities** • Know how to adapt available technical conditions to extend the proposed experience, within competencies of his key roles (such as Audience, Synthesizer or Modifier) • Handle as a challenge and use all the interactive tools provided by the Producer, within key role it includes • Cognitively interpret fragments of history, clues, and small rewards, for deeper involvement with the narrative • Pay adequate attention during fruition, to explore the parts of the product as much as possible, and to discover hidden content (such as hidden scenes or frames, undisclosed websites, other forms of easter-eggs) • Be able to collect achievements over time and group them to return to the producer as feedback • Engage in networks of Players, through social media, to share achievements, tips and questions

developers, editors, graphic designers). Even if a person cannot master all technologies involved in a production, it is important to be aware that the workflow incorporates these technologies. User interfaces depend on a sense of continuity of actions, and sudden ruptures caused by the simple articulation of independent products can compromise the whole fruition process.

The success of this implementation depends, above all, on the incorporation into the system of responses to individuals' necessities arisen from the set of observed uses, considering content characteristics and fruition process. Data to be read by Producers can be generated after the distribution of audiovisual product; in other words, after the audience has watched or heard the content, and used the interaction tools. This is the audience data and feedbacks mentioned above. In this case, Producers' actions are limited to future works or, in the case of long-running productions, to slow transformations as the narrative unfolds. Furthermore, it is possible to use the technique of personas creation, with some previous data and with the development of scenarios, to predict some reactions. In this case, it is of great value to read the data obtained from the application of digital tools. Access and understand the data is an important tool (although not fundamental) for the Producer.

4 Conclusion

This article described core competencies needed by an individual to occupy one of the roles indicated in the Audiovisual Design model. The competences listed here may not be the only ones in certain contexts of fruition of an audiovisual workpiece. They are, however, ample descriptions that correspond to a significant part of situations which require an action or reaction by user. Each competency involves knowledge about the object of interaction, capacities (potentialities) for use of these objects, skills to engage in a particular action and behaviors in face of situations that arise during audiovisual consumption.

These competencies are relevant to the contemporary media ecosystem, which, when merging different means in the same environment, dominated by digital technologies, give rise to new affordances related to use of interfaces and computers. In this sense, the HCI field contributes to analyze situations where media consumption occurs, pointing out what these new affordances are and how audiovisual production should be organized. Hence those affordances can be incorporated in content and interface planning of interaction, as a prediction of actions taken by individuals. In other words, HCI's problem-solving chain allows a complex use of different platforms and media by indicating potential properties within each audiovisual content distribution channel.

Such confluence of two distinct fields (Media Studies and HCI) demonstrates the need to review school core curriculum, especially in the field of audiovisual production, which today does not contemplate important topics for the activities of content producers for interactive systems.

Although the methodology of Audiovisual Design incorporates technologies available today, we must keep our mind open for assimilation of new technological or social dynamics that interfere in the proposed flow. We also understand that financial limitations may prevent the application of this methodology in its most complete version: on the production side, by the costs that may be high depending on the type of resources it proposes to use; by the audience, by the imbalance in the access to the Internet and digital tools that allow interaction at the levels foreseen here.

However, Audiovisual Design remains an appropriate methodological set, allowing to ponder over each phase of planning as an isolated process, but connected through the Lines of Design. In addition, it adds questions relating to training professionals who will produce content since the relations among system, content and individual are becoming increasingly close. Finally, it is worthwhile to emphasize the application of AD as a theoretical tool for evaluation and understanding of media usage scenarios. AD methodology encourages discussion about concrete aspects of audio and video systems use and theoretical aspects such as layers of identity, perception of value and sense people attribute during the enjoyment of the content.

Acknowledgement. Daniel Gambaro recognizes and thanks São Paulo Estate Research Foundation - FAPESP, for the grants received and to which this and other researches are connected: Processes 2015/20430-5 and 2016/17884-7.

References

1. Millerson, G., Owens, J.: Video Production Handbook, 4th edn. Focal Press, Burlington (2008)
2. Jenkins, H., Ford, S., Green, J.: Cultura da Conexão: Criando Valor e Significado Por Meio da Mídia Propagável. Aleph, São Paulo (2014)
3. Meixner, B.: Hypervideos and interactive multimedia presentations. ACM Comput. Surv. **50**, 1–34 (2017). https://doi.org/10.1145/3038925
4. Becker, V., Gambaro, D., Ramos, T.S.: Audiovisual design and the convergence Between HCI and Audience Studies. In: Kurosu, M. (ed.) HCI 2017. LNCS, vol. 10271, pp. 3–22. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58071-5_1
5. Becker, V., Gambaro, D., Ramos, T.S., Toscano, R.M.: The Development of media affordances in the audiovisual design. In: Abreu, J.F. de, Guerrero, M.J.A., Almeida, P., Silva, T. (eds.) Proceedings of the 6th Iberoamerican Conference on Applications and Usability of Interactive TV - jAUTI 2017. pp. 91–103. UA Editora, Aveiro, Portugal (2017)
6. Preece, J., Rogers, T., Sharp, H.: Design de Interação – Além da Interação Homem-computador. Bookman, Porto Alegre (2013)
7. Oulasvirta, A., Hornbæk, K.: HCI research as problem-solving. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI 2016, pp. 4758–4769. ACM Press, New York (2016)
8. Kapteliinin, V.: Affordances and Design. The Interaction Design Foundation (2014)
9. Richards, R.: Users, interactivity and generation. New Media Soc. **8**, 531–550 (2006). https://doi.org/10.1177/1461444806064485
10. Murray, J.H.: Inventing the Medium: Principles of Interaction Design as a Cultural Practice. MIT Press, Cambridge (2011)
11. Boyatzis, R.E.: Competencies in the 21st century. J. Manag. Dev. **27**, 5–12 (2008). https://doi.org/10.1108/02621710810840730
12. Woodruffe, C.: What is meant by a competency? Leadersh. Organ. Dev. J. **14**, 29–36 (1993). https://doi.org/10.1108/eb053651
13. Sanders, J., Albers, P.: Multimodal literacies: an introduction. In: Literacies, The Artes, and Multimodality, pp. 1–25 (2010)
14. Sousa, J.P.: Elementos de Teoria e Pesquisa da Comunicação e dos Media. Edições Universidade Fernando Pessoa, Porto, Portugal (PT) (2006)
15. Yonn, C.Y.: The effect factors of end-user task performance in a business environment: Focusing on computing competency. Comput. Human Behav. **25**, 1207–1212 (2009)
16. Gosciola, V.: Roteiro para as novas mídias - Do cinema às mídias interativas. Senac - SP, São Paulo (2008)
17. Block, B.: A narrativa visual: criando a estrutura visual para cinema TV e mídias digitais. Elsevier, São Paulo (2010)
18. Manovich, L.: Software takes Command : Extending the Language of New Media. Bloomsbury Academic (2013)
19. Aumont, J., Marie, M.: Dicionário teórico e crítico de cinema., campinas, SP (2003)
20. Costa, A.: Compreender o cinema., São Carlos, SP (2003)
21. Cirino, N.N.: Cinema Interativo: Problematizações de linguagem e roteirização (2012)

22. Toscano, R., Becker, V.: Potenciais narrativos de interatividade para vídeos de altíssimas resoluções no campo da videocolaboração (2015)
23. Toscano, R., Becker, V., Ferreira, L., Coutinho, S., Burgos, L.: Arquitetura de design colaborativo para imersão temporal e espacial em vídeos de altíssimas resoluções e HFR. In: O futuro da videocolaboração: perspectivas, pp. 13–53. Sociedade Brasileira de Computação (SBC), Porto Alegre, RS (2017)
24. Bronsztein, K.P., Cirino, N.: A metamorfose do virtual: conceito e experiências de fruição. Rev. Eco Pós. Arte, Tecn, pp. 148–160 (2015)
25. Bieging, P., Busarello, R.I.: Experiências de consumo contemporâneo pesquisas sobre mídia e convergência (2013)
26. Nogueira, L.: Narrativas fílmicas e videojogos. LabCom, Beira interior (2008)
27. Nogueira, L.: Cinema múltiplo, figuras, temas, estilos, dispositivos (2015)
28. Faro, P.: Cinema, vídeo e videoclipe: relações e narrativas híbridas. Rumores - Rev. Comun. Ling. e Mídias, 4 (2010)
29. Ala-Mutka, K.: Mapping digital competence: towards a conceptual understanding. Inst. Prospect. Technol. Stud. 60 (2011)
30. Newton, K., Soukup, K.: The Storyteller's Guide to the Virtual Reality Audience. https://goo.gl/Byy8MX
31. Maike, V.R.M.L., de Sousa Britto Neto, L., Goldenstein, S.K., Baranauskas, M.C.C.: Heuristics for NUI revisited and put into practice. In: Kurosu, M. (ed.) HCI 2015. LNCS, vol. 9170, pp. 317–328. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20916-6_30

Check for updates

How to Bring the 'Myth' of Cultural Awareness into Enterprise Software – Challenges and Approaches

Ulla Geisel(✉)

SAP SE, SAP Design GD UI Concepts, Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany
ulla.geisel@sap.com

Abstract. An increasing number of companies see a strong business need to get deeper insights into the cultural settings of their consumers as they are coming to the realization that meanings, behavior and activities of people might differ from region to region, and that this can have an impact on the acceptance and the usage of their products. Sometimes, however, the impact culture can have, is still seen as a myth. This paper focuses on how we at SAP Design Global Design UI Concepts tackle the challenge to bring cultural awareness to the teams.

Keywords: Intercultural User Experience · Intercultural design
Cultural awareness · Culture · Design anthropology

1 Introduction

Imagine if in ancient times, a craftsman from Walldorf in Germany was asked by Odysseus and the Greeks to design the Trojan Horse. Most likely this craftsman would not have had any idea about the concept or meaning of a horse in this specific context, about socio-cultural habits, the environmental situation or the equipment needed. Perhaps he would have chosen another vehicle to hide and transport the warriors, as he may not have known that the horse was revered as a symbol for Athena and that because of this, the Trojan people would pull it into their city in front of the temple of Athena. How would he know if the material he was planning to use to create the horse would be available in that region, or if the horse would be the right size to fit in all the Greek warriors and their weapons, or if the horse would have been too big to fit through the city gate? What if the mechanism to open the horse from inside couldn't be operated by the warriors because they were used to other technologies?

We could continue with more illustrations to make it obvious that this thinking doesn't only apply to this myth. When designing, or creating an artefact, a product or a

This paper represents the view of the anthropological perspective within SAP Design - Global Design UI Concepts for SAP Fiori design language.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 82–90, 2018.
https://doi.org/10.1007/978-3-319-91238-7_7

service, it makes a real difference to have specific insights into cultural and environmental situations and the context of when they will be used. It can be the difference between success or failure.

This simplified example often helps to make designers, enterprise product owners, software developers, and others, understand why there is a good reason to get deeper insights into cultural and intercultural aspects. They can more easily recognize that there is a value-add when taking these aspects into account at the beginning. An increasing number of companies are investing in this area, as they do not only see the competitive advantage but also the need to understand the cultural context of their users in the various regions worldwide. Neglecting them can cause severe damage.

The Designers in our User Experience concepts team, who provide the SAP Fiori design language for business applications that are used around the world, already aim to have the cultural and intercultural aspects in mind when working on specifications for design concepts and floorplans. These will then be used by application designers and development to build applications that fit best the peoples' actual experience.

The challenge is not only to stress the benefits when intercultural design is implemented correctly versus costs when it is done wrongly or not at all, but seed the importance of cultural and intercultural aspects within the company so that it is not seen as a 'myth' but an essential aspect to deliver successful user experience. We are building a bridge between the technological focus and cultural phenomena and bring awareness and knowledge to the teams so that they can integrate it appropriately in their work, be it design concepts or enterprise software applications.

Before detailing out our approach however, first we must be clear on the underlying assumptions.

2 Underlying Assumptions

This section focuses on the underlying assumptions for our approach, explaining the notion of culture and User Experience design, as well as covering how we view intercultural User Experience, Users and Business Software.

2.1 Culture and User Experience Design

Culture as we understand it follows Geertz' (1973) concept of culture as "webs of significance that man himself has spun" and their meanings (Geertz 1973). The analysis of these webs is generative; it is concerned with the ways people find meaning in the world. Therefore, culture embraces meanings, behaviors and practices that groups of people develop and share over time (Sun 2012). We acknowledge that we also have several aspects of local cultures to consider when trying to get an understanding of the meanings, practices and behavior: these are socio-cultural aspects (e.g. age, organization), individual factors (e.g. education, values) as well as the different ways of living and acting.

Although culture takes a vivid role in the design of products and services, the tendency is often to remain on the surface of the famous iceberg invented by Hoft (1995) which concentrates on the visible tip, and on issues such as e.g. translation, color, layout, aesthetics, punctuation. The submerged part of the iceberg (per Hoft 90% of it) consists of unspoken and unconscious rules and sets of meaning that are often not considered by product and service designers and developers. The result is a disconnect between meaning and actions (Sun 2012). As Sun (2012) points out concentrating only on the tip of the iceberg can result in Dos and Don'ts and lead to guidelines, advice, standards etc. that often remain static and only represent the dominant culture of a nation-state, which can lead to stereotyping. The dynamics of culture and its ever-changing aspect are neglected. People might follow these guidelines and rules without even realizing that they are no longer valid for example, the postulation that in China the color white is only used in the context of death or grief. Chinese brides in some regions nowadays wear white wedding dresses, sometimes in addition to the traditional red ones as red is still seen as a color describing luck. Through western influence, young people, especially those in urban areas, have come to prefer this more western style.

These types of examples show us that culture is not a static concept, that cultures evolve and change over time and that there is a constant influence on local cultures. The hidden part of the iceberg needs to be considered to understand actions performed by people and the meaning they ascribe to it.

Bell and Dourish (2011) outline three key problems when we try to detect 'cultural differences' which have an impact on the use of technologies, and thus the design for different cultural settings. Firstly, where can the boundaries between cultural regions be drawn? Is it on nation-state level, religion, language, others? What do we do about the traffic of culture across boundaries? Secondly, how can we understand culture in an increasingly interconnected world? And thirdly, we are all pulling from many cultural systems to understand our worlds.

Such problems led Bell and Dourish to a different, generative view on culture: "[…] we encounter the world through cultural lenses, which bring it into focus in particular ways while also rendering it meaningful and accountable to us. These lenses frame what we see, and how we see and understand it." (Bell and Dourish 2011).

We prefer these concepts of active and performative cultural practice over those that use cultural dimensions, like the dimensions Hofstede defined based on survey results collected within IBM in 72 countries in the 1970's (Hofstede 2001). Hofstede's five defined dimensions described by more static values are widely used, not only for intercultural communication but also in the web and software design area (Sun 2012). Our concern here is that the dimensions are bound to nation-states and their definition is based on surveys conducted in one company, thus showing one corporate culture. The participants were all chosen from one corporate area; only their attitudes in the corporate context were explored so that the wider socio-cultural perspective was missed.

The concepts of culture laid out above use a wider and much more flexible concept and approach than the value-oriented cultural dimensions; they incorporate meaning and activities into the 'role' of our users.

The task of User Experience design is to create user interfaces that mirror parts of the 'web of significance' and do this as closely as possible, to ensure that people have the best possible experience with the technology.

We therefore argue, that the daily life we encounter through 'cultural lenses' and that we experience should be aligned with the User Experience and that the User Interface should include cultural acceptable actions and meanings. Technology and its usage per se is a cultural production.

2.2 The User

In software design, we work with 'Personas' that reflect the 'user' of a system. To derive a 'Persona' usually one to a few users are observed and interviewed to see how they accomplish a specific task. From that learning a 'typical' user is created.

The 'Persona' comprises of these findings and serves as the 'ideal' of a user who performs a specific business role which would then guide designers and developers building applications. The 'Persona' can be a valuable basis to get a first understanding of the users in their specific working context, and see how they perform certain activities with the software, what hinders their work and what is helpful. It is, however, only a very generic artifact that is used for products used across multiple regions and thus misses the specific cultural settings of the people using our software.

Within intercultural design we are specifically interested in the 'how' and 'why', and not only the 'what' of when people do things, or use technology and applications. Details help to understand the facts, but they can't explain the reason and rationale behind something (Ladner 2012).

We believe it our task to not only 'see' the user in their work context, but include their 'web of significance'.

2.3 Intercultural User Experience

We understand that to implement intercultural User Experience we must create a design language for enterprise software that considers different meanings, behaviors, practices, languages and how they are used in communication situations, as well as specific economic standing. At the same time, we want to ensure Usability and User Experience across intercultural boundaries.

We prefer to use the term 'intercultural' instead of 'cross-cultural', as we believe 'intercultural' reflects a collaborative joint approach of people from different cultural backgrounds, rather than the 'cross-comparing' and potential building of an average.

The term intercultural implies that it is all about shared experiences and interactions of people with and between different cultural backgrounds. It requires immersion and personal engagement from design and development. Exchange between teams from different countries and regions who are designing and developing enterprise software is of utmost importance to us.

We work in close collaboration with people from different countries and regions as we need to understand how people interpret and use products and services based on their cultural backgrounds as technology is always contextualized within those. User Experience design should thereby not only consider peoples' behavior and thinking but take the technical and infrastructural environment as well as the business context into account.

2.4 Enterprise Software

Besides understanding the benefit cultural and intercultural awareness brings, we need to consider, that we earn our money by building business software that is used around the globe.

One constraint therefore is profitability. As in every business there needs to be a balance of what we invest versus what we get out. Building software that can be used in different cultures and domains without too many necessary specific adaptations is therefore preferred over building thousands of individual custom solutions that are not only expensive to implement but also to enhance and maintain.

On the other hand, many of our customers need to work worldwide, which means that people working in different countries and with different backgrounds still need to work collaboratively with our software and within a corporate setting in order to communicate with each other. Individual software solutions that perfectly support one individual, might therefore hinder collaborative work within global organizations.

With that in mind, we still do not strive for a one-size-fits all approach that builds a cross-cultural average, but seek for a way to find an intercultural basis (culturally agnostic foundation), from which we can anticipate and provide options to easily adapt our software to specific cultural needs.

Today, we already try to have a good definition for the 'tip of the iceberg' and consider how to reach what lies beneath the surface. This is provided by a culturally agnostic basis derived from intercultural work, which allows flexibility for cultural specific changes on top.

As such, the SAP Fiori design language supports:

- Responsiveness via the adaptability of the design controls, patterns and floorplans to the size of the screen and different devices
- Touch/keyboard support
- Over 30 different languages
- Different calendar types (e.g. Gregorian, Islamic, Japanese)
- Right-to-Left languages and the ability to change the reading direction of the screen
- Themability to adapt the appearance of the UI like colors and fonts to the needs of the companies and their users
- Two themes for visually impaired people (high contrast white, high contrast black)
- Personalization of tables
- Key users can change the sequence or labels of controls via extensibility

We are in continuous dialog with development teams on these aspects. In addition, this foundation can be enhanced and complemented by new aspects derived out of the anthropological design work.

With this framework, we can limit the number of individual custom solutions our customers might still need. Where they are so specific that we cannot cover them with existing means or with adaption, then the decision for custom solutions is usually taken, especially if the benefit of supporting a specific use case and its cultural aspects outweigh the higher investment.

3 Approach

When starting as an anthropologist in a software company you might have the idea of doing your own fieldwork, either for a specific topic sponsored by one or several stakeholders, or to come up with new ideas to shape (or even create) products based on your research data. You might want to go outside to make the point that it is not only workplaces that are interesting to design and development but you need to see how people behave in their cultural settings. You will want to bring the notion of culture and what you have learned in the field back to your teams.

This is not always the case. You might learn that your task is seen to provide Design and Development teams with a list of cultural or intercultural 'do's and don'ts' that will help them to build their applications.

To show clear value to our teams, and take them beyond their assumption that a Do & Don't list is sufficient, we decided on the approach detailed out below. We deeply invest in informing and engaging with people in the organization.

As Roberts (2010) points out it is a long process and a long conversation for anthropologists in the corporate world to transfer their knowledge or their findings to the many stakeholders who are differentially located from a geographical, hierarchal and functional perspective. It is about establishing an ethnographic or anthropological sensibility and a distinctive point of view within the organization. It is about changing perception, language and challenging assumptions within the organization. It is about enhancing appreciation for what Hasbrouck (2018) calls 'ethnographic thinking' and understanding what this can bring in addition to the solution-focused research work. Ethnographic thinking as described by Hasbrouck "isn't simply a front-end research component positioned to feed the design process, but an ongoing inquiry that helps shape design solutions while simultaneously observing and interpreting the evolution, or pulse, of human interactions – always asking "why?" [sic]" (Hasbrouck 2018).

As suggested by Roberts (2018) 'embedded' anthropologists need to learn "to account for and express their contributions in ways that combine soft and hard outcomes" (Roberts 2010). Soft outcomes might include education, awareness, etc. Hard outcomes might be product ideas or influence on product and market strategy, etc. We concentrate on the soft outcomes first which include education, production and amplification of awareness and sensibility.

3.1 Spreading the Word Within the Company – Information and Education

Information and Storytelling. As stated previously we don't want to provide the teams with lists of do's and don'ts or checklists because we see this as problematic. These lists often derive out of stereotypes, personal experiences, informal observations, surveys; they might even contain information like "Germans are punctual whereas Italians are late" which is already a stereotype. Do's and don'ts tend to stay on the surface, to stereotype, are often built on nation-state level and neglect nuances and dynamics of ever-changing culture (Sun 2012). Such lists are often used in intercultural communication and business; in our understanding, they might not be of help for designers and developers working on products and services.

We want to provide our teams with the information that encompasses our own cultural settings and those more alien to us, and raise sensitivity for the topic. To make this explicit where possible, we tell this information in stories. Stories describe events, explain the meanings that lie behind actions and set them into a larger context. A good story combines causes and effects in a narrative way which makes people more likely to remember the knowledge conveyed versus a list of pure facts.

People should not only get the facts but the interpretations of those facts and what these could mean for us, as responsible designers of the SAP Fiori design language, and for development.

The current basis for the information is either secondary research or examples from other companies that already have anthropological and ethnographic experiences. In addition, we use findings from in-house teams who are already aware of the importance of cultural and intercultural settings. We are building on this knowledge-base as we get more examples shared with us, and that can be used to successively build a veritable treasure trove of such stories.

This information is shared on an internal site for designers and developers who are using the SAP Fiori design language. In addition, we provide a collaboration platform that is open to the whole company to encourage exchange on cultural and intercultural topics. We believe that building-up a community of anthropologically sensitive and interculturally concerned people plays a vital role in strengthening and reinforcing the value of our approach.

Building a Community. Within the company, we connect, engage, and exchange with people globally to work on intercultural topics and learn about meaning and activities in various cultural settings. They come from several areas like design, globalization services, development and consulting.

Our current approach is to learn, educate and produce awareness internally. Currently we have to learn about cultural settings ‘secondhand’, through people who guide us. Building a network of proxies is therefore a good means to get to the good information. We either find proxies as informants within our local area, especially those who have just arrived from other countries and still have a good sense of the cultural meanings and activities of their original cultural setting, or else from people within other regions. Doing this, we consider their working environment within a global company and thus the corporate influences as well as other socio-cultural and individual aspects (e.g. education, technology and design knowledge) that might not reflect the broader view we strive for.

We know that colleagues from other countries who live longer in our local area might change as they are more exposed to the influences and adaptations different to their original locality and mix up their former and the new cultural meanings and behaviors; they themselves state that it would be better to find a proxy in their homeland and volunteer to connect us with them. Usually from such contacts more and more arise.

In addition, we establish an external network with anthropologists and researchers in academia and business to get a much broader view on cultural settings outside our corporate environment.

So, even without running field research we can use such local and remote resources to get an understanding of local cultural settings as well as meanings and activities of people.

The tool we intend to use for exchanges within the community is an internal collaboration platform where our findings will be exchanged, can be discussed and/or enriched, and questions can be raised.

Hold Presentations and Write Blogs. Diffusing such a topic within a huge company, with lots of areas and regions worldwide was very challenging, so we began with presentations to introduce the general aspects of the topic within our design community. This proved to be a good starting point. As all our application designers are usually assigned to a specific business area, they are already able to multiply what they learn within their teams. By disseminating this high-level knowledge, we were then contacted by teams or internal forums to talk about intercultural aspects or we pro-actively offered to talk about the topic in sessions.

Slide presentations are a very common way that information is shared within a company, often presented to smaller groups, sometimes larger forums. As distribution channel slides are not the best medium we need to think of other tools and ideas to spread the word instead of just using slides, information sharing areas on internal sites for the SAP Fiori design language or the collaboration tool. The idea that Dautcher and Griffin (2010) presented at EPIC 2010 to create videos out of slides to make the stories more appealing, interesting and consumable to a broader audience will also be used within our team.

Furthermore, internal and external blogs will be used as distribution channels, to expand the network and exchange on the topic.

Give Trainings. We were asked by a colleague from Globalization Services to integrate the intercultural aspects into a training we developed on UX basics for non-designers. To meet this need we created a one-day training on the intercultural aspects of User Experience.

This basic and interactive training not only gives insights into intercultural design in general using examples from both our own and other companies, but has a broader approach in that we start with our understanding of what 'culture' is enriched by, with stories, with exercises for stereotyping generally and in communication in research projects. If we get this wrong it can lead to huge misunderstandings and thus result in either issues between teams, company and customer, or poorly designed applications or missing functionalities.

To make it an interactive training the participants get the research results of a project in a South American region, the SAP Fiori design concepts and floorplans the colleagues used as basis and then have to adapt these concepts and floorplans to the cultural meanings and activities that were analyzed based on the research data.

The training started last year and has been held five times in Brazil and Germany. Feedback, ideas and stories shared in the training sessions by participants coming from their own cultural setting continuously find their way into the training material.

We plan to make this training an official part of the companies learning platform. To do so we will also start a 'train-the-trainer' series. This not only helps to scale but also to get more insights into further intercultural aspects that will enrich the training and the community platform even more.

4 Conclusion

Bringing cultural awareness to Enterprise Software designers, developers, product managers and others can be a challenging endeavor. It can be done in several ways: through e.g. analysis of data gathered through field research tackling specific issues, or research concerning a broader view on meaning and activities of people in specific cultural contexts, or other forms of ethnographic research that might have an impact on products and services of the company; this is probably the preferred way of anthropologists working in the corporate environment to show their value to the organizations.

In addition, it can also be done by engaging with people in the company through long conversations and education to bring the notion and impact of culture, 'ethnographic thinking' and cultural sensibility to the teams working globally. This is the approach we follow as a first step within the SAP Design Global Design UI Concepts organization for our SAP Fiori design language.

By providing and evangelizing information on the impact of culture on products and services, educating on what 'ethnographic thinking' is, by establishing a community internally and an external network, creating exchange platforms, holding trainings, giving presentations and figuring out various ways and output means for circulating findings we bring the cultural awareness to the teams within our company so that it is no longer seen as a myth.

References

Bell, G., Dourish, P.: Divining a Digital Future: Mess and Mythology in Ubiquitous Computing. Massachusetts Institute of Technology, Cambridge (2011)

Dautcher J., Griffin, M.: Ethnography, storytelling and cartography of knowledge in a global organization: how a minor change in research design influenced the way our team see, and is seen by our organization. In: EPIC 2010 Proceedings (2010)

Geertz, C.: The Interpretation of Cultures. Basic Books Inc., New York (1973)

Hasbrouck, J.: Ethnographic Thinking. From Method to Mindset. Routledge, New York (2018)

Hofstede, G.: Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations, 2nd edn. SAGE Publications, Thousand Oaks (2001)

Hoft, N.L.: International Technical Communication: How to Export Information about High Technology. Wiley, New York (1995)

Ladner, S.: Practical Ethnography. A Guide to Doing Ethnography in the Private Sector. Left Coast Press Inc., Walnut Creek (2014)

Roberts, S.: Embedded Anthropology: My Talk from the RSA Design Society. http://www.ideasbazaar.com/embedded-anthropology-my-talk-from-the-rsa-design-society-event/. Accessed 17 Jan 2018

Sun, H.: Cross-Cultural Technology Design. Creating Culture-Sensitive Technology for Local Users. Oxford University Press, New York (2012)

A Method to Make an Existing System Adaptive

Francesca Gullà(✉), Alessandra Papetti, Roberto Menghi, and Michele Germani

Department of Industrial Engineering and Mathematical Sciences,
Università Politecnica delle Marche, Via Brecce Bianche, 12, 60131 Ancona, Italy
{f.gulla,a.papetti,r.menghi,m.germani}@univpm.it

Abstract. The Adaptive systems are becoming essential into our daily life, thanks to the fast improvement of computing technology and the deployment of Internet of Things (IoT) devices. Despite many efforts have been made to improve adaptive systems design methods they remain very heterogeneous and mostly limited to each domain of application. Moreover, the most of existing adaptive applications propose specific approaches for the development of new systems, without considering the opportunity to convert existing smart systems into adaptive systems. To overcome this limitation, this paper introduces an approach able to support the designers in adapting an already existing system. To this purpose, a new design method, consisting of the three following steps, has been developed: (1) context analysis; (2) adaptive module design; and (3) system re-design. The described approach is applied, as an example, to a Wi-Fi system for the control and management of household appliances developed by a large Italian company.

Keywords: Adaptive system · Adaptive interface · Methodology
Smart environment · Universal design · IoT

1 Introduction

Nowadays, the smart systems are becoming essential into our daily life, thanks to the fast improvement of computing technology and the deployment of Internet of Things (IoT) devices. The success of the smart systems is tightly linked to features such as easy handling and user-friendly graphical interface (GUI). Primarily designing GUIs is an important step in the development of smart systems, as it allows the user to interact effectively and mutually with the intelligent system.

Adaptive Interface (AI) represents a new way to improve the GUI usability and to help the user to carry out their tasks efficiently and enjoyable [1]. Indeed, the AI adapts and customizes automatically its features and structure to the user characteristics, to facilitate users with different skill levels and impairments [2]. In order to make the user experience more enjoyable and efficient, AI is designed and employed in many different contexts: for example, in domestic environments [3, 4], in workplaces application [5, 6] and healthcare domain [7, 8]. However, all these applications propose specific approaches for the development of new systems, without considering the opportunity to convert existing smart systems into adaptive systems.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 91–101, 2018.
https://doi.org/10.1007/978-3-319-91238-7_8

For this reason, this research work aims to provide a method able to support the designers in adapting an already existing system. In particular, starting from existing methodologies developed for new applications [9, 10], a new design method, consisting of the three following steps, has been developed: (1) context analysis; (2) adaptive module design; and (3) system re-design.

2 Research Background

Adaptive systems are entities that adapt their displays and available actions to the user's current goals and abilities, by monitoring user status, the system state and the current situation" [11]. Benyon and Murray [12] define Adaptive Systems as "systems which can automatically alter aspects of their functionalities and/or interface in order to accommodate the differing needs of individuals or groups of users and the changing user needs over time". To achieve this, such systems have to learn the context in which they evolve and to adapt their interfaces and interaction modalities to communicate with users. In detail, an adaptive interface is defined as the collection of software system, algorithms and graphical interface that can modify the data representation, the way of interaction, and the assignment of tasks to be performed as a function of the ability of the user, the state of the system, and the environment. The research literature describes many approaches that can be used to design an adaptive interface.

Dieterich et al., defines the adaptation strategy as a decision-making process, which answer to the following questions [13]: Where? Why? What? To what? How? The first attribute identifies the application areas in which the adaptive system can be helpful; the second attribute identify the reason to make an adaptive system: depending on the context and on the system type, the "adaptation goals" concern a search goal, a problem-solving or learning goal [14]; the third attribute refers to what aspects are subject to adaptation in a particular system; in detail, what features undergo a change depending on which users: "adaptation constituents" [11]; the fourth attribute refers to what aspects of the user consider to make adaptation: in general, the adaptation depends on the user's characteristics and preferences, on the user's goal, on the user's background and on the utilization context [15]; the fifth attribute identifies how to make adaptation through the definition a set of rules "adaptation rules" [14]; finally, an additional dimension has been added to those highlighted by Benyon and Murray: when to adapt? The adaptation can be carried out in run-time (during the use of the system) or statically (at the first access of the system) [12]. Stephanids et al., define these attributes as "adaptation determinants" [14]. Liu et al., design a dynamic, seamlessly personalized adaptive user interface through the use of an approach based on episodes identification and association (EIA) [16].

Namgoong et al. propose an adaptive user controller exploiting the representation of devices through semantic descriptions [17]; Yen and Acay introduce a generalized framework for the adaptation of user interface in complex environments [18]; new framework to design an AUI in the Web Application has been proposed [19–21]; Bongartz et al., define a new architecture for the AUI design in the Smart environment context [22]; Song and Cho provide a new AUI through the use of Bayesian Network in a ubiquitous environment [23]; Wang et al., provide an adaptive user interface model based perceptual control theory [24]; in the world wide web Rebai uses learning algorithms and Bayesian network to provide a new AUIs [25].

There are several contexts in which the adaptive systems have been used. In the health context, for example, the adaptive user interface has been used in a hypermedia data browsing to develop a clinical workstation prototype for the cardiology domain [26], in cancer domains to develop an interactive system which provide patients information [27] in e-health domain for tracking and monitoring health data for patients [7]. In smart environments context, the adaptive user interface has been used to manage and control various devices simultaneously in a modern home theater system [23]. Yen et al. provide a generalized framework for the UI adaptation in complex environments [18]. El-Bakry et al. present a new adaptive interface design for web application [20]. In addition, the adaptive user interfaces have been used in a work smart environments context, for a warehouse picking system [22].

In the mobile device context, the adaptive user interface has been used to adapt a graphical interface to a various type of user [24, 28, 29]. The literature on the adaptive interfaces research context is very heterogeneous and closely linked with each domain of application; however, all these experimental applications propose methods and specific approaches for the creation of new adaptive systems, without considering the adaptation of existing systems.

3 The Design Method

Not every system can just be turned into an adaptive one, so it is worth to specify some requirements that the existing system should meet for the proposed method to be applicable. They are the following:

- Communication and storing capabilities should be supported by the system to ensure the proper data flow;
- Capabilities of sensing contextual data from the user, the environment, the system itself. The system should be able to recognize user and human-system interaction information;
- Adaptable elements and actions should be supported by the system. They allow introducing variability in system behavior based on the changeable requirements.

In addition, the economic feasibility to make a system adaptive should be verified.

The method consists of three main steps: (1) context analysis; (2) adaptive module design; and (3) system re-design (Fig. 1).

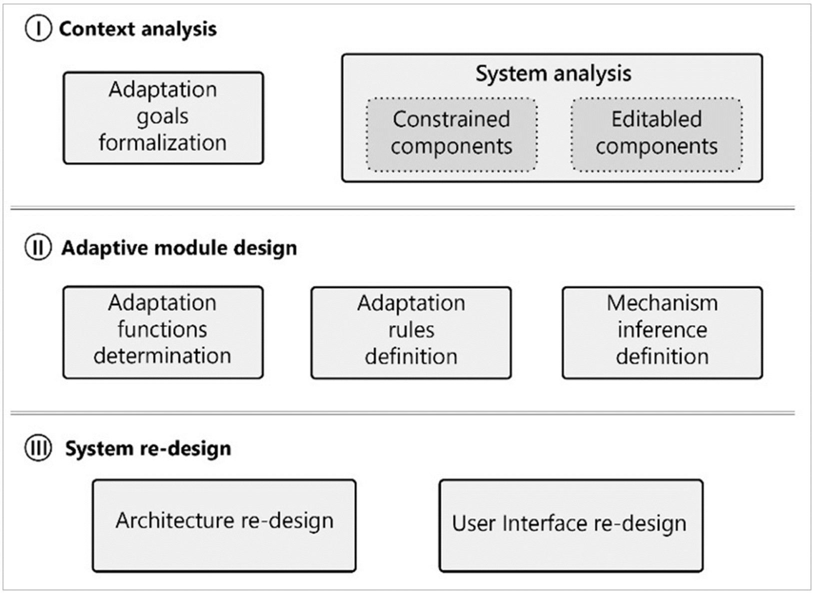


Fig. 1. Design method

The context analysis consists of two interdependent activities. The former identifies the reason to make a system adaptive by defining the "adaptation goals", which could refer to search goals, problem-solving or learning goals. On the basis of the final target, there are different type of adaptations such as user-centered adaptation, target-oriented adaptation, software architecture adaptation. In general, the profile of the target user plays a key role. Therefore, it is essential to understand his/her characteristics (e.g. goals, needs, moods, preferences, intentions, etc.) and investigate how to acquire the missing information. In case of potential conflicting system goals, it is suggested to perform a trade-off analysis. At the end of this activity, the main functionalities (both existing and new) of the adaptive system must be clear. The latter involves the study of the main system features such as architecture, framework, database, script code and layout, etc. It includes the characteristics of the application, the devices, and the physical features of a system as well as the relationships between elements. According to the adaptation purpose and the company's needs, the following components have to be identified: existing elements or features that cannot be modified (i.e., constrained component) and existing requirements that can be changed (i.e., editable components). In particular, the main constraints concern technical issues, marketing requirements and economic strategies.

In order to satisfy the adaptation goals, all possible design variabilities should be explored by constructing a design space with the help of experts' knowledge. The design of the adaptive module consists of three sub-modules. The first one refers to the determination of what aspects of the system are subject to adaptation or what new features, which undergo a change depending on users interaction, should be implemented. They may refer to semantics, syntactic (i.e. sequencing of the interaction) or lexical aspects of interaction (i.e. interface objects, interaction techniques, media and modalities). They include generic functions (error correction, help, etc.) and interaction (task simplification, improved user experience, etc.). At the same time, it is necessary to establish what aspects of the user consider to make adaptation. The second sub-module identifies how to make adaptation. It consists on the definition of a set of rules that involves the logics formalization for a proper adaptive behavior [30]. The last sub-module concerns the mechanism inference definition that consists in the identification of the domain-independent algorithm used to draw conclusions or perform actions using the knowledge base and the user's responses. It also defines the right time for adaptation, which is a crucial point in the adaptive process. The adaptation can be carried out in run-time (during the use of the system) or statically (at the first access of the system).

Finally, the system re-design consists of modifying the current architecture by inserting the adaptive module and editing the graphical interface. The system architecture redesign includes the database restructuring, the addition of the adaptation module and possible sensors, the data flow optimization, and the enabling of new system actions. The method implications on the user interface redesign could refer to graphics (e.g. colors, brightness, contrast, font, etc.), navigation (e.g. wizard mode), interaction mode (e.g. voice, gesture, etc.), and content (e.g. pre-selected function, help, notify, etc.).

4 The Case Study

The described method is applied, as an example, to a Wi-Fi system for the control and management of household appliances developed by a large Italian company.

4.1 Context Analysis

First of all, the final adaptation goals have been defined: make the system adaptive on the basis of the user behavior and needs. Subsequently, a system analysis to identify the specifics of the smart system has been carried out. Three different modules make the current system architecture: the mobile application, the household appliances and the Cloud (Server-DB) (Fig. 2).

The user-household appliance interaction takes place in two different ways: Direct Interaction (performed on the household appliance) or Mobile Interaction (performed on the mobile application). The application allows the user-household appliance communication both in the home (through Local Network) and outside (through Remote Control). In the first case queries are sent directly to the household appliances and saved on the Cloud DB (View HAs status and HAs Programming). In the case of Remote Control use, however, the queries are sent to the Household Appliances through the Server on Cloud. In detail, the HAs programming queries will be sent through the Appliance Actuator in Active Remote Control mode, while the status will be displayed on Cloud Database. The status update on the Database takes place every 30 s (polling).

On the basis of the defined adaptation goals, the constrained components (CC) and editable components (EC) have been identified. Specifically, the defined CC are: User-Household appliances communication modality, Home-appliance Program Temperature

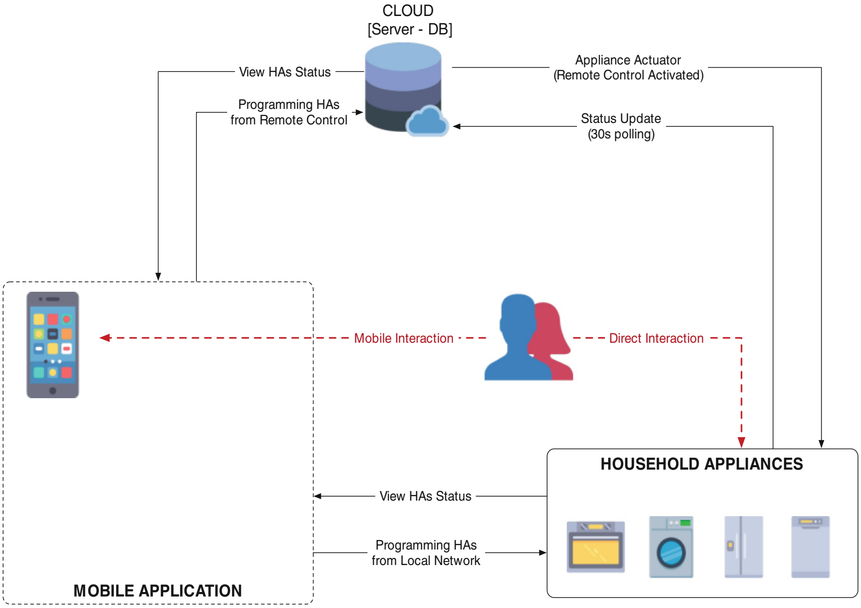


Fig. 2. The architecture design

and Cycle, Mobile Application Interface color and font. Instead, the defined EC are: Database, Mobile Application Interface contrast and layout, mobile application browsing mode, adaptive functions addition (i.e., programs or temperature suggestion).

4.2 Adaptive Module Design

Starting from the current system architecture and based on the defined requirements, the adaptive module has been designed. Specifically, the Adaptive module consists of:

- **Adaptation Rules,** which involves the:
  - **Back-end Translation for Bayesian Network**. This module is responsible for reading the updated Database, filtering and translating it. Specifically, the filtering action selects from the DB the states needed to perform the adaptation. Instead, the translation action consists of translating the states, previously filtered, into a compatible language to be sent to the Adaptive Middleware module.
  - **Adaptive Middleware.** This second module consists of Adaptive Engine, characterized by the Decision-Making Algorithm (DM) [12]. The adaptive middleware is responsible for Bayesian network learning; through this module data received from the Middleware Translation are processed and sent to the network. Once trained the network and obtained the specific conditional probabilities, these will be elaborated through the Decision-Making algorithm. The DM will identify the appropriate adaptation actions to be taken on the system by upgrading network node dependencies on the mobile application.

- **Mechanism Inference**, this module consists of the **Bayesian Network** customized for the System. The nodes number constituting the network is affected by the devices number involved and by the user's monitoring actions. The network learning is made via the adaptive middleware.

4.3 System Re-design

4.3.1 Architecture Re-design

Finally, the system architecture has been re-designed. In details, the Adaptive module introduction has been carried out on two different levels: the first level involves the DB Server on Cloud, the second level involves the mobile application. The adaptation actions suggested by the adaptive module will be shown on the mobile device through the user interface.

Figure 3 shows the adaptive module addition in the system architecture.

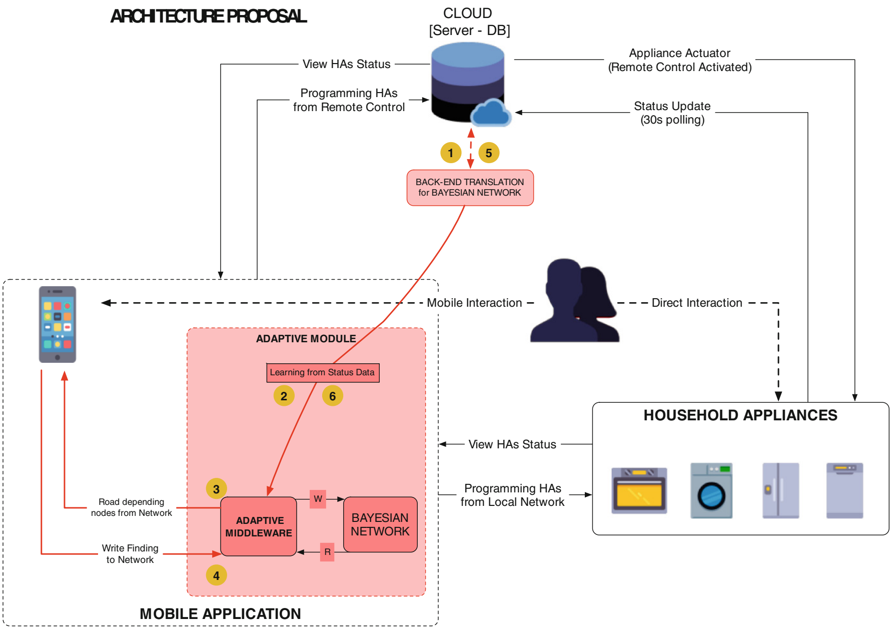


Fig. 3. The architecture re-design (in red the adaptive module addition) (Color figure online)

4.3.2 User Interface Re-design

In order to simplify the control and the setting of household appliances, the mobile application UI has been redesigned. The UI re-designing consists in removing features that are not required by the user and adapting others to his preferences. Based on the user profile, the previously interaction and the monitored actions, the adaptive module introduction allows finding the optimal features combination for the interface (Table 1).

Table 1. User interface re-design.

Current UI	Adaptation goal	UI adaptation action
UI contrast and zoom: fixed	Increase UI accessibility and readability	Adapt UI contrast and zoom at the User Login
–	Improve user experience	Add/replace image with text to help user in programming the device
Single page layout	Optimize the layout (to highlight relevant information and minimize irrelevant and potentially distracting information)	Change UI layout at the User Login (i.e. propose the most fitting option for the user to provide an easy interaction and usage)
No Program suggestion	Best Program Suggestion	Depending on the user login, day and time range, the adaptive network suggests the best Program to the end user. The application will set as the first choice the Program suggested by the network, as a second choice the Program suggested by the Company and in cascade all the others
No Temperature suggestion	Best Temperature Suggestion	Depending on the user login, day, time range and suggested Program, the adaptive network highlights the best Temperature to the end user. The application will set as the first choice the Temperature suggested by the network, as a second choice the Temperature suggested by the Company and in cascade all the others
No Delay Start suggestion	Delay Start and cooking duration Suggestion	Depending on the user login, day, time range and suggested Program, the adaptive network suggest the delay start and cooking duration (oven case). The application will set as the first choice the most used delay start time and cooking duration

The UI adaptations concerned two main features: the first one consists of changing the interface's characteristics, and the last consists of changing the content displayed by the interface. The first UI adaptations is based on varying the graphical features (contrast, zoom, etc.,) or its layout configuration. In this case, the platform is aware of the user experience with the interface and it makes the corresponding adaptation based on these interactions. The second UI adaptation deals with the interface contents. Contents represent all interface items editable according to uses and actions the user acts on interface, own preferences and needs.

As a result, the user interface has been split into two presentation modalities: Default (i.e. the current UI) and New UI. The New UI has been designed by creating separate user interfaces (starting from Default UI) and by dynamically adapting the graphical features and the content according to the adaptation rules.

5 Conclusion and Future Works

In order to support the development of smart systems able to ensure a better user experience, this paper proposes a design methodology for making an existing system adaptive. It consists of three steps: from the context analysis in terms of goals, constraints, and opportunities to the redesign of the system architecture and GUI, through the definition of adaptation algorithms.

The methodology has been experimented by an Italian company, which aimed at improving its Wi-Fi system for the control and management of household appliances. It led designers in their activity, supporting the rules definition and reducing the overall efforts. Such approach allowed performing adaptation activities (i.e., monitoring, analysis, decision-making, execution, etc.) without affecting system performance and availability.

Different applications will be carried out to highlight the benefits of using the proposed methodology as opposed to the traditional ones and to validate it. Aspects related to usability, technology, efficiency and costs will be addressed. The final aim is the development of a set of guidelines and a user-friendly tool that effectively support the universal design.

References

1. Lavie, T., Meyer, J.: Benefits and costs of adaptive user interfaces. Int. J. Hum. Comput. Stud. **68**(8), 508–524 (2010)
2. Haas, M.W., Hettinger, L.J.: Current research in adaptive interfaces. Int. J. Aviat. Psychol. **11**(2), 119–121 (2001)
3. Kushwaha, N., Kim, M., Kim, D.Y., Cho, W.D.: An intelligent agent for ubiquitous computing environments: smart home UT-AGENT. In: Proceedings of the Second IEEE Workshop on Software Technologies for Future Embedded and Ubiquitous Systems, pp. 157–159. IEEE, May 2004
4. Gullà, F., Ceccacci, S., Menghi, R., Cavalieri, L., Germani, M.: Adaptive interface for smart home: a new design approach. In: Cavallo, F., Marletta, V., Monteriù, A., Siciliano, P. (eds.) ForItAAL 2016. LNEE, vol. 426, pp. 107–115. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54283-6_8
5. Letsu-Dake, E., Ntuen, C.A.: A case study of experimental evaluation of adaptive interfaces. Int. J. Ind. Ergon. **40**(1), 34–40 (2010)

6. Evans, D., Arasu, V.A., Fendley, M.: A model-based framework for an adaptive user interface in supervisory control. IFAC Proc. Vol. **46**(15), 37–43 (2013)
7. Vasilyeva, E., Pechenizkiy, M., Puuronen, S.: Towards the framework of adaptive user interfaces for eHealth. In: 2005 Proceedings 18th IEEE Symposium on Computer-Based Medical Systems, pp. 139–144. IEEE, June 2005
8. Shakshuki, E.M., Reid, M., Sheltami, T.R.: An adaptive user interface in healthcare. Procedia Comput. Sci. **56**, 49–58 (2015)
9. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. Adapt. Hypertext Hypermedia, 1–43 (1998)
10. Gullà, F., Cavalieri, L., Ceccacci, S., Germani, M., Bevilacqua, R.: Method to design adaptable and adaptive user interfaces. In: Stephanidis, C. (ed.) HCI 2015. CCIS, vol. 528, pp. 19–24. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21380-4_4
11. Rothrock, L., Koubek, R., Fuchs, F., Haas, M., Salvendy, G.: Review and reappraisal of adaptive interfaces: toward biologically inspired paradigms. Theor. Issues Ergon. Sci. **3**(1), 47–84 (2002)
12. Benyon, D., Murray, D.: Applying user modeling to human-computer interaction design. Artif. Intell. Rev. **7**(3), 199–225 (1993)
13. Dieterich, H., Malinowski, U., Kühme, T., Schneider-Hufschmidt, M.: State of the art in adaptive user interfaces. Hum. Factors Inf. Technol. **10**, 13 (1993)
14. Stephanidis, C., Karagiannidis, C., Koumpis, A.: Decision making in intelligent user interfaces. In: Proceedings of the 2nd International Conference on Intelligent User Interfaces, pp. 195–202. ACM (1997)
15. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. User Model. User Adap. Inter. **6**(2–3), 87–129 (1996)
16. Liu, J., Wong, C.K., Hui, K.K.: An adaptive user interface based on personalized learning. IEEE Intell. Syst. **18**(2), 52–55 (2003)
17. Namgoong, H., Sohn, J.C., Cho, Y.J., Chung, Y.K.: An adaptive user interface in smart environment exploiting semantic descriptions. In: 2006 IEEE Tenth International Symposium on Consumer Electronics, ISCE 2006, pp. 1–6. IEEE (2006)
18. Yen, G.G., Acay, D.: Adaptive user interfaces in complex supervisory tasks. ISA Trans. **48**(2), 196–205 (2009)
19. He, J., Yen, I.L., Peng, T., Dong, J., Bastani, F.: An adaptive user interface generation framework for web services. In: Congress on Services Part II, SERVICES-2, pp. 175–182. IEEE (2008)
20. El-Bakry, H.M., Riad, A.M., Abu-Elsoud, M., Mohamed, S., Hassan, A.E., Kandel, M.S., Mastorakis, N.: Adaptive user interface for web applications. In: Proceedings of the 4th WSEAS International Conference on Business Administration Recent Advances in Business Administration (ICBA 2010), pp. 20–22 (2010)
21. Hervás, R., Bravo, J.: Towards the ubiquitous visualization: adaptive user-interfaces based on the semantic web. Interact. Comput. **23**(1), 40–45 (2010)
22. Bongartz, S., Jin, Y., Paternò, F., Rett, J., Santoro, C., Spano, L.D.: Adaptive user interfaces for smart environments with the support of model-based languages. In: Paternò, F., de Ruyter, B., Markopoulos, P., Santoro, C., van Loenen, E., Luyten, K. (eds.) AmI 2012. LNCS, vol. 7683, pp. 33–48. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34898-3_3
23. Song, I.J., Cho, S.B.: Bayesian and behavior networks for context-adaptive user interface in a ubiquitous home environment. Expert Syst. Appl. **40**(5), 1827–1838 (2013)
24. Wang, X., Hua, Q.Y., Zou, F., Guo, L.: An adaptive user interface model for mobile devices based on perceptual control theory. In: 2014 5th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 908–911. IEEE (2014)

25. Rebai, R., Maalej, M.A., Mahfoudhi, A., Abid, M.: Building and evaluating an adaptive user interface using a Bayesian network approach. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **14**(7), 548–565 (2016)
26. Ghédira, C., Maret, P., Fayn, J., Rubel, P.: Adaptive user interface customization through browsing knowledge capitalization. Int. J. Med. Informatics **68**(1), 219–228 (2002)
27. Bental, D., Cawsey, A., Pearson, J., Jones, R.: Adapting web-based information to the needs of patients with Cancer. In: Brusilovsky, P., Stock, O., Strapparava, C. (eds.) AH 2000. LNCS, vol. 1892, pp. 27–37. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44595-1_4
28. Mitrović, N., Mena, E.: Adaptive user interface for mobile devices. In: Forbrig, P., Limbourg, Q., Vanderdonckt, J., Urban, B. (eds.) DSV-IS 2002. LNCS, vol. 2545, pp. 29–43. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36235-5_3
29. Nivethika, M., Vithiya, I., Anntharshika, S., Deegalla, S.: Personalized and adaptive user interface framework for mobile application. In: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1913–1918. IEEE (2013)
30. Gullà, F., Cavalieri, L., Ceccacci, S., Germani, M.: A BBN-based method to manage adaptive behavior of a smart user interface. Procedia CIRP **50**, 535–540 (2016)

Towards an Integrated Approach to Studying Virtual Reality-Mediated Social Behaviors

Jeffrey C. F. Ho(✉)

School of Design, Hong Kong Polytechnic University, Hung Hom, Hong Kong
jeffrey.cf.ho@polyu.edu.hk

Abstract. The study on social behaviors that are mediated by virtual reality (VR) has become a major challenge for researchers given the increasing popularity of VR in the mass market. In this study, we reviewed and discussed certain issues on studying social behavior toward multi-user VR application in different settings. This study is aimed at proposing an integrated approach to studying VR-mediated social behaviors. The diversity in contexts of multi-user VR applications guarantees extensive research foci, such as team dynamics, trust, and persuasion. An integrated approach that can address this new class of research questions on social behavior in VR is necessary. The traditional research methods may capture various aspects of VR-mediated social behaviors, whereas several methods may demonstrate limitations. This paper explores the possibilities and challenges of an integrated approach to studying VR-mediated social behaviors.

Keywords: Virtual reality · Social behavior · Methodology

1 Introduction

Virtual reality (VR) has become a popular research topic since the 1990s. VR has recently received a considerable attention from the mass market. VR technologies allow users to enter a virtual environment, and numerous VR applications even allow multiple users to enter a virtual world simultaneously. For example, *Star Trek: Bridge Crew* (2017) has a VR mode that allows four players to be on the bridge of a starship and collaborate on exploratory missions in an imaginary world. Social behaviors that are mediated by VR technologies may involve different combinations of communication formats (e.g., voice and gestures vs. voice only) and in various location settings (e.g., co-location vs. virtual teleportation). Thus, a gap in the literature remains to be filled. However, the traditional research methods used in social sciences demonstrate limitations in studying VR-mediated social behaviors. An integration of methods can aid researchers in measuring and observing social behaviors on VR. This study is aimed at exploring such integrated approach.

Various scenarios in using VR involve social behavior. Multiuser VR applications support multiple users who are physically co-located, whereas several VR applications are designed to support remotely located users. VR systems are also designed to support social interactions between VR users and outsiders [5, 8]. Stream gameplay is a popular phenomenon that includes VR gameplay. The interactions between a VR broadcaster

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 102–113, 2018.
https://doi.org/10.1007/978-3-319-91238-7_9

and the audience can result in new forms of social behaviors. Different types of social behavior manifest in all these scenarios. Furthermore, the forms of communication vary from verbal (e.g., spoken words) to non-verbal communication (e.g., gestures, orientations, positions of physical bodies and avatars, and use of virtual objects). Multiuser VR applications, such as multiplayer VR games and collaborations in design, have diverse contexts [14]. Such diversity in contexts of multiuser VR applications guarantees extensive research foci, such as team dynamics, trust, and persuasion. Thus, an integrated approach that can address this new class of research questions on social behavior in VR is necessary.

This study focuses on the VR applications based on head-mounted displays (HMDs). Therefore, examples such as [15] are disregarded.

A discussion of the traditional research methods and their suitability for different steps of a typical experimental session is introduced in this paper. A classification of various scenarios of VR-mediated social behaviors is proposed. The challenges and considerations of integrating research methods into studying each type of scenarios are discussed. This paper primarily discusses the issues in the data collection stage. However, analysis methods of interactions (e.g., [2, 17]) are excluded.

2 Four Traditional Methods

This study focuses on four research methods that researchers can adopt in an experiment or user study on VR-mediated social behavior and multi-user VR applications. These methods, namely, questionnaire, interview, observation, and focus group, are well established and commonly used in studies on social science and human–computer interaction.

The focus on these methods is due to they are well established, thereby indicating that researchers can use previous works as methodological references. The findings from these methods can be compared with previous studies that applied the same methods. Furthermore, these methods do not require special tools, such as eye-tracking and electroencephalography devices. Researchers who have relatively simple settings in their laboratory can adopt these methods, which generally do not require changes in or amendments to the studied VR applications. The enhancement of a VR application if the source code is unavailable is very difficult, if not infeasible; this scenario is frequently the case for commercial VR applications.

Questionnaire. Questionnaires [3] allow researchers to collect quantitative and qualitative feedback from participants. Established questionnaires measure different aspects of the internal states of participants, such as immersion [9]. Questionnaires can collect responses from participants in a structured manner without the assistance of a research staff. The questionnaire can be in a paper- or computer-based form. Participants answer several questions by themselves. Thus, many research staffs are not required even if several participants are involved in a session.

Questionnaires require the full attention of participants. These participants must pause from their errand and focus on responding to the questions on a questionnaire. In the context of VR applications, the time to ask participants to fill out a questionnaire should be after their experience with a VR application. The drawback is that their immediate response cannot be captured. This scenario is particularly important in social behaviors because a typical trial or experimental session would involve a series of interactions and communications among the participants. Capturing the participants' response to one another's behavior is important.

Simple questions to the participants are possible during exposure to a VR application. This process is not equally disrupting as asking them to fill out a questionnaire. Researchers can ask brief questions to all participants who are in a VR world during a session and ask these participants to respond via hand or body gestures (e.g., "Please indicate with your fingers how excited you are now. Five means extremely excited. One means not excited at all."). Responses with hand or body gestures can prevent participants from knowing one another's responses and being influenced. This brief questionnaire approach can only apply to simple questions that require simple answers. Simple answers are expressed with simple gestures, such as a thumb up or down for yes or no and showing fingers to rate on a 5-point Likert scale.

Interview. Interviews [10] allow researchers to guide participants to express their viewpoints in a humanistic manner. Thus, researchers can carefully control the pace and even adjust content and/or order of questions in accordance with the behavior of the participants (e.g., asking the participants for the reasons for performing certain actions in a VR usage session). Participants can ask questions if the questions are unclear to them. An interview can be an option for VR-mediated social behaviors. This method is useful in collecting qualitative feedback from the participants but suffers the limitation of requiring the attention of participants and thus might interrupt the participants' exposure to VR stimuli. Interviews can be conducted retrospectively and are typically performed after the participants have experienced a VR application. Retrospective interviews can include recordings of a session of VR usage to remind participants about their experience and serve as a reference of their answers (e.g., referring to moments when they perform certain actions). Retrospective interviews with recordings of VR usage can assist participants in revisiting their experience and expressing their emotions and cognitions in a social interaction in a VR context.

An interview requires at least one researcher for each participant. The participants in a session that involves multiple participants (which is typically the case for studying VR-mediated social behaviors) are interviewed nearly at the same time, thereby requiring numerous research staffs to support such procedure.

Observation. Observation [7] can be a method that is used to observe social interactions, including hand gestures, body movements, and facial expressions. Researchers must observe behaviors in the real and VR worlds when applied to VR contexts during VR usage. If participants are co-located, then researchers can directly observe the behavior of multiple participants simultaneously. Behaviors in the real world denote differently in a VR world. Observations of VR-mediated social behaviors require a mapping between the two worlds. The mapping, such as direct observation or via a live video camera, can be achieved by showing a live view on a monitor that displays the

VR world and a live view of the real world. A live video camera can also record a session and offer recordings to support a retrospective interview.

Focus Group. Researchers can consider conducting focus groups [12] with a group of participants retrospectively and obtain collective feedback from multiple participants simultaneously. Similar to interviews, recordings of the participants' VR usage can support retrospective focus groups. The drawback is that responses from individual participants can influence or be influenced by their peers. Every participant experiences different parts of a VR usage session. A retrospective focus group may require multiple recordings from different views to be displayed in a synchronized manner.

2.1 Integrating Traditional Methods

This section describes a framework for integrating the four traditional research methods in an experiment session of multi-user VR applications. A typical experimental session that explores social behavior in a multi-user VR application involves the participants' exposure to a VR application. A session can be divided into three stages, namely, "before exposure," "during exposure," and "after exposure" (Fig. 1).

	Before Exposure	During Exposure	After Exposure	
			Individual Response	Collective Response
Questionnaire	Questionnaire on background	Simple questions	Questionnaire	
Interview			Retrospective interview	
Observation		Observation with notetaking and recording		
Focus Group				Retrospective focus group

Fig. 1. Integrated framework of the traditional methods in an experiment session

Before Exposure. The administration of a questionnaire that requires the participants' demographics, previous experience with VR and similar applications, and reference values for measures involved in the study (e.g., pre-exposure attitude towards certain topics) must be performed before the participants are exposed to a VR application.

During Exposure. Researchers can observe and record the behavior of the participants during an exposure to a VR application. The researchers can take video recordings and screen captures while observing to support the retrospective interview and focus group. Researchers can also ask simple questions to participants at certain points during the exposure. However, this activity should be maintained to a minimum to reduce severe intrusion to the participants' exposure to the VR application.

After Exposure. Researchers can measure the participants' response after the participants' exposure to a multi-user VR application. Researchers should focus on individual response first by using a questionnaire and retrospective interview and then proceed to the collective response by conducting a focus group. The participants can express their own views in the questionnaires or individual retrospective interviews before listening to other participants in the focus group. This activity is performed in this order to prevent participants from influencing one another while individual responses are collected.

3 Classification and Dimensions

This section presents the proposed classification scenarios of the VR-mediated social behaviors. The classification is aimed at identifying and discussing the challenges faced by researchers who attempt to investigate social behavior in each scenario (Fig. 2).

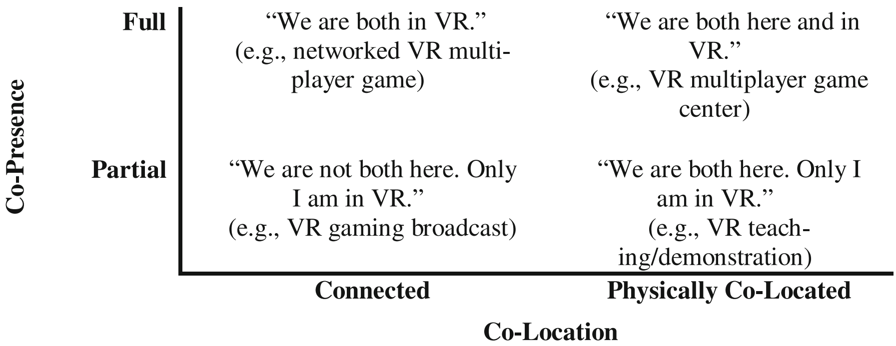


Fig. 2. Classification of VR-mediated social behaviors

3.1 Co-presence

The first dimension of the classification is "co-presence," which refers to the degree that the users are all co-present in a VR world. One end of the dimension is full co-presence; in this dimension, all the users utilize the HMD VR devices and are in a VR world. Each user wears an HMD VR device to enter a VR world.

Certain situations occur when not all users are in a VR world. The HMD VR technologies are becoming popular, but not every user owns this device. There are cases where only one or some users of a VR application wear HMDs and enter the VR environment. Other users can still join the VR application via other devices, such as tablet devices [1, 8], but are not similar to the presence of a user(s) that utilize an HMD VR device. In a digital game context, this scenario is called asymmetric VR gaming [8]. In these situations, the co-presence is partial.

The co-presence dimension is based on the classification by Kraus and Kibsgaard [11]. However, their classification disregards cases in which several users are partially co-present with other users but are all physically co-located [5, 8].

3.2 Co-location

The second dimension "co-location" refers to the extent in which the users of a VR application are physically co-located. That is, one end of this dimension is physically co-located. For certain multi-user VR applications (e.g., *Star Trek: Bridge Crew* (2017)), if multiple sets of HMD VR devices are co-located, then multiple users can enter the VR world together. If users are physically co-located, then they can select to communicate directly in the real world (e.g., shouting at one another while playing a VR game together) or via the VR application. Researchers who study such social behavior must capture and observe the interactions in the VR and the real worlds (and potentially the transitions between the two worlds).

The other end of the dimension is connected. That is, several or all users of a multi-user VR application are connected but not physically co-located. The rationale for distinguishing the aspect of co-location is the difference in considerations when studying the social behavior involved. If all the users are not physically co-located and only connected via a network, then these users must communicate via the application. Researchers may further focus on the types of information (e.g., verbal and social cues) that can be transmitted via the studied network-capable VR applications.

4 Methodological Considerations in Each Scenario

This section discusses the considerations of using the research methods discussed in each category. The considerations before, during, and after the participants' exposure to a multi-user VR application of the category are presented.

4.1 Physically Co-located and Full Co-presence in VR

VR applications that support multiple co-located users to enter a VR world enable many types of social interactions. This type of VR application allows users to interact via the VR (via the application) and the real world. For example, Chaqué and Charbonnier [4] developed a multi-user VR platform that combines real and virtual worlds. This platform uses tangible objects in the real world as props for the users to interact and is designed for physically co-located users to enter a VR world together. Two players in the example game demonstrated in their paper interact with a simple cardboard box in the real world, but this box appears to be an Egyptian chest in the VR world. Players in the gameplay can cooperate on moving the chest; this scenario is a VR-mediated interaction. The players can verbally communicate with one another directly in the real world.

Zaman et al. [19] developed and user-tested a collaborative VR application that allows multiple physically co-located users to collaborate on a spatial design task in a VR world. The participants performed in pairs several collaborative design tasks in a series of user-testing sessions in the system. The researchers audiotaped the sessions for content analysis of verbal communication and performed a content analysis of the transcripts of the audio tapes. After the test, each participant filled out a questionnaire with questions based on the Likert scale.

Before Exposure. Before participants are exposed to a VR application of this category, researchers can ask each participant to fill out a questionnaire. As the participants are physically co-located, the questionnaire can be conducted by one research staff.

During Exposure. Researchers can consider directly observing any interactions in the VR and real worlds. Studying behavior in the real world can be challenging. If a behavior is tracked by an application, such as the cardboard box in [4], then the application can record tracking information. Otherwise, researchers need other tools to observe and capture interactions in the real world. The interactions may include body movement (with or without props) and verbal communication. As participants are co-located, that the interactions occur in a 3D space presents a challenge. Setting up several video cameras may capture the interactions of two to three users. One risk is that the participants' body may block the view of the cameras. Such a risk will increase along with the number of participants in a session. Although direct observation by researchers can help fill the gaps not captured by the cameras, the risk of occlusion remains.

The cognitive and emotional states of the participants is difficult to study, as in other immersive technologies. Asking the participants to fill out a questionnaire is possible, but this approach disrupts the experience. One alternative is pausing the exposure to VR application and ask participants simple questions. As the participants are co-located, this course of action can be completed easily. The questions should be short to minimize interruption to the participants' experience. As discussed in Sect. 2, the questions should be straightforward so that the participants can respond with simple hand gestures to prevent the participants from influencing one another.

After Exposure. Retrospective interview with video recordings of a session may provide researchers a window to the internal states of the participants. However, participants' immediate response in their interactions cannot be examined. In addition, the session spreads across the VR and real worlds. Body gestures and movements observed in the real world should correspond to certain actions in the VR world. In [19], partners appear as virtual humanoid hands in the VR world. Recordings of the real and VR worlds are needed to support retrospective interviews.

In the scenarios of this category, participants are co-located. They enter a VR world together in the same physical location. A retrospective focus group can be conducted right after the collection of individual feedback (with either questionnaire or individual interview or both) to allow the participants to discuss their shared experience.

4.2 Connected and Full Co-presence in VR

VR applications that fall under this category support multiple users with HMD VR devices to remotely enter a VR world. All the interactions between the users are mediated and transmitted by the VR applications. Sra [18] put forward a mechanism to present different physical constraints faced by different users who are remotely connected to a VR world.

One challenge is that players or users only see the representation of one another. The representation can be an avatar (e.g., Facebook's Space), an object, or a 3D scan of a player. For example, Facebook Spaces[1] allow users of HMD VR devices to remotely meet on a VR world. The users can draw 3D objects, chat, and view panoramic imagery together. Users are co-present in the VR world as avatars.

Before Exposure. Researchers can consider a background questionnaire before an exposure to a VR application. The challenge lies in the distributed nature of the experimental setup. To simulate a remotely connected scenario, participants must be separated from one another and enter a VR world. Each participant in a session needs a separate location. Therefore, additional research staff and space are necessary. As the participants, researchers, and equipment are not co-located, the researchers need to communicate efficiently with one another so that the timing of executing the procedure can be coordinated. For instance, multiple participants need to enter the VR world at the same time to prevent individual participants from being exposed to the world alone. Such situations require the researchers to carefully manage time and communicate efficiently so that the procedure for individual participants can be coordinated.

During Exposure. As participants are separately located, multiple researchers are required to observe their interactions. A researcher can observe a participant in one location. The notetaking of the researchers should be synchronized. The participants' behavior (e.g., waving hands) may represent an interaction with other participants in a VR world. Therefore, mapping is necessary between the behavior in the physical and VR worlds. The researchers must observe the participants' behavior with a reference to a view in the VR world, which can be a live view of the participants' HMD view or a spectator view provided by a VR application.

After Exposure. After exposure to a VR application, each researcher can administer a post-exposure questionnaire or conduct a retrospective interview with each participant. If a focus group is planned, then the researchers need to bring the participants together after collecting individual response. As the participants are not physically co-located, they may have not seen one other before the exposure. In this case, the focus group may need a warm-up period. The participants may also need an introduction to understand that the other participants are actually the other users they meet during the exposure.

4.3 Physically Co-located and Partial Co-presence in VR

In cases under this category, all users are physically co-located. Therefore, part of their interactions can happen in the real world. Recently, researchers have gained interests in VR applications that allow HMD VR users to interact with non-HMD users. One reason is that HMD VR devices are not as popular as smartphones. Not every user has a HMD VR device. A few VR games allow multiple users to engage in a VR world even if only one set of VR equipment is available. Other users can engage in the VR world via commonly available devices, such as tablets. Gugenheimer et al. [8] proposed a system called ShareVR that allows non-HMD users to interact with HMD users and be part of the VR world experience. ShareVR consists of a floor projection showing the VR scene

[1] https://www.facebook.com/spaces.

in the real world, mobile displays with tracking capabilities for non-HMD users, and a HMD VR device. Gugenheimer et al. [8] called this type of interaction co-located asymmetric interaction. They presented use cases of digital games and drawing application. In their user study evaluating ShareVR, Gugenheimer et al. [8] asked the participants to fill out a questionnaire after each experiencing ShareVR.

Another motivation of VR applications under this category is that HMD VR devices are designed to fully immerse users. The HMD blocks the attention of users to the real world. If a user is wearing an HMD and enters a VR world, he or she is blocked from interactions with anyone in the real world. The HMD user cannot interact with the real world even if necessary. Chan and Minamizawa [5] proposed and developed a HMD called FrontFace that supports communication between co-located HMD users and outsiders. The innovative HMD consists of an external display that shows a user's visual attention (eyes) and reference to the user's position in the VR world. One of the proposed application is VR classroom in which instructors can observe learners wearing FrontFace to examine their learning status. Chan and Minamizawa [5] proposed methods for outsiders to interact with a FrontFace user. They conducted a trial with a small group of three participants to experience and give feedback on the innovative HMD.

Before Exposure. Researchers can administer a background questionnaire in the beginning of an experiment session.

During Exposure. The first challenge is that exposure to a VR system differs between HMD and non-HMD users. As all participants are co-located, researchers can directly observe the participants' interactions. The researchers need to carefully observe how the behaviors of non-HMD users are represented in a VR world for comprehension. The researchers should directly observe users' behavior in the real world and their view in the VR world to scrutinize their representation in the VR world. To capture all the interactions, interactions in the real and VR worlds should be included. The methods to capture the behavior of HMD and non-HMD users may vary.

After Exposure. After exposure, researchers can collect individual response from the participants through a questionnaire and retrospective interview. HMD users and non-HMD users are exposed to different aspects of a VR application. The researchers should consider whether all the participants address the same questions. Alternatively, the questions for HMD and non-HMD users can differ and be specific to their mode of engaging in a VR world.

4.4 Connected and Partial Co-presence in VR

VR applications under this category support users with and without HMD VR to remotely enter a VR world. One use case is supporting HMD users and non-HMD users to remotely collaborate on tasks. Cergeaud et al. [6] proposed a method that allows HMD and non-HMD users to collaborate remotely in a common VR meeting. The system consists of physical objects (or props) to support object manipulation. Cergeaud et al. [6] conducted an interview after letting participants try the system.

Another scenario is online streaming of VR gameplay. Streaming platforms, such as Twitch, support VR gameplay, which allows a streamer wearing HMD to play a VR game while the audience watch the gameplay streaming with regular screens.

Before Exposure. Before entering a multi-user VR application, researchers can ask participants to fill out a pre-exposure questionnaire. As the participants are separately located, multiple research staff are necessary.

During Exposure. As discussed above, pausing a session and asking all participants simple questions are possible. In cases under this category, the participants are not co-located. The pause needs to be synchronized. However, the participants should answer verbally instead of non-verbally.

Direct observation of remotely located participants requires multiple researchers. The researchers should also consider how the non-HMD participants are represented in a VR world and how they can interact and communicate with the HMD participants. Doing so determines how the researchers observe the participants' behavior. For example, if they are spectators who can comment on the HMD participants (e.g., spectators of VR gamers on Twitch), their commenting behavior needs to be observed in the application.

Recordings of the whole session should involve video recordings of the behavior of HMD and non-HMD participants in the real world and screen recordings of live views of HMDs and non-HMD devices.

After Exposure. Researchers can conduct retrospective interview individually with each participant. Similar to the procedure before the exposure to a VR application, multiple researchers are needed. In a retrospective focus group, introduction and warm-up may be necessary because the participants have not met before the exposure. They may not realize that the peer participants in a focus group are those who interacted with them in the VR world. The recordings of HMD and non-HMD devices must be shown in the session to remind them about their experience. As such a retrospective focus group session should happen shortly after an exposure to a VR application, editing the videos is challenging. The recordings from different devices may not be nicely combined. The researchers may need to prepare multiple monitors to play the recordings simultaneously.

5 Toward an Integrated Approach

Scenarios of VR-mediated social behavior are classified. The proposed classification consists of two dimensions distinguishing the extent in which the participants of multi-user VR applications are located physically in the same place and the extent in which they are virtually present in a VR world. There are existing classifications of immersion environments [11] and shared-space technologies [1]. Kraus and Kibsgaard [11] classified the communication of multiple users in immersive environment. Benford et al. [1] classified shared-space technologies according to transportation, artificiality, and spatiality. However, these two studies are not focused on methodological considerations.

The current study provides a framework to systematically consider the concerns in the major steps of experimental procedures involving multi-user VR applications in different settings.

References

1. Benford, S., Greenhalgh, C., Reynard, G., Brown, C., Koleva, B.: Understanding and constructing shared spaces with mixed-reality boundaries. ACM Trans. Comput. Interact. **5**, 185–223 (1998)
2. Bowers, J., O'Brien, J., Pycock, J.: Practically accomplishing immersion: cooperation in and for virtual environments. In: Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work - CSCW 1996, pp. 380–389. ACM Press, New York (1996)
3. Bradburn, N.M., Sudman, S., Wansink, B.: Asking Questions: The Definitive Guide to Questionnaire Design - For Market Research, Political Polls, and Social and Health Questionnaires. Jossey-Bass, San Francisco (2004)
4. Chagué, S., Charbonnier, C.: Real virtuality: a multi-user immersive platform connecting real and virtual worlds. In: Proceedings of the 2016 Virtual Reality International Conference on VRIC 2016, pp. 1–3. ACM Press, New York (2016)
5. Chan, L., Minamizawa, K.: FrontFace: facilitating communication between HMD users and outsiders using front-facing-screen HMDs. In: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI 2017, pp. 1–5. ACM Press, New York (2017)
6. Clergeaud, D., Roo, J.S., Hachet, M., Guitton, P.: Towards seamless interaction between physical and virtual locations for asymmetric collaboration. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology - VRST 2017, pp. 1–4. ACM Press, New York (2017)
7. Goodman, E., Kuniavsky, M., Moed, A.: Observing the User Experience: A Practitioner's Guide to User Research. Morgan Kaufmann, Burlington (2012)
8. Gugenheimer, J., Stemasov, E., Frommel, J., Rukzio, E.: ShareVR: enabling co-located experiences for virtual reality between HMD and non-HMD users. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI 2017, pp. 4021–4033. ACM Press, New York (2017)
9. Jennett, C., Cox, A.L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A.: Measuring and defining the experience of immersion in games. Int. J. Hum. Comput. Stud. **66**, 641–661 (2008)
10. King, N., Horrocks, C.: Interviews in Qualitative Research. SAGE Publications Inc., London (2010)
11. Kraus, M., Kibsgaard, M.: A classification of human-to-human communication during the use of immersive teleoperation interfaces. In: Proceedings of the 2015 Virtual Reality International Conference on ZZZ - VRIC 2015, pp. 1–8. ACM Press, New York (2015)
12. Krueger, R.A., Casey, M.A.: Focus Groups: A Practical Guide for Applied Research. SAGE Publications Inc., Thousand Oaks (2008)
13. Latoschik, M.E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., Botsch, M.: The effect of avatar realism in immersive social virtual realities. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology - VRST 2017, pp. 1–10. ACM Press, New York (2017)

14. Narayan, M., Waugh, L., Zhang, X., Bafna, P., Bowman, D.: Quantifying the benefits of immersion for collaboration in virtual environments. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology - VRST 2005, pp. 78–81. ACM Press, New York (2005)
15. Nawahdah, M., Inoue, T.: Virtually dining together in time-shifted environment: KIZUNA design. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW 2013, p. 779. ACM Press, New York (2013)
16. Nguyen, C., DiVerdi, S., Hertzmann, A., Liu, F.: CollaVR: collaborative in-headset review for VR video. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST 2017, pp. 267–277. ACM Press, New York (2017)
17. Rahimian, F.P., Ibrahim, R.: Impacts of VR 3D sketching on novice designers' spatial cognition in collaborative conceptual architectural design. Des. Stud. **32**, 255–291 (2011)
18. Sra, M.: Asymmetric design approach and collision avoidance techniques for room-scale multiplayer virtual reality. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST 2016 Adjunct, pp. 29–32. ACM Press, New York (2016)
19. Zaman, C.H., Yakhina, A., Casalegno, F.: nRoom: an immersive virtual environment for collaborative spatial design. In: Proceedings of the International HCI and UX Conference in Indonesia on CHIuXiD 2015, pp. 10–17. ACM Press, New York (2015)

Check for updates

Tool-Mediated HCI Modeling Instruction in a Campus-Based Software Quality Course

Christos Katsanos[1,2], Michalis Xenos[3(✉)], and Nikolaos Tselios[4]

[1] Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] HCI Group, Electrical and Computer Engineering Department, University of Patras, Patras, Greece
ckatsanos@ece.upatras.gr
[3] Computer Engineering and Informatics Department, University of Patras, Patras, Greece
xenos@ceid.upatras.gr
[4] Educational Sciences and Early Childhood Education Department, University of Patras, Patras, Greece
nitse@ece.upatras.gr

Abstract. The Keystroke Level Model (KLM) and Fitts' Law constitute core teaching subjects in most HCI courses, as well as many courses on software design and evaluation. The KLM Form Analyzer (KLM-FA) has been introduced as a practitioner's tool to facilitate web form design and evaluation, based on these established HCI predictive models. It was also hypothesized that KLM-FA can also be used for educational purposes, since it provides step-by-step tracing of the KLM modeling for any web form filling task, according to various interaction strategies or users' characteristics. In our previous work, we found that KLM-FA supports teaching and learning of HCI modeling in the context of distance education. This paper reports a study investigating the learning effectiveness of KLM-FA in the context of campus-based higher education. Students of a software quality course completed a knowledge test after the lecture-based instruction (pre-test condition) and after being involved in a KLM-FA mediated learning activity (post-test condition). They also provided post-test ratings for their educational experience and the tool's usability. Results showed that KLM-FA can significantly improve learning of the HCI modeling. In addition, participating students rated their perceived educational experience as very satisfactory and the perceived usability of KLM-FA as good to excellent.

Keywords: Project-based learning · Educational tool · Learning activity
Web form design · Keystroke level model · Fitts' law

1 Introduction

The integration of core human computer interaction (HCI) concepts such as interface design and evaluation into the computer science/engineering curriculum is not well anticipated. Such an integration should balance effectively HCI theory instruction as well as hands-on experience. Nevertheless, at the end of the learning process the students should be able to effectively use HCI knowledge to design and evaluate software. As a

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 114–125, 2018.
https://doi.org/10.1007/978-3-319-91238-7_10

result, teaching of independent, isolated conceptual entities without offering a coherent conceptual context to provide the student the ability to create meaningful associations and abstractions and subsequently apply the obtained knowledge to increase quality of interaction should not be considered as effective [1, 2].

The keystroke level model (KLM) [3] provides an accurate estimation of the time required to perform an interaction task for an expert and error-free modeled user. The designer can employ this model to choose among design alternatives and optimize a user interface in terms of task completion time. As a result, an efficient KLM learning activity should provide the opportunity to the student to: (a) learn the basic aspects of the KLM model, (b) to apply efficiently the KLM to calculate the time to complete a task with a given user interface, (c) to understand the impact of specific design decisions on usability, and (d) to understand the differences of alternative interaction strategies. Web form filling tasks provide a rather good learning context to introduce students to KLM modeling because people are typically experienced in performing themselves such tasks and form filling tasks are well-organized and executed in a serial manner.

A tool which automates the process of calculating the time required to complete a form may greatly assist the educational process. It can provide the ability to test alternative scenarios and to reflect upon specific design approaches without the need to carry out tedious and repetitive calculations. The Keystroke Level Model-Form Analyzer (KLM-FA)[1] has been introduced [4, 5], as a practitioner's tool to facilitate web form design and evaluation based on established HCI predictive models, namely the Keystroke Level Model (KLM) [3] and the Fitts' Law [6]. Such models constitute a core teaching subject of HCI courses, as well as of courses like software design, interaction design and software quality. In a previous work [7], it was found that KLM-FA can support teaching and learning of HCI modeling in the context of distance education. In this paper, we employ a pre-post research design in the context of traditional campus-based education. The tool is briefly presented hereinafter.

1.1 The KLM Form Analyzer Tool (KLM-FA)

KLM-FA is a Windows desktop application for web form design and evaluation. The tool automatically detects the form elements and presents them in an embedded web form preview. The user of the tool can select the form elements involved in the modeled task–they are auto-highlighted in the web form preview– and KLM-FA calculates the predicted time for task completion (see Fig. 1). The tool provides various profiles for modeled users depending on their age and typing skills following established KLM modeling conventions. In addition, KLM-FA can easily model various form filling interaction scenarios depending on input devices usage (keyboard, mouse) for filling form elements.

[1] Available at http://klmformanalyzer.weebly.com.

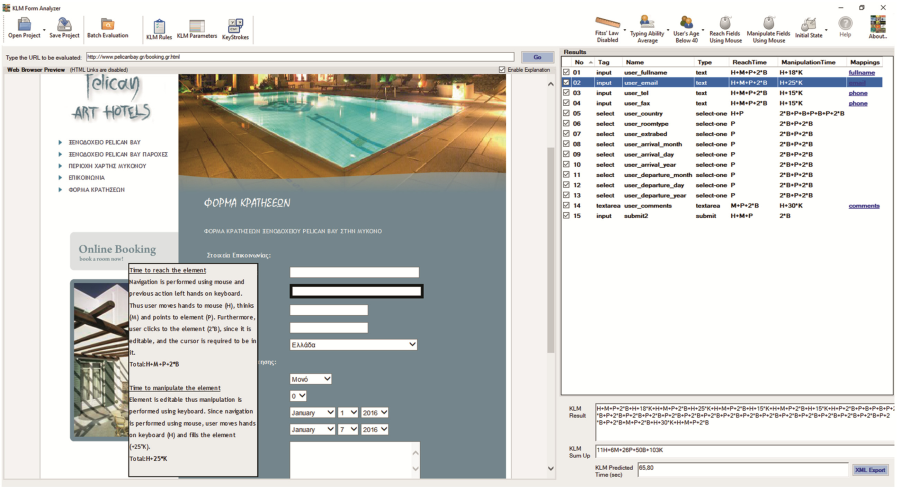


Fig. 1. The KLM-FA main user interface: an interactive web form preview shows the web form (left part) and an interactive table of modeling results shows KLM operators and time estimation per element and in total (right part). The “Enable Explanation” functionality displays a tooltip explaining the KLM modeling sequence next to the currently focused element.

KLM-FA is based on a two-phase modeling of every user action on a web form element: one has to first *reach* the form element and next to *manipulate* it. A typical usage scenario of the tool is the following. First, the designer inputs a URL of a web form either stored online in a web server or locally in the designer’s filesystem. Then, KLM-FA parses the webpage and identifies all forms and form elements. Subsequently, the designer selects all or some of the elements which are involved in the modeled user interaction task, specifies related parameters (e.g. whether the user is using the mouse to reach form elements) and may optionally change KLM modeling defaults (e.g. rules for placing mental operators).

Next, KLM-FA simulates the series of modeled user’s actions to first reach each element and then manipulate it. In this context, KLM operators are being detected and furthermore Fitts’ Law may be automatically applied for each of the simulated mouse movements to reach or manipulate an element. Then, a summarized result presents the series of simulated actions to complete the form filling task and estimates the required time to do so.

KLM-FA is useful for educational purposes, apart for professional design or evaluation practices, since it supports the following educational scenarios [7]:

- It provides *step-by-step KLM application*. The student can select a form element either from the web browser or from the sequential list and observe the specific actions to reach and manipulate the element. As a result, the student can distinguish what element of the form contributes to what extent in the total KLM calculation.
- It provides *concrete feedback* to the user. The KLM-FA user can alter some of the assumptions describing the modeled user (i.e., age, typing ability) and observe the difference.

- It provides the ability to *test alternative designs*. Thus, the user can investigate the design approaches which lead to an optimum result, in terms of time to complete the form-filling task.
- It provides the ability to *explore the impact of form elements' positioning and size*. By using Fitts' law to compute the time to point an element, the user may observe the impact of element's size and relative positioning into the form.
- It provides the ability to *test alternative form manipulation strategies* such as keyboard usage, mouse usage or a combination of them.

This paper reports a pre-post study that investigates the learning effectiveness of KLM-FA (post-test) compared to lecture-based instruction (pre-test) in the context of campus-based higher education. In specific, the research questions investigated by this study are the following:

- RQ1: Is there any effect of the KLM-FA mediated learning activity on students' learning performance in the context of campus-based higher education?
- RQ2: Were students with lower pretest score benefited from the KLM-FA mediated learning activity at least to the same extent as students with higher pretest score?
- RQ3: Did students find the KLM-FA a useful educational tool?
- RQ4: Did students find the KLM-FA a usable tool?

The rest of the paper is organized as follows. The method of the study is presented in the next section, followed by the findings which are analyzed according to the aforementioned research questions presented. Finally, the paper concludes with a discussion of the obtained findings as well as directions for future work.

2 Methodology

The study presented in this paper provides insights on the students learning gain (academic performance), as well as on the students' perceived educational experience. In addition, it provides the results of the students' assessment of the KLM-FA, the tool used to mediate the learning activity. A one group pre-post design was adopted.

The study took place in the context of campus-based classroom education and in specific in the course named "*CEID_NE5577: Software Quality Assurance and Standards*". This is an elective course offered to the students of the Computer Engineering and Informatics Department (CEID) at the University of Patras, during the first semester of their final (5th) year of studies. CEID is a 5-year B.Sc. degree with an Integrated M.Sc., corresponding to 300 European Credit Transfer System (ECTS) units.

The CEID_NE5577 course includes 14 lectures and 5 compulsory essays and offers 5 ECTS units to the students. The essays are graded and contribute to 50% of the final course grade, while a passing grade in all essays is a prerequisite for participating in the final exams for the other 50% of the course grade. Students are introduced to basic HCI concepts during lectures 2 to 5. These lectures include, inter alia, the presentation of the topics related to this study: The Human Information Processor model, the KLM and the Fitts' Law.

2.1 Participants

There were 108 students enrolled in the CEID_NE5577 course. However, only 22 of them attended the lectures, participated in all essays and successfully completed the course. These 22 students were aged from 21 to 26 (mean age = 22.1, SD = 0.97) and 7 were female. Although the lecture attendance in CEID is not obligatory, this course had a high attendance rate with approximately 17 out of the 22 students attending each lecture (mean number of students per lecture = 17.2, SD = 3.8) and an equally high participation on the course section of the eClass, which is the University of Patras Learning Management System (LMS).

The eClass LMS offers means of asynchronous communication (e-messages and e-fora) and the appropriate infrastructure for receiving, submitting and grading the essays. A total of 287 students' messages were recorded in the eClass (mean number of messages per student = 13.0, SD = 5.2) which is interpreted as a very active online presence, especially since CEID_NE5577 is a campus-based and not a distance course. All communication between students and the professor was through eClass and all essays were submitted and graded on the eClass.

2.2 Materials

The KLM and Fitts' Law are part of the second essay, in which the students were required to demonstrate their knowledge, by solving some modeling problems in interaction design using KLM-FA. The students had 15 days to complete the essay.

Regarding the KLM part of the essay, students were provided with the KLM-FA tool and a short (6 m 23 s) video available on YouTube further explaining the use of the tool. They were asked to use KLM-FA to evaluate signup forms of diverse complexity–from easy signup forms of hotels as shown in Fig. 1, to more complex forms of online services– using various parameters of the tool (i.e. user ages, typing skill), as well as using Fitts' Law to model the pointing device movement time. Students were also asked to conduct a "pen-and-paper" KLM modeling for a non-working prototype of a, rather complex, form.

All students were asked to complete a knowledge test (pre-test) after the end of the lectures and before downloading their essay description from the course LMS. They had access to their essay description only after completing the knowledge test. The knowledge test included all issues related to KLM and comprised of 14 multiple-choice questions with four answer options each. The same test (post-test) was offered to them after they had submitted their essays and completing it was required to formally finalize their essays into the course LMS. They were also asked to complete three additional scales as part of the post-test: (a) a 5-items scale rating their educational experience with KLM-FA from 1 to 5, (b) the standardized System Usability Scale [8] provided in participants' native language [9, 10], and (c) the 7-point adjective rating question [11] with wordings from "worst-imaginable" to "best-imaginable".

2.3 Procedures

The learning material, including the lectures slides, KLM-FA demo video and the essay description were announced on the eClass. All the questionnaires were also incorporated in the same LMS. After the deadline, the 22 submitted essays were graded by the professor and indicative solutions and results were uploaded in eClass. Finally, each student received personal remarks on their submitted essay.

The collected data were organized and preprocessed using Microsoft Excel 365 ProPlus and were analyzed using IBM SPSS Statistics v20.0. The materials offered to students using the eClass LMS were available to all enrolled students until the end of the semester.

3 Results

First, reliability analysis was conducted for the questionnaires used in the study. To this end, the Cronbach's alpha measure of internal consistency was used [12]. The knowledge test had low reliability with all items included; Cronbach's $\alpha = 0.402$, N = 14 items. This was attributed to an ambiguity in the way the content related to two questions was presented in the lecture slides and in the references at the course book. Removing these two questions resulted in a scale with adequate reliability; Cronbach's $\alpha = 0.782$, N = 12 items. The educational experience scale had very good internal consistency; Cronbach's $\alpha = 0.873$, N = 5 items. SUS is a standardized scale [9, 10, 13–15] and had also adequate reliability for our dataset; Cronbach's $\alpha = 0.760$, N = 10 items.

Following the rationale reported in [16], we produced a composite variable for the normalized learning gain defined as the difference between posttest score and pretest score ("observed gain" [16]) divided by the difference between the max possible score and the pretest score ("amount of possible learning that could be achieved" [16]).

Table 1 presents descriptive statistics of the dependent variables measured in this study. In all subsequent statistical analyses, the effect size r was calculated according to the formulas reported in [17].

Table 1. Descriptive statistics of the dependent variables in this study. Sample size N = 22 university students in campus-based education.

Variable	M	Mdn	SD	95% CI
Pretest score (0–100)	62.9	66.7	15.4	[56.0, 69.7]
Posttest score (0–100)	72.0	75.0	13.0	[66.2, 77.7]
Normalized learning gain (%)	19.1	18.3	37.0	[2.7, 35.5]
KLM-FA educational experience rating (1–5)	4.0	4.0	0.8	[3.7, 4.3]
SUS score for KLM-FA (0–100)	82.0	85.0	10.2	[77.6, 86.6]
Usability adjective rating for KLM-FA (1–7)	5.3	5.0	0.8	[5.0, 5.7]

3.1 RQ1: Learning Performance and KLM-FA

A two-tailed dependent samples t-test showed that the difference between students' pretest and posttest scores in the knowledge test was statistically significant; $t(21) = 2.890$, $p = 0.009$, $r = 0.533$. This large observed effect size [18] demonstrates the learning effectiveness of KLM-FA in the context of campus-based higher education. A parametric test was used, because Shapiro-Wilk analysis found that the distribution of the differences in the posttest and pretest scores did not deviate significantly from a normal distribution; $W(22) = 0.939$, $p = 0.187$.

However, a Kolmogorov-Smirnov test of normality showed that this distribution of differences in scores was significantly non-normal; $D(22) = 0.202$, $p = 0.02$. Although the Shapiro-Wilk test is more reliable for small sample sizes [17], we also conducted a non-parametric analysis of the same data due to the contradicting results in the normality tests. In agreement to the parametric test, a two-tailed Wilcoxon signed rank test also found that students improved significantly their scores in the knowledge test after using KLM-FA; $z = 2.601$, $p = 0.009$, $r = 0.390$.

3.2 RQ2: Learning Gain for Students with Low and High Pretest Score

A median split analysis was conducted to investigate whether students with low initial performance achieved higher learning gain compared to students with high initial performance.

To this end, students with pretest score lower to the median (Mdn = 66.7) were assigned in the low pretest performance condition (N = 10), whereas the rest were assigned in the high pretest performance condition (N = 12). Table 2 presents pretest score, posttest score and normalized learning gain grouped by students' initial performance.

Table 2. Descriptive statistics of study variables grouped by students' initial performance.

Group	Variable	M	Mdn	SD	95% CI
Low initial score	Pretest score (0–100)	48.1	50.0	9.1	[41.1, 55.1]
High initial score	Pretest score (0–100)	74.2	75.0	8.7	[68.4, 80.1]
Low initial score	Posttest score (0–100)	66.7	66.7	9.3	[59.5, 73.8]
High initial score	Posttest score (0–100)	80.3	83.3	9.3	[74.0, 86.6]
Low initial score	Normalized learning gain (%)	34.0	40.0	21.1	[17.7, 50.2]
High initial score	Normalized learning gain (%)	21.2	0.0	27.2	[2.9, 39.5]

A two-tailed independent samples t-test found an effect of student's initial performance on their post-test score; $t(20) = 2.589$, $p = 0.018$. A parametric test was used, because both the assumptions of normality and homogeneity of variance were not violated; Shapiro-Wilk tests, $p > 0.05$ and Levene's test, $F(1,20) = 0.157$, $p = 0.696$ respectively. In addition, a two-tailed Man-Whitney U test found no effect of students' initial performance on their normalized learning gain; z = 1.292, p = 0.197. A non-parametric test was selected, because the assumption of normality was violated for the high initial performance group; W(12) = 0.787, p = 0.006.

These results tend to support that students of lower initial performance improved significantly more their posttest score compared to students of higher initial performance. However, students of both low and high initial performance were equally benefited from the KLM-FA activity in terms of their normalized learning gain.

3.3 RQ3: Educational Experience with KLM-FA

Participating students rated their learning experience with KLM-FA in the post-test questionnaire. Table 3 presents descriptive statistics of these ratings per question and overall.

Table 3. Descriptive statistics of students' self-reported ratings of their educational experience with KLM-FA.

Question (1: strongly disagree; 5: strongly agree)	M	Mdn	SD	95% CI
Q1. The KLM-FA helped me to understand the KLM model and Fitts' Law	3.9	4.0	0.9	[3.5, 4.3]
Q2. During the activity, I am satisfied with my learning progress and effectiveness	3.7	4.0	0.9	[3.3, 4.1]
Q3. I think that KLM-FA is useful as an educational tool	4.1	4.0	0.9	[3.7, 4.5]
Q4. I would recommend KLM-FA to a colleague or friend who wants to learn the KLM and Fitts' Law	4.0	4.0	1.0	[3.5, 4.4]
Q5. I would recommend KLM-FA to a colleague or friend who wants to learn how to design web forms or evaluate their usability	4.4	5.0	0.9	[4.0, 4.8]
Overall scale (Cronbach's α = 0.873)	4.0	4.0	0.8	[3.7, 4.3]

Participants self-reported ratings about their learning experience with KLM-FA were rather high (M = 4.0, SD = 0.8). In specific, students agreed that KLM-FA helped them to understand the KLM model and Fitts' Law (M = 3.9, SD = 0.9) and that their perceived learning progress and effectiveness was satisfactory (M = 3.7, SD = 0.9). Students also found KLM-FA to be a useful educational tool (M = 4.1, SD = 0.9). The collected data suggest that they would probably recommend KLM-FA to colleagues or friends who want to be educated on established HCI models (M = 4.0, SD = 1.0) and web form design (M = 4.4, SD = 0.9). These results are in agreement with perceived educational experience ratings provided for KLM-FA by distance education university students in our previous work [7].

Correlation analysis found that students' education experience ratings and their grade in the KLM essay were not significantly associated; $r_s = -0.360$, p = 0.100. Spearman's coefficient was used, because the assumption of normality was violated by both variables; W(22) = 0.903, p = 0.035 and W(22) = 0.828, p = 0.001 respectively. This finding provides tentative support that students agreed that their educational experience was good without considering their academic performance in the essay.

In addition, students provided rather positive comments for their educational experience with KLM-FA in a related open-ended question of the post-test questionnaire. For instance, one student mentioned that "*It [KLM-FA] helped me understand the complexity of web forms design and factors that affect their usability*" and another mentioned that "*The 'Enable Explanation' feature helped me understand the calculation of the times to reach a field and to manipulate a field*".

In sum, these findings demonstrate that KLM-FA was perceived by the students as a useful educational tool that fits well into the educational context of the campus-based software quality course.

3.4 RQ4: Perceived Usability of KLM-FA

After interacting with KLM-FA, the participating students completed the SUS questionnaire and the adjective rating scale, both measures of a system's perceived usability.

KLM-FA received a mean SUS score of 82.0 (SD = 10.2). According to a dataset of nearly 1000 SUS surveys [11], this means that students found KLM-FA as "Good to Excellent" (SUS score from 71.4 to 85.5) in terms of perceived usability. Students' usability adjective ratings were also rather high (M = 5.3, SD = 0.8), confirming that KLM-FA was perceived from "Good" (corresponds to 5) to "Excellent" (corresponds to 6). In a previous study with distance education students [7], KLM-FA was also perceived as "Good to Excellent" with a mean SUS score of 73.6 (SD = 13.2) and similar usability adjective ratings (M = 5.4, SD = 0.8).

In the post-test questionnaire, students were also offered the chance to write down three positive and three negative characteristics regarding the KLM-FA tool. A total of 41 positive and 16 negative characteristics were reported respectively. Thematic analysis of participants' answers resulted in groupings. KLM-FA was found to be a usable and easy to use tool (8 students), provide quick and accurate results (8 students), educationally valuable (8 students), flexible and parametrizable (7), simple and understandable (6) and useful for Fitts' Law calculations (3). By contrast, students reported that KLM-FA needs: (a) improvements in user interaction (6 students), such as adding shortcuts for frequently used functionality, (b) changes in KLM modeling (5 students), such as modeling user errors, (c) enrichment in modeling explanation (3 students), and (d) additional flexibility by not requiring an implemented DHTML form (2 students). Finally, the most frequently-mentioned students' general suggestion for further KLM-FA improvement, as mentioned in a separate question, was to make it available in different operating systems other than Windows (3 students).

All in all, these findings demonstrate that KLM-FA was perceived as a usable tool by the students of the software quality course.

4 Discussion and Conclusions

The goal of this paper was to examine the effectiveness of a tool mediated learning activity in the context of a campus-based HCI modeling instruction. A one group pre-posttest design was adopted. Students improved significantly their scores in the knowledge assessment test after using KLM-FA, the tool that mediated the learning activity. Their performance was significantly improved and jumped, on average from 62.9% to 72%. Also, there were no significant differences in the students' learning gain of different initial performance (low versus high initial performance). The tool which has been adopted (KLM-FA) was perceived as a useful and usable tool by the students regardless of their academic performance. These findings are in agreement with our previous work both on using KLM-FA in the context of distance education [7] and implementing other tool-mediated activities [19, 20] to support HCI instruction.

However, this study is not without limitations. First, the sample was rather small and therefore the confidence interval for the results is rather wide. Moreover, an experiment with a control group using paper and pencil and a treatment group using the KLM-FA tool will help us to examine in more detail the exact learning phenomena taking place during the activity. In addition, utilization of learning analytics methods [21] to examine the students' low level actions and their possible relation with the learning outcome will provide a deeper understanding of the way the students benefited from the tool. Furthermore, exploration of other instructional design approaches which adopt technologies such as wikis [22, 23], combined with the introduction of KLM-FA in the HCI modeling instruction constitute an additional future goal. Finally, we also plan to conduct studies that monitor students' fixations [24, 25] and other physiological signals, such as skin conductance which is a reliable indicator of stress [26–29], in an attempt to evaluate and further improve the KLM-FA user experience.

References

1. Haan, G.D.: DevThis: HCI education beyond usability evaluation. In: Proceedings Chi Sparks (2011)
2. Nordahl, R., Serafin, S.: Using problem based learning to support transdisciplinarity in an HCI education. In: Proceedings of HCIed (HCI in Education) Conference, p. 94 (2008)
3. Card, S.K., Moran, T.P., Newell, A.: The keystroke-level model for user performance time with interactive systems. Commun. ACM **23**, 396–410 (1980)
4. Karousos, N., Katsanos, C., Tselios, N., Xenos, M.: Effortless tool-based evaluation of web form filling tasks using keystroke level model and fitts law. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 1851–1856. ACM, New York (2013)
5. Katsanos, C., Karousos, N., Tselios, N., Xenos, M., Avouris, N.: KLM form analyzer: automated evaluation of web form filling tasks using human performance models. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013. LNCS, vol. 8118, pp. 530–537. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40480-1_36
6. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. **47**, 381 (1954)
7. Katsanos, C., Tselios, N., Karousos, N., Xenos, M.: Learning web form design by using the KLM form analyzer: a case study. In: Proceedings of the 19th Panhellenic Conference on Informatics, pp. 44–49. ACM, New York (2015)
8. 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. Taylor and Francis, London (1996)

9. Katsanos, C., Tselios, N., Xenos, M.: Perceived usability evaluation of learning management systems: a first step towards standardization of the System usability scale in Greek. In: 2012 16th Panhellenic Conference on Informatics (PCI), pp. 302–307 (2012)
10. Orfanou, K., Tselios, N., Katsanos, C.: Perceived usability evaluation of learning management systems: empirical evaluation of the system usability scale. Int. Rev. Res. Open Distrib. Learn. **16**, 227–246 (2015)
11. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. **4**, 114–123 (2009)
12. Nunnally, J., Bernstein, I.: Psychometric Theory. McGraw-Hill, Humanities/Social Sciences/Languages (1994)
13. Bangor, A., Kortum, P., Miller, J.: An empirical evaluation of the system usability scale. Int. J. Hum.-Comput. Interact. **24**, 574–594 (2008)
14. Kortum, P., Bangor, A.: Usability ratings for everyday products measured with the system usability scale. Int. J. Hum.-Comput. Interact. **29**, 67–76 (2013)
15. Lewis, James R., Sauro, J.: The factor structure of the system usability scale. In: Kurosu, M. (ed.) HCD 2009. LNCS, vol. 5619, pp. 94–103. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02806-9_12
16. Nelson, L., Held, C., Pirolli, P., Hong, L., Schiano, D., Chi, E.H.: With a little help from my friends: examining the impact of social annotations in sensemaking tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1795–1798. ACM, New York (2009)
17. Field, A.P.: Discovering Statistics Using SPSS. SAGE, Los Angeles, [Calif.]; London (2009)
18. Cohen, J.: A power primer. Psychol. Bull. **112**, 155–159 (1992)
19. Katsanos, C., Tselios, N., Tsakoumis, A., Avouris, N.: Learning about web accessibility: a project based tool-mediated approach. Educ. Inf. Technol. **17**, 79–94 (2012)
20. Katsanos, C., Tsakoumis, A., Avouris, N.: Web accessibility: design of an educational system to support guidelines learning. In: Proceedings of PCI 2009, pp. 155–164, Corfu, Greece (2009)
21. Kotsiantis, S., Tselios, N., Filippidi, A., Komis, V.: Using learning analytics to identify successful learners in a blended learning course. Int. J. Technol. Enhanced Learn. **5**, 133–150 (2013)
22. Altanopoulou, P., Tselios, N., Katsanos, C., Georgoutsou, M., Panagiotaki, A.: Wiki-mediated activities in higher education: evidence-based analysis of learning effectiveness across three studies. Educ. Technol. Soc. **18**, 511–522 (2015)
23. Tselios, N., Altanopoulou, P., Katsanos, C.: Effectiveness of a framed wiki-based learning activity in the context of HCI education. In: Proceedings of PCI 2011, pp. 368–372. IEEE Computer Society, Kastoria (2011)
24. Katsanos, C., Tselios, N., Avouris, N.: Evaluating website navigability: Validation of a tool-based approach through two eye-tracking user studies. New Rev. Hypermedia Multimed. **16**, 195–214 (2010)
25. Poole, A., Ball, L.J.: Eye tracking in Human-Computer Interaction and usability research: current status and future prospects. In: Encyclopedia of Human Computer Interaction, pp. 211–219. Idea Group Reference (2006)
26. Liapis, A., Karousos, N., Katsanos, C., Xenos, M.: Evaluating user's emotional experience in HCI: the PhysiOBS approach. In: Kurosu, M. (ed.) HCI 2014. LNCS, vol. 8511, pp. 758–767. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07230-2_72
27. Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M., Karousos, N.: Recognizing emotions in human computer interaction: studying stress using skin conductance. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (eds.) INTERACT 2015. LNCS, vol. 9296, pp. 255–262. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22701-6_18

28. Lunn, D., Harper, S.: Using galvanic skin response measures to identify areas of frustration for older web 2.0 users. In: Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A), pp. 34:1–34:10. ACM, New York (2010)
29. Liapis, A., Katsanos, C., Sotiropoulos, D.G., Karousos, N., Xenos, M.: Stress in interactive applications: analysis of the valence-arousal space based on physiological signals and self-reported data. Multimed. Tools Appl. **76**, 5051–5071 (2017)

Check for updates

Approaches to Interface Icon Classification

Teemu Korpilahti(✉)

Aalto University, Helsinki, Finland
Teemu.korpilahti@aalto.fi

Abstract. Various attempts have been made at creating a classification for user interface icons, with some focusing on their pictorial presentation, and others on the signs' relation to their intended meaning. This paper provides a review of the existing classification systems, and discusses their strengths and weaknesses. Based on this review, we then propose an alternative practical approach to icon classification, which is aimed towards designers of user interface icons rather than the research community, and evaluate its usability based on data gathered from two online surveys.

Keywords: Interface icon · Taxonomy · Classification · Semiotics · Semiology

1 Introduction

In this age of digitalisation, we are increasingly surrounded by devices and interfaces. The design of the user experience has become a key asset in the competition for consumers. An integral, but underrated, part of the experience, is the design of the interface icons. Many attempts have been made to create a classification for user interface icons, but none of them sufficiently for the benefit of the icon designer. The field of research is like a semantic puzzle, where researchers invent their own terms to describe the same things, or use similar terms to describe different relations. What is needed is a more concrete and practical system. One that can be adopted by designers, even without a very profound understanding of semiotic theory.

2 Semiotics or Semiology

The classification of signs in general is usually either related to semiotics or semiology, or both. Although these two terms are sometimes mixed with one another, they are based on two different models. While both were developed during the same time period and handle the same subject matter, semiology was created by Ferdinand de Saussure [1], and is connected to linguistics. Semiotics, on the other hand, was created by Charles Sanders Peirce [1] as part of his full philosophical system, composed of phaneroscopy (phenomenology), normative sciences, and metaphysics.

Following the Saussurean terminology, the signifier and the signified are both components of a sign [2]. For Peirce, on the other hand, a semiosis involves an operation of three subjects, a sign, its object and its interpretant. Furthermore, according to Peirce

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 126–137, 2018.
https://doi.org/10.1007/978-3-319-91238-7_11

all signs can be classified into **icons, symbols** and **indexes** based on the signs relation to the referent [3]. In this traditional system, an icon is limited to a representation that resembles its object, an index carries an actual connection to its object (e.g. a sign of a telephone on the door of a phone booth), and a symbol in contrast has no visual connection to its object, and is an abstract sign that can only be understood through learning its meaning. In the context of user interface icon design the relevance of classifying a sign as an index is somewhat irrelevant, since it could be argued that there is no variation in the level of connectedness between a referent and a sign.

In describing the relationship between the signifier and the signified, *arbitrary* is one of the terms Saussure uses. Another closely related term used in linguistics is *unmotivated.* These two terms describe two different levels of logical relation. A system is arbitrary, when its signs are founded not by convention, but by unilateral decision. Similarly, a sign is unmotivated, when the relation between its signified and signifier is not analogical. According to this logic, it is possible to have systems and icons that are at the same time arbitrary and motivated, and others which are non-arbitrary and unmotivated. For example, if a designer decided to use an image of a bloodhound to signify the search function, the relation would be arbitrary but motivated. In contrast, the stop icon is a good example of a non-arbitrary and unmotivated sign, and its use is heavily based on convention, whereas the relation of its appearance to the stop function is logically difficult to explain.

3 The Classification of Icons

As is widely known, in 1973 Xerox released the first computer system with a graphical user interface, introducing bitmap graphics, menus and icons, along with the desktop metaphor with its trash can, folder, file and printer icons. The first attempts to create a classification or taxonomy of icons started in the 1980's. Most of these classification systems are to some extent related to Peirce's semiotics. One main difference, however, is that in these systems the "index" sign type usually isn't a top level classification. Additionally, what Peirce classified as icon is commonly divided into two or more classes or subclasses based on the level of abstraction or the type of logical connection between the signifier and signified.

One of the first articles on icon classification is Keith Lodding's iconic interfacing [4], published in 1983. According to Lodding, at the time of the writing, the era of graphical interfaces had arrived, but no unique discipline of "iconic communication" existed. Lodding recognized three types of interface icons: **representational, abstract, and arbitrary.** Representational icons serve as an example for a general class of objects. Abstract icons attempt to present a concept through the use of metaphor, and by reducing the image to its essential elements that carry the intended concept. Arbitrary icons are "invented" and assigned a meaning, which must then be learned. Interestingly, Lodding also notes that once arbitrary icons are entered into use, they appear to enjoy a long lifetime when compared to either representational or abstract icons.

Webb et al. [5] have also suggested three main categories for visual icon classification. Their classifications are: **picture, symbol,** and **sign.** Picture is a realistic depiction of a system object or function, and as such, pictures are the most detailed forms of icons, making them easiest to interpret and remember. Symbols emphasize critical features by analogy or symbolism, and therefore, they are simplified and are affected the most by context. It should, however, be noted that the use of the term symbol or symbolic by Webb et al. is contradictory to Peirce, who used the same term to describe an arbitrary relation. The third category, defined as sign, is used instead by Webb et al. to describe icons with no intuitive connection between the icon and its referent. Signs are therefore abstract and simple, but their association with their referent must be learned.

Yet another icon classification has been proposed by Rogers [6], with the aim of discovering a primitive syntax and semantics for icons, which could then be used as a basis for designing a set of icons for a given application. According to Rogers, a common problem with iconic interfacing is the different meanings that can be attributed to a single icon. Verbal languages have syntactic and semantic rules which help to disambiguate their meaning. Pictorial languages do not yet have equivalent rulesets. Paradoxically, while pictorial communication could potentially be universally understood, it lacks the rules to guide this process. As such, context is a key factor in disambiguating the meanings of icons. Rogers presents Hungarian toilet signs as an example, where a stiletto shoe is used to represent women or men. Because of the user's prior knowledge of where to find a toilet in a restaurant, and the fact that there are generally separate toilets for men and women, these kinds of signs are immediately recognisable. Indeed, many restaurants and pubs use such signs, often in a funny way, to represent male and female toilets.

Rogers also identifies four ways in which interface icons can refer to their referent. **Resemblance icons** depict the referent through an analogous image, **exemplar icons** serve as a typical example for a general class of objects (e.g. a knife and fork represent a restaurant), **symbolic icons** show the relationship between the referent and referrer in a more abstract or metaphoric manner (e.g. image of a wine glass with a fracture conveys the concept of fragility), and **arbitrary icons** bear no resemblance to the referent, and therefore, their associations must be learned.

Furthermore, Rogers recognizes the usefulness of icon-based interfaces, and proposes the development of a taxonomy and grammar for icons to assists designers with overcoming the challenges of choosing when to use icons and how to design them.

Similarly, Lidwell et al. [7] base their classification for iconic presentation on the work of Rogers in their Universal Principles of Design. They just replace the terms **resemblance icons** with **similar icons** and **Exemplar icons** with **example icons**.

Finally, Purchase [8] proposes a semiotic system for icons based on Peirce's definitions, in which she introduced three classifications: **symbolic, concrete-iconic,** and **abstract-iconic.** Symbolic icons here are arbitrary in their relation to the concept they signify. Note that Rogers' classification used the term **arbitrary** and Webb et al. **sign** for roughly the same purpose. Concrete-iconic icons can be seen as identical to the concept that they represent. Abstract-iconic icons are similar, but not identical to the concept, and lie higher up the abstraction scale.

4 Earcons and Auditory Icons

In one of the earliest articles on the use of auditory icons in computer interfaces, published in 1986, William Gaver proposes that icons could be categorized as either **nomic, symbolic** or **metaphorical** [9]. Gaver uses the term nomic to describe an icon's relation to its referent, which is direct and descriptive. Symbolic is used for icons with an arbitrary relation to their referent; similar to Peirce's and Purchase's use of the term, but contradictory to Webb et al. Metaphorical icons can be either structure-mappings or metonymic mappings. Structure-mappings make use of similarities between the referent and the icon (e.g. an image of a tree can represent genealogy). While metonymic mappings use a part or a feature to represent the whole (e.g. a hissing sound can represent a snake).

A few years after the above publication, in 1989, Blattner et al. also studied auditory icons, or earcons, as they called them [10]. They suggested, that visual icons design techniques could be useful for earcons, even though at the time, there was little published material on visual icon design. They also suggested that the systems they developed for earcons could provide a useful basis for visual icons. A major difference they recognized between earcons and icons, is the fact that earcons are only used for informational purposes, whereas icons are both informational and interactive. Furthermore, multiple visual icons can be presented simultaneously, which is not possible for earcons. Blattner et al. recognized three main icon types: **abstract**, **representational**, and **semi-abstract.** Abstract icons consist of geometric shapes that are not recognizable as real-world things. Representational icons are pictures of familiar objects or operations. Whereas semi-abstract icons are simply everything in between, and they either consist of both abstract and representational elements, or they are representational images that are visually highly abstracted.

5 The Need for a New Classification

Wang et al. [11] recognized some limitations and weaknesses in the aforementioned systems for icons classification. According to them, in the existing classifications categories overlap, and different classifications use different terminologies, sometimes to describe the same concept. To highlight these problems, they compare nine existing taxonomies and their terminologies, and then they propose a classification of their own. In their taxonomy, they identify four icon categories: **concrete, abstract, combination,** and **alphabetic icons.** Concrete icons consist only of real-world objects. Abstract icons are the absolute opposite consisting only of shapes, metaphysical arrows and "arbitrary symbols"; the last of which is an interesting choice of terminology if you consider all the connotations that have been previously associated with the terms "symbol" and "arbitrary". Combination icons are a fusion of concrete and abstract elements. Alphabetic icons are any icons that show written characters regardless of what other elements they contain.

Similarly, Nakamura et al. [1] also identify problems with existing icon classifications, and argue that previous icon taxonomies are not granular enough to capture differences among representation strategies that can be visually subtle but semantically significant. Studies such as that of Blattner et al. focus only on the pictorial representation and overlook the relation between concept and representation, while others such as Rogers, Gaver and Lidwell et al. are more focused on the signs' relation to their intended meaning.

Nakamura et al. also identify and propose their own classification using three basic representation strategies: **visual similarity, semantic association** and **arbitrary convention.** Representation through visual similarity has no subcategories, but semantic association is divided to eight subcategories: comparison or contrast, exemplification, semantic narrowing, physical decomposition, temporal decomposition, body language, metaphor, and contiguity. Contiguity is further narrowed to physical contiguity, container, source, use, tool, cause or effect, and object. Arbitrary convention is also further divided to abstract, concrete and transposed conventions. Transposed convention refers to a situation where the convention is in the referent, not in the pictorial representation. This relation however is not explained in a very detailed manner.

6 A Taxonomy for Trademarks

Mollerup's Marks of Excellence [12], thoroughly studies trademarks, their design, functions and history. The function of trademarks is of course different to that of icons, but there are many similarities in the aspects of their design. Mollerup has also created a complete taxonomy system for trademarks. This system is derived from Aristotelian logic, following these five rules:

1. It must consist of classes that are distinct. The differences between the classes must be clear so that there is no room for misunderstanding as to which class an item belongs.
2. The characteristics on which the classes are based should be used consistently, and each step in the classification should be based on a single principle of division.
3. There should be no overlap between classes. Parallel (co-ordinate) classes should be exclusive.
4. Co-ordinate classes should be able to collectively cover all possible entries.
5. The classes should be relevant to the purpose of the taxonomy.

Using these logical rules, Mollerup proposes the following taxonomy.

1. Trademarks
    a. Graphic marks
        (1) **Picture marks**
            i. **Figurative marks**
                (a) **Descriptive marks**
                (b) **Metaphoric marks**
                (c) **Found marks**
            ii. **Non-figurative marks**
        (2) Letter marks
            i. Name marks

   (a) Proper names
   (b) Descriptive names
   (c) Metaphoric names
   (d) Found names
   (e) Artificial names
  ii. Abbreviations
   (a) Initial abbreviations
    A. Acronyms
    B. Non-acronym initial abbreviations
   (b) Non-initial abbreviations
b. Non-graphic marks

Icons rarely consist only of letters but sometimes letters and words may be a part of an icon. Therefore, branch 1.a.(1) (Picture marks), is the most relevant to the purpose of icon taxonomy. If we examine the subclasses it contains, figurative marks are marks that depict an object, and non-figurative marks are pictures in their own right. Non-figurative marks are marks that appear completely abstract. For instance, the stop icon by itself might be considered non-figurative. Also, icons whose graphical meanings are unclear to a user, might be considered non-figurative. The power icon for instance, is a logical graphical reference to zero (off) and one (on), but to some users it might seem completely abstract.

Descriptive marks refer directly to their object. The printer icon is a clear example of a descriptive icon. Metaphoric marks refer to their object through a shared quality. The file and thrash can are metaphoric icons. Found marks show something recognizable that is arbitrary to its object. The relationship of the symbol and its object might often have a historical explanation that is logical. Over time the connection is lost and the symbol becomes arbitrary. For instance, the floppy disk as an icon for the save function, is becoming arbitrary to younger users who have never used the floppy disk storage media, and as such, it might even be considered a found mark.

The major difference in Mollerup's system is the top level division to figurative and non-figurative marks. The semantic relation between the trademark and its object is only relevant to figurative marks in this system.

7 A Practical Approach on Icon Classification

This paper proposes an alternative approach on icon classification that addresses the concerns put forward by Nakamura et al., and builds on the same Aristotelian logic and aim for practicality as Mollerup's system. The proposed taxonomy is aimed towards designers of user interfaces and icons rather than the research community.

The semiotic or logical relationship, **logical** vs. **arbitrary**, is the top level of classification of the system. Logical icons are divided to two subclasses: **descriptive** and **metaphoric**, which further describe the relation of the icon to its object. This is similar to Mollerup's division of figurative marks. Arbitrary icons are divided into **concrete** and **abstract** subclasses. These subclasses are more concerned with the type of pictorial presentation, since there is no logical relationship to define. The purpose of developing this system was to evaluate whether the terminology of even such a basic classification can be understood and mastered by test subjects, who are not familiar with the field of research.

The proposed terminology can be explained in the following manner:

- **Concrete:** The icon is visually recognizable as an image that represents a real-world object. To put it more simply, it is possible to clearly name the object that image shows.
- **Abstract:** The icon is not visually recognizable as an image that represents a real-world object. The icon is just a combination of shapes that does not resemble any recognizable real-world object.
- **Logical:** There is a clear logical connection between the icon image and the functionality it represents. The icon is self-explanatory and can be interpreted with little previous knowledge.
- **Arbitrary:** There is no logical connection between the icon image and the functionality it represents. Any association between the image and its functionality must be learned.

As you may notice there is an overlap and direct conflict with Saussure's terminology. Even this proposed system is becoming entangled in the same semantic debate that was mentioned before. The term arbitrary here, is similar to what Saussure described as unmotivated, and logical equals to Saussure's motivated. Instead, Saussure used the term arbitrary to describe a system where the signs are not founded by convention, but by unilateral decision. For the purpose of designing user friendly interfaces with legible icons, the whole premise of making such unilateral decisions should only come to play, when the designer needs to introduce an icon for a completely new element or action of the interface. It is not as useful for the purpose of classifying existing interface icons (Fig. 1).

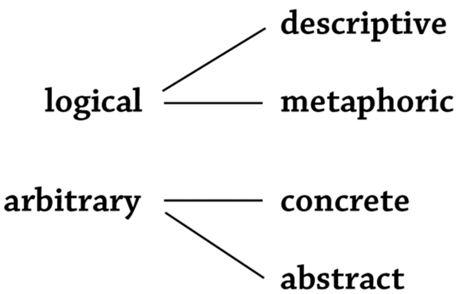


Fig. 1. Proposed taxonomy system.

8 Online Surveys

The functionality of the system was evaluated through two online surveys. The first survey tested two basic classifications questions that concerned both the semiotic relationship between the sign and its object, and the visual type of the sign. The results were published in full detail in a poster article for HCI Toronto 2016 [13].

The respondents were recruited from university students and staff, LinkedIn and Facebook GUI interest groups and the author's personal networks. The majority of the 119 participants were designers, design students and IT professionals. In all there were participants with 22 different nationalities ranging from United States and Germany, to Iran and China. The largest group (66%) were Finnish. 58% of the participants were male and 42% Female. The majority of the respondents (49%) were 30 to 40 years old, with the entire range being between 10 and 57 years.

The participants were asked to describe their skills in using digital devices on a five step range: very good, good, average, poor, or very poor. 66% of the respondents described their skills as very good, 28% good, and the rest 6% average. They were also asked how often they found it difficult to understand the meaning of interface icons or buttons. This was also measured on a five step range: very rarely, rarely, occasionally, often, or very often. 29% of the respondents replied very rarely, 40% rarely, 29% occasionally, and 2% often. So it seems, that even skilful users sometimes encounter problems in understanding interface icons.

Due to the relatively small sample size, it is not of great worth to compare small percentage differences in the classification of individual icons, but rather to have an overall understanding of how accurate and applicable such a classification can be (Fig. 2).

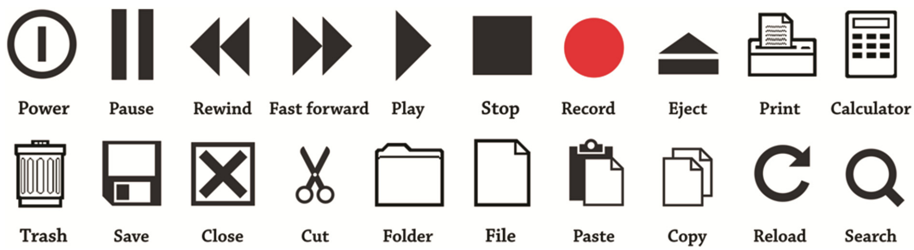


Fig. 2. The icon set that was used in the surveys.

The participants were asked to classify a set of twenty icons through the following questions:

1. Is the icon **abstrac**t or **concrete?**
2. Is the icon **logical** or **arbitrary?**

8.1 Importance of Context

The participants were also asked to shortly name what they thought each of the icons stands for, i.e. what is its meaning or function. The purpose of this question was to verify, that the user had correctly recognized the icon. It was clear from the answers, that some of the icons could be understood in a variety of ways, and others were not always recognized at all. The lack of context is one important factor in this. For instance, the magnifying glass was recognized as both a search and magnifying tool. Similarly, the cross icon that stands for closing or deleting, was also recognized as the symbol for irritating substances. The total number of icon classifications was 2380 (119 participants classified 20 icons). Out of these, there were 242 cases where the icon had not been correctly recognized. Therefore, 89.8% of the data was valid, and this part of it was analysed further.

8.2 Abstract and Arbitrary

In the icons that were classified as abstract, there appeared to be two clusters. One comprised of strongly arbitrary icons. The other cluster of icons was above 58% in the logical scale. This group consisted of the arrow icons of the survey: play, fast forward, rewind, and reload. It is clear, that arrows were considered as abstract representations. Yet they are so commonly used, that users intuitively understand their meaning. The origin of the arrow symbol most likely derives from the concrete archer's arrow object. So, the case seems to be, that the appearance of the arrow has just become so simplified and abstracted, that it is no longer considered concrete.

8.3 Arbitrary and Concrete

In this icon set there were no occurrences of icons that are clearly arbitrary and concrete at the same time. The paste icon comes closest to this, being 81% concrete and only 58,3% logical. It would seem, that icons are most commonly in this group, if the metaphor or descriptive relation of the icon (signifier) and its signified is weak from the beginning, or becomes unclear over time (Table 1).

Table 1. Summary of the first survey data.

Icon	%Concrete	%Logical	Icon	%Concrete	%Logical
Calculator	99,1	90,6	Play	14,4	58,5
Close X	2,3	36,8	Power	9,1	24,2
Copy	83,1	77,1	Printer	98,1	91,3
Cut	96,6	94,1	Record	9,1	20,2
Eject	10,7	36,9	Reload	13	59,3
Fast forward	14,3	65,5	Rewind	14,3	65,5
File	94	76,1	Save	94	70,9
Folder	98,2	92	Search	87,8	81,6
Paste	81	58,3	Stop	4,8	21,2
Pause	7	23,5	Trash can	99,2	97,5

8.4 Descriptive and Metaphoric

The second survey received 105 responses, 70% of which were from Finnish respondents. In all there were participants with 18 different nationalities. 67% of the respondents were male and 33% female. The largest age group of the respondents were 30 to 40 years old (39%), with the entire age range varying from 21 to 57 years. 68% of the participants described their skills as very good, 23% good, 8% average and 1% poor. To the question whether the respondents found it difficult to understand the meaning of interface icons or buttons, 28% replied very rarely, 40% rarely, 29% occasionally, 2% often and 1% very often.

In general, the demography was similar to the previous survey. The portion of male respondents was however 9% higher and the 30–40 age group was less dominant, 39% versus the previous 49%.

The total number of icon classifications was 2100 (105 participants × 20 icons). Out of these, there were 219 cases, where the icon had not been correctly recognized. Therefore, 89.6% of the data was valid, and this part of it was analyzed further.

In the second survey the respondents were asked to classify the same set of interface icons as **metaphoric, descriptive** or **not applicable.** The hypotheses was, that the icons that were classified as arbitrary in the first survey, should be classified as not applicable in the second survey, since there apparently is no logical relation between the icon and its object. The icons that were classified dominantly arbitrary were: **stop, record, power, pause, eject,** and **close.** All of these were indeed classified as dominantly not applicable. The only flaw in the logic was, that two additional icons (fast forward and rewind) were also classified as not applicable (Table 2).

Table 2. Icons classified as not applicable.

Icon	% Metaphoric	% Descriptive	% Not applicable
Stop	29.2	13.5	**57.3**
Power	34.4	10.0	**55.6**
Pause	33.0	13.6	**53.4**
Record	36.0	11.6	**52.3**
Eject	39.8	8.0	**52.3**
Fast forward	34.6	18.3	**47.1**
Rewind	36.3	16.7	**47.1**
Close	42.7	11.0	**46.3**

The icons that were classified as mostly descriptive were: print, calculator, folder, trash, copy, file, and play. Perhaps the most surprising result here was, that calculator, folder, and trash icons were seen as descriptive rather than metaphoric. The calculator application is clearly a part of the classic desktop metaphor, but perhaps the icon can be also considered as descriptive since it looks similar to the application itself. Interestingly also the play icon was the only playback icon in this group, as all the others (stop, pause, rewind, fast forward, eject) were classified as not applicable (Table 3).

Table 3. Icons classified as not descriptive.

Icon	% Metaphoric	% Descriptive	% Not applicable
Print	5.5	**91.2**	3.3
Calculator	11.5	**88.5**	0.0
Folder	37.5	**62.5**	0.0
Trash	41.2	**58.8**	0.0
Copy	47.4	**52.6**	0.0
File	46.8	**51.1**	2.1
Play	36.3	**47.1**	16.7

The metaphoric classification group didn't offer as many deviations. The group consisted of the icons: **search, save, paste, cut,** and **reload.** What is surprising, is the fact that cut and paste were a part of this group, but copy was seen as a descriptive rather than metaphoric icon (Table 4).

Table 4. Icons classified as metaphoric.

Icon	% Metaphoric	% Descriptive	% Not applicable
Search	**75.0**	22.6	2.4
Save	**60.8**	34.3	4.9
Paste	**58.3**	31.9	9.7
Cut	**56.9**	42.2	1.0
Reload	**54.9**	17.6	27.5

9 Discussion

Although there are logical patterns in the classification data, it is also apparent that the respondents found the task of classifying icons according to this terminology challenging. This was also clear from the open feedback that was received from the respondents. It might explain why the differentiation between the classes is not so decisive. Certainly, this type of classification can only be useful for a designer or researcher, who is more used to dealing with such logical relations, and even that can be questioned. The classification corresponds to Nakamura et al.'s aspiration for a system that takes into account both the type of pictorial representation, as well as the relation between the icon and its object. The level of granularity at present, is not as ambitious as that of Nakamura et al. Future research can add depth and granularity to the system, but it is perhaps not necessary, bearing in mind the original goal of creating a concrete and practical system that can be adopted by designers without a very profound understanding of semiotic theory and terminology.

10 Conclusions

The motivation for the creation of the existing classification systems has been to discover or create logical rulesets for the creation and interpretation of interface icons. What

Rogers [6] called a primitive syntax and semantics for icons, a taxonomy and grammar for icons, to assists designers with overcoming the challenges of choosing when to use icons and how to design them. The effectiveness and clarity of such systems has proven to be questionable, but such an approach is also insufficient to guide the design of icons and icon sets. The design process is also affected by other factors, such as the physiology and mechanics of the human eye, neural processes of how the brain interprets images, and cognitive psychology of how visual symbols are linked to non-visual concepts. Outside of these human centered factors there are other forces at play. Device and OS manufacturers, such as Apple, Microsoft, and Google provide their own interface guidelines that include specifications for icons. The never ending technical development of devices and displays affects what is necessary and possible in the reproduction of icon designs. Icons are also commonly extensions of a company's or product's brand. Icons and icon systems need to consistently follow a specific visual style that derives from brand guidelines. Still the stylistic changes need to be subtle, so that commonly used icons don't lose their recognizable connection to the original form. Further research needs to address how all these factors affect the way in which interface icons take shape through intentional design choices and environmental factors.

References

1. Nakamura, C., Zeng-Treitler, Q.: A taxonomy of representation strategies in iconic communication. Int. J. Hum. Comput. Stud. **70**, 535–551 (2012). https://doi.org/10.1016/j.ijhcs.2012.02.009
2. Eco, U.: Semiotics and the Philosophy of Language. Indiana University Press (1986)
3. Abdullah, R., Hübner, R.: Pictograms, Icons and Signs. Thames & Hudson Ltd., London (2006)
4. Lodding, K.N.: Iconic interfacing. IEEE Comput. Graph. Appl. **3**, 11–20 (1983). https://doi.org/10.1109/MCG.1983.262982
5. Webb, J.M., Sorenson, P.F., Lyons, N.P.: An empirical approach to the evaluation of icons. SIGCHI Bull. **21**, 87–90 (1989). https://doi.org/10.1145/67880.67894
6. Rogers, Y.: Icons at the interface: their usefulness. Interact. Comput. **1**, 105–117 (1989)
7. Lidwell, W., Holden, K., Butler, J.: Universal Principles of Design. Rockport Publishers, Massachusetts (2003)
8. Purchase, H.: Defining multimedia. IEEE Multimed. **5**, 8–15 (1998). https://doi.org/10.1109/93.664737
9. Gaver, W.W.: Auditory icons: using sound in computer interfaces. Hum.-Comput. Interact. **2**, 167–177 (1986). https://doi.org/10.1207/s15327051hci0202_3
10. Blattner, M.M., Sumikawa, D.A., Greenberg, R.M.: Earcons and icons: their structure and common design principles (abstract only). SIGCHI Bull. **21**, 123–124 (1989). https://doi.org/10.1145/67880.1046599
11. Wang, H.-F., Hung, S.-H., Liao, C.-C.: A survey of icon taxonomy used in the interface design. In: Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore!, pp. 203–206. ACM Press, London (2007)
12. Mollerup, P.: Marks of Excellence. Phaidon Press Limited, London (1997)
13. Korpilahti, T.: A practical approach to icon taxonomy. In: Stephanidis, C. (ed.) HCI 2016. CCIS, vol. 617, pp. 259–263. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40548-3_43

User Experience Evaluation by ERM: Experience Recollection Method

Masaaki Kurosu[1(✉)], Ayako Hashizume[2], and Yuuki Ueno[3]

[1] The Open University of Japan, Chiba, Japan
nigrumamet-s23@mbr.nifty.com
[2] Tokyo Metropolitan University, Tokyo, Japan
[3] Otsuka Business Service, Tokyo, Japan

Abstract. Evaluation of user experience (UX) is the key for improving the quality of artifact that people use in their everyday life and thus improve their quality of life. Two types of evaluation methods can be classified, i.e. the real-time method (e.g. ESM by Larson & Csikszentmihalyi) and memory-based method (e.g. UX curve by Kujala et al.). The ERM proposed here belongs to the latter and asks users to remind their past experience from the time when they had the expectation for targeted artifact and to rate the degree of satisfaction by −10 to +10 rating scale. Two ways of analysis are proposed, one is classifying the result by the aspect of experience and another is classifying the result on the time phase. Web-based application was developed so that the data collection will become efficient.

Keywords: User experience (UX) · Evaluation
Experience recollection method (ERM) · Rating scale · Satisfaction

1 Introduction

Since 1998 when Norman proposed the concept of user experience [20], evaluating the user experience (UX) has become the key activity for improving the artifact, i.e. products and services, that people are using in their everyday life and thus improve the quality of life. In the website of AllAboutUx [1], total of 86 different methods are listed as of June 2017. They include field study methods, laboratory study methods, online study methods and questionnaires to investigate the UX. before usage, as the snapshot, as the episode and as the long-term UX.

This suggests that the UX evaluation methods can be classified into two groups including real-time methods and memory-based methods. The former includes some of field study methods and laboratory study methods in addition to online study methods and questionnaires. And the latter includes the rest of field study methods and laboratory study methods. The real-time method evaluates the experience at the time of evaluation. But by considering the fact that the value of UX may vary depending on the episode that the user may face, it would be better to evaluate the UX, at least, along with a certain period of time or for several times. In this respect, the memory-based method that can evaluate the UX covering a long period of time is

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 138–147, 2018.
https://doi.org/10.1007/978-3-319-91238-7_12

better than the real-time method while there are many reports that the memory sometimes lies and is incomplete since Munsterberg pointed it out in 1908 [19]. Because there are different phases of UX including the expectation before the actual usage, episodes at the time of purchase or short-time after the purchase, episodes that may occur during the long-time usage, and the current impression, the UX evaluation should cover the whole range of experiences for some artifact.

All about UX website [1] includes such developmental phases as the concept, early prototype, functional prototype as well as products on the market, authors think that the real UX can only be evaluated for products on the market and products in use. Foreseeing the UX would be the silver bullet for manufacturers for minimizing the risk of providing unwanted products or services that will give users the poor experience. But what can be called as the experience should be based on the real interaction with the artifact in the past in the real context of use by the real users.

Based on these ideas, the ERM, a new method for evaluating the UX, will be explained in this article.

1.1 Satisfaction as the Measure

In ISO9241-11:1998 [5], the concept of usability has three sub-concepts, i.e. the effectiveness, efficiency and satisfaction. In this 1998 version, it was defined as “freedom from discomfort, and positive attitudes towards the use of the product”. But during the discussion for its revision that was held in the spring of 2017, the trial definition of satisfaction was changed to “person’s perceptions and responses that result from the use of a system, product or service.” Interesting thing is that the definition of UX in the same version was “person’s perceptions and responses that result from the use and/or anticipated use of a system, product or service”. In essence, these definitions are almost the same. Although the trial definition of satisfaction was changed to a different sentence thereafter, this fact shows that the concept of satisfaction and UX are quite closely related with each other even among ISO standard professionals

This is related to the idea that the satisfaction is a super concept of various hedonic aspects including the pleasure, joy, delight, beauty, cuteness, etc. The fact that the satisfaction is a topmost concept among hedonic descriptive word and other quality characteristics was also confirmed by the concept dependency analysis [13].

Based on this idea, the concept structure of quality characteristics as shown in Fig. 1 was proposed [14]. This figure is rather complex, but the basic structure is as follows: on the left, there are “qualities in design” of which ISO/IEC25010:2011 [7] referred to as “product quality”. The reason why it is not called the product quality is that the quality characteristics included in these two boxes are the qualities that should be considered and manipulated during the design activity. And the product quality is just the result of this activity. On the right, there are “qualities in use” that include quality characteristics while the real user is using the artifact in the real context of use. This is the reason why the UX is related to the quality in use.

On the upper side of the figure, there are qualities that are overt and can be measured objectively, thus they are named as objective qualities. On the lower side of the figure there are qualities that are covert because they occur in the human mind, thus are named as subjective qualities. This differentiation is similar to the dichotomy of pragmatic attributes vs. hedonic attributes by Hassenzhal. Based on the combination of these two dimensions, the quality in design vs. quality in use and the objective quality vs. subjective quality, there are four regions of quality characteristics. As can be seen in the lower right area or the subjective quality in use, there is the satisfaction on top of such hedonic characteristics as joyfulness, delight, pleasure, etc.

One more thing that should be noted is that other three areas are linked to the subjective quality in use based on the direct influence or the perception. In other words, the satisfaction is the final and topmost quality characteristics of all. This is the reason why we decided to use the satisfaction as the measure of UX in the ERM.

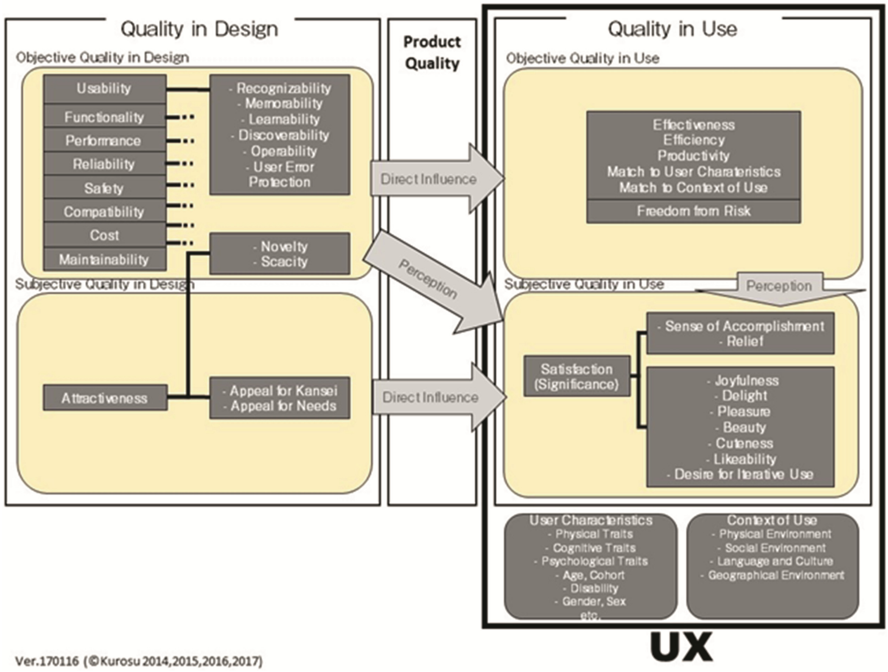


Fig. 1. Four categories of quality characteristics that shows the relationship between the UX and satisfaction. Cited from [14].

1.2 Real-Time Methods and Memory-Based Methods

As was suggested before, we classified UX evaluation methods into two categories: real-time methods and memory-based methods.

The real-time methods include such methods as ESM, DRM [9] and TFD [15]. These methods have the merit of being able to measure the non-memory-biased and non-distorted impression regarding the UX. But, they have the demerit of invasiveness to the life of informant. To take an example of ESM, informants will have to stop their everyday behavior while answering to the phone call. Because of this invasiveness, it is usually said that the adequate maximum span for ESM is 2–3 weeks in the field of clinical psychology.

On the other hand, the memory-based methods including CORPUS, Co-drawing of Chronological Table of Usage [2], iScale [10] and UX curve [12] can evaluate the UX for a long period of time. What is common to these methods is that it is a graph-based method where the abscissa is time and the ordinate is the scale value for evaluation such attributes as the attractiveness, ease of use, etc.

Because it is based on memory, there might be the influence of memory decay and distortion of memory. These deficiencies, however, was typically found in the field of forensic psychology in, for example, the work of Loftus [18]. The case of Prof. von Liszt cited in Munsterberg [19] is the same, though it was in an experimental situation.

But we will have to notice that there is a big difference in the forensic psychological situation and in the UX evaluation situation. In the former situation, what witnesses experience and remember is the external event that, at least initially, nothing to do with them. But in the latter UX situation what users experience and remember is the internal events that were initiated, in most cases, by their own will. In other words, in the UX evaluation, the degree of self-involvement is far stronger than in the case of the witness. Of course, there might be a loss of memory and some distortions. But the degree of this influence is expected to be smaller compared to the case of witness. Furthermore, from some viewpoint, what users report may not always be true. It is paradoxical, but when we obtain the recollection of past experiences of users, we should interpret it as the evidence of current reflection of users' integration of past events and it is the evaluation at the time of survey. In other words, it is the overview of the past events regarding experiences of users from the current standpoint.

With this in mind, we started to improve the memory-based method for evaluating the UX.

1.3 Temporal Model of UX

It can be observed every day that the evaluation of our experience on some artifact (product/service) is changing depending on what happens in the use of that artifact and how it happened. In other words, UX cannot be measured as a single value but should be measured as a series of dynamic fluctuation.

In this sense, the temporal model of UX is necessary for the purpose of evaluating the UX as a fundamental conceptual framework. Although UX White Paper [21] proposed the model of UX over time, the model describing the time spans of UX (Fig. 2 in [21]) looks inadequate because the combination of momentary UX during usage and the episodic UX after usage is not shown as a repetition, and because the concept of cumulative UX over time is not adequate considering the peak end rule proposed by Kahneman et al. [8]. Peak end rule tells that people judge the experience based on the most intense point (peak) and at the end of experience (end). That is, the UX is not just the result of simple accumulation of the total experience, hence the concept of cumulative UX is against this conceptual model. We then thought to evaluate the UX during the usage of artifact but not to summate it to get such a value as the cumulative UX.

Regarding the temporal sequence, another figure in the UX White Paper is suggestive. Because of inclusion of such inadequate keywords as cumulative UX and other UXs, we revised the figure to create a new one. Expectation is included in the total sequence of UX because it is the result of mental activity based on the external information such as the catalogue, information on the web, TV commercial, magazine article, information from the friend, etc. in addition to the internal information stored in the memory of user regarding the experience with the similar or previous version of the artifact. Then comes the phase of purchase. In some cases, people get the artifact by not purchasing at the retail store but by getting it from the friend or family member. Anyways, this is the change point when the consumer becomes the user. Consumer behavior theory such as the EBM model [3] usually deals with the phase up to this point. Generally, the experience value at the purchase is positive because of fulfillment of the purchase motivation. But in some case, for example, when the student entered the university that he rated as a second or third level, the satisfaction level may be a bit negative even at the beginning based on the law of level of aspiration [17].

After the purchase or obtaining the artifact, the phase changes to the initial use. In the case of person-to-person service, its lifecycle usually terminates here. During the initial use, some trouble may occur such as to experience the difficulty while installing the product in user's environment. After the initial use, the long-term use follows. During this phase, many types of episodes may occur including positive ones and negative ones. The duration of this phase will vary based on the type of products. Some products will only be used for months (e.g. battery), and some others will be used for a year or more than that (e.g. laptop computer, telephone, washing machine, university education, being hospitalized etc.).

When the performance of the product degrades, the user becomes bored of using it, or the new product catches the eye of the user, and the product will be wasted. This is the end of the product lifecycle.

The UX evaluation should usually be conducted 6 to 12 months after the installment as is described in ISO9241-210:2010 [6]. It is the time when most users get accustomed to the product and the location of that product may be fixed in the life of the user.

We thought that the memory-based UX evaluation should be conducted at 6 to 12 months after the user started to use the product. But sometimes, the length of time may be much shorter or longer after the start in some special cases.

2 Previous Trials Based on UX Curve

As the best convenient memory-based UX evaluation method that have been proposed, we focused on the UX curve [12] because of its simplicity and visual impact.

After using this method for more than a year, we made a revision of this method and gave it another name, UX graph [14]. The UX graph was proposed because there were many aspects in the UX curve that need to be improved as follows:

1. Abscissa representing the time is not uniform.

Although abscissa is assigned the time, the original UX curve did not show the unit of time scale and the uniformity of time is not guaranteed. In the UX graph, time scale by the year is shown so that the arbitrary change of angle of inclination may not influence the wrong impression.

2. Curve was arbitrarily drawn.

Users could draw the curve in any way they liked in the UX curve. But to give more exactness to the graph, we asked users to draw the curve as exact as possible in the UX graph.

3. Coordinate of each episode is not exactly determined.

Because the curve was arbitrarily drawn, the coordinate of each episode point is not exact in the UX curve. For this reason, we asked users to give the value from +10 to −10 so that the coordinate may be more exact in the UX graph.

4. Drawing similar curves for 3 times is tiresome.

In the UX curve, users were asked to draw three similar curves in terms of attractiveness, ease of use, and utility. Another graph is about degree of usage but its structure is different from three graphs described above. Anyways, we found that three curves looked similar in many cases.

Considering the burden of drawing similar curves for three times, users were asked only to draw the graph regarding the level of satisfaction in the UX graph. The reason for this was written in the Sect. 1.1.

5. Expectation phase is not included.

Although the phase of expectation is included in the UX White Paper as one of the UX phases, the UX curve did not ask users to draw the curve from the expectation phase. In the UX graph, they were asked to draw the graph from the expectation phase.

Based on these considerations we proposed the UX graph and also provided the online version [22] that the users can enter the data from their laptops or smartphones.

In the UX graph method, users are asked to describe episodes they encountered in terms of the artifact they are using. Episodes are written in text and the rating scale value from +10 to −10 is also given. Then they are asked to plot the coordinate of episode on the graph sheet by the solid line using the time and scale value as the coordinate. Finally, they are optionally asked to draw the curve of frequency of use by the dashed line.

An example of UX graph on the use of smartphone by a male user aged 43, the whole graph can roughly be split into two parts. In the first part, the user purchased the smartphone for the first time with a certain degree of expectation and he became satisfied as he got accustomed to its use. But as he continued to use, he encountered some software/hardware troubles. As a result, the frequency of use lowered down a bit. And finally, he now thinks to buy another one.

We used the UX graph for about two years to get more than 200 samples and noticed following point: it might be less meaningful of asking users to specify the time when the episode occurred.

Because of the ambiguous nature of memory, it would be less meaningful to ask users to remind of the time when the episode regarding the artifact occurred. Even though the graph is seemingly exact, it should be called as the quasi-exactness for that reason.

And we decided to discard the idea of asking the exact time to informants. It seemed to be enough to roughly distinguish the phases.

Then we came up to the idea of the experience recollection method or simply the ERM.

3 Experience Recollection Method (ERM)

We have developed the ERM based on the above described idea. In the ERM, the user will be shown only the blank slots on the paper. The time scale is not shown by the year but by the rough phase including the expectation, impression at the time when the user started to use the artifact, experiential episodes along with the long phase of using the artifact, recent episodes and the prospect for the future. Every time when the user fills in the cell with an episode, they are also requested to give the satisfaction rating from +10 to −10.

The format with example is shown in Fig. 2. Because there is no curve nor the graph, it is difficult to grasp the general tendency at a glance. But when we use this data at the interview session afterwards, it will give us many suggestions and further research questions.

Recording Sheet for ERM: Experience Recollection Method Target Item *University* Male/Female Age *18*

1. Write what you experienced at each phase and fill in the evaluation by +10 to -10 rating.

Phase		What you experienced	Evaluation
Expectation before the purchase		*Expecting that I can learn what I could not learn at high school.*	3
Evaluation at the time of start of usage	Year 2016	*May be able to learn various subjects.*	4
Evaluation at early days from the start of usage		*It's located very far from home*	-7
		Interesting to study, but teachers don't let me know how to study.	-3
Evaluation during the use		*Science and technology are interesting*	2
		Could get many good friends.	4
		Very interesting to go to the university.	6
		Too much home tasks and have little time to do them.	-8
Recent evaluation		*There are many interesting lectures.*	7
		There is a limit of the number of lectures that I can take.	-2
Present evaluation	Year 2017	*There are many restrictions.*	2
Estimation in the near future		*Would like to study more seriously.*	4

Fig. 2. An example of ERM regarding the university education. (translated from Japanese)

What we found in this paper version is: in the paper version, the space for writing the episode is limited to write up as many episodes as were reminded.

Because the first version was printed on the paper, there was not the flexibility for giving additional spaces for writing episodes. Paper version has a merit that the group data can be obtained in a short period of time. But we had to think of the improvement regarding this space problem by using the software.

3.1 Development of On-Line Version

Developing the on-line version, as was in the UX graph, was thought to be the solution to the space issue. In the case of UX graph, the space issue was not much critical because the user can add any episodes after the last line regardless of the time sequence. In the on-line version, episodes are written in chronological order. But it is not necessary for the user to write down the episode in the order of occurrence.

In the ERM, the space issue was crucial because the area for episodes are blocked along with the time phase. On-line version will give us the solution to this issue. Based on the adequate programming, a new line will appear on the screen one after another as necessary. We first developed the Japanese version [23] and released it in the spring of 2017.

3.2 Various Application Field

Because the ERM can be applied to any kind of artifact, it can be applied, for example, to the education at university as a service activity. This kind of information will give university staffs the feedback informative for understanding problems that students may have and for improving the educational environment or the curriculum. In one example, the student is highly motivated for the study at the university. But he has complaints about the location of the university that is very far from his home and is suffering from the burden of the home task. Though the second point was rated as negative, it is not a big issue because he has a high level of motivation.

But there may exist some other cases where the students point out fundamental deficiency of the university system that may give university staffs a good opportunity to improve the education if they take the information seriously.

3.3 Feedback to the Development

After the ERM (and interview research) is conducted, the information obtained should be categorized in terms of the type of issue (e.g. battery life, waterproof, etc.) considering the phase of occurrence by using KJ method [11] or the affinity diagram. Summarized information should be fed back to the planning section or the technological section. For example, it may be revealed that prolonging the battery life of the smartphone while maintaining the current thickness and weight is technically very difficult, but its importance is quite high. Hence the development of the long-life battery may lead the company to the top position in the smartphone market.

Figure 3 shows this relationship. The upper flow is the industry development process starting from planning, designing, manufacturing and selling. During the designing stage, there are understanding user, specifying the requirement, designing the solution and evaluating the design included.

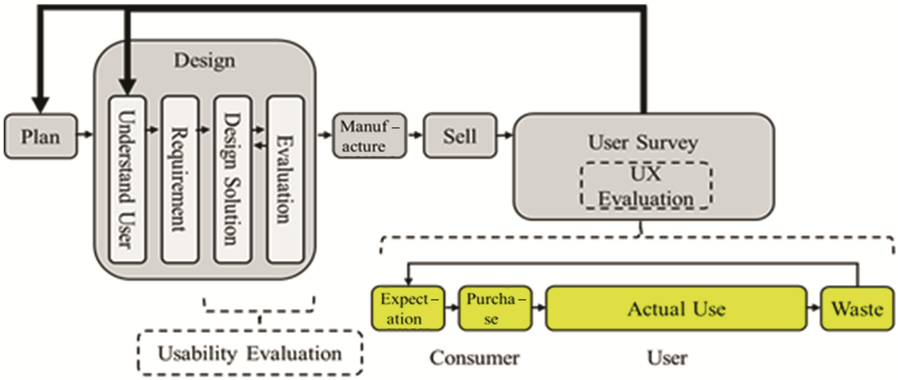


Fig. 3. Relationship of UX evaluation and planning. The upper flow represents industry process and the lower flow represents user experience process.

The user survey on the UX by using, e.g. the ERM, should be conducted after the release of the product/service. The lower flow is the user experience process starting from the expectation, then the purchase and actual use follow, and finally reach the waste. All these user experience process should be evaluated by the UX evaluation method. The information obtained at the UX evaluation should be fed back to the planning section or the design team (heavy line). This feedback is crucially important for reflecting the user experience to the development of next version of the product/service.

4 Conclusion

Based on the consideration on the concept of satisfaction as an integrative measure for the UX, the consideration on the time phase of the UX, we developed the ERM or the Experience Recollection Method. This method is based on our experience with the UX curve and its revised version, UX graph. It was revealed that this method will give us much information on the UX in terms of various artifacts including products (e.g. smartphone) and services (e.g. university education). The result of this method is useful for further interview research and, then, will be effective for giving the adequate feedback to the planning activity of the next version of products/services in their development process.

References

1. All About UX. http://www.allaboutux.org/all-methods
2. Ando, M.: Dynamic Change of Long-term Usability – Change of Context of Use and Its Influence, pp. 31–45, Sokendai Bunka-kagaku Kenkyu (2007). (in Japanese)
3. Blackwell, R.D., Engel, J.F., Miniard, P.W.: Consumer Behaviour. South Western, Div of Thomson Learning International (2005)
4. Hashizume, A., Ueno, Y., Tomida, T., Suzuki, H., Kurosu, M.: Web tool for the UX graph. In: ISASE 2016 Proceedings (2016)

5. ISO9241-11:1998: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) – Part 11: Guidance on Usability
6. ISO 9241-210:2010: Ergonomics of Human-System Interaction - Human-Centred Design for Interactive Systems
7. ISO/IEC25010:2011: Systems and Software Engineering – Systems and Software Quality Requirements and Evaluation (SQuaRE) – System and Software Quality Models
8. Kahneman, D., Fredrickson, B.L., Schreiber, C.A., Redelmeier, D.A.: When more pain is preferred to less: adding a better end. Psychol. Sci. **4**(6), 401–405 (1993)
9. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B.: User experience over time: an initial framework. In: ACM SIGCHI 2009 Proceedings, pp. 729–738 (2009)
10. Karapanos, E., Martens, J.B., Hassenzahl, M.: Reconstructing experiences with iScale. Int. J. Hum. Comput. Stud. **70**, 849–865 (2012)
11. Kawakita, J.: Idea Generation Method - for the Development of Creativity. Chuo-koron sha (1967). (in Japanese)
12. Kujala, S., Roto, V., Vaananen-Vainio-Mattila, K., Karapanos, E., Sinnela, A.: UX curve: a method for evaluating long-term user experience. Interact. Comput. **23**(5), 473–483 (2011)
13. Kurosu, M.: Usability, quality in use and the model of quality characteristics. In: Kurosu, M. (ed.) HCI 2015. LNCS, vol. 9169, pp. 227–237. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20901-2_21
14. Kurosu, M.: Theory of User Engineering. CRC Press, Boca Raton (2016)
15. Kurosu, M., Hashizume, A.: TFD (Time Frame Diary) – a new diary method for obtaining ethnographic information. In: APCHI 2008 Proceedings (2008)
16. Larson, R., Csikszentmihalyi, M.: The experience sampling method, In: Reis, H.T. (ed.) Naturalistic Approaches to Studying Social Interaction, New Directions for Methodology of Social and Behavioral Sciences, pp. 41–56. Jossey-Bass (1983)
17. Lewin, K., Dembo, T., Festinger, L., Sears, P.S.: Level of aspiration. In: Hunt, J.McV. (ed.) Personality and the Behavioral Disorder, vol. 1, pp. 333–378 (1944)
18. Loftus, E.F.: Memory: Surprising New Insights into How We Remember and Why We Forget. Addison-Wesley, Reading (1980)
19. Munsterberg, H.: On the Witness Stand: Essays on Psychology and Crime, Doubleday, Page & Co., Garden City (1908)
20. Norman, D.A.: The Invisible Computer – Why Good Products Can Fail, the Personal Computer is So Complex, and Information Appliances are the Solution, The MIT Press, Cambridge (1998)
21. Roto, V., Law, E.L.-C., Vermeeren, A., Hoonhout, J. (eds.): User Experience White Paper – Bringing Clarity to the Concept of User Experience (2011). http://www.allaboutux.org/uxwhitepaper/
22. Ueno, Y., Kurosu, M., Hashizume, A.: On-line version of UX graph (2016). (in English). https://ux-graph.com/ver4/index.php?lang=en
23. Ueno, Y., Kurosu, M., Hashizume, A.: On-line version of ERM (2017). (in Japanese). https://ux-graph.com/erm/

Don't Leave Me Alone: Retrospective Think Aloud Supported by Real-Time Monitoring of Participant's Physiology

Alexandros Liapis[1(✉)], Christos Katsanos[2,3], and Michalis Xenos[4]

[1] School of Science and Technology, Hellenic Open University, Patras, Greece
aliapis@eap.gr
[2] Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[3] HCI Group, Electrical and Computer Engineering Department, University of Patras, Patras, Greece
ckatsanos@ece.upatras.gr
[4] Computer Engineering and Informatics Department, University of Patras, Patras, Greece
xenos@upatras.gr

Abstract. Think aloud protocols are widely applied in user experience studies. In this paper, the effect of two different applications of the Retrospective Think Aloud (RTA) protocol on the number of user-reported usability issues is examined. To this end, 30 users were asked to use the National Cadastre and Mapping Agency web application and complete a set of tasks, such as measuring the land area of a square in their hometown. The order of tasks was randomized per participant. Next, participants were involved in RTA sessions. Each participant was involved in two different RTA modes: (a) the strict guidance, in which the facilitator stayed in the background and prompted participants to keep thinking aloud based on his judgement and experience, and (b) the physiology-supported interventions, in which the facilitator intervened based on real-time monitoring of user's physiological signals. During each session, three participant's physiological signals were recorded: skin conductance, skin temperature and blood volume pulse. Participants were also asked to provide valence-arousal ratings for each self-reported usability issue. Analysis of the collected data showed that participants in the physiology-supported RTA mode reported significantly more usability issues. No significant effect of the RTA mode was found on the valence-arousal ratings for the reported usability issues. Participants' physiological signals during the RTA sessions did not also differ significantly between the two modes.

Keywords: Human-Computer Interaction · Physiological signals
Usability evaluation · Retrospective Think Aloud

1 Introduction

Software development industry has been increasingly focusing on usability as one of the most critical quality characteristics of an interactive system. Usability evaluation

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 148–158, 2018.
https://doi.org/10.1007/978-3-319-91238-7_13

constitutes the key process to improve usability [1, 2]. Quantitative usability metrics, such as 'time on task' and 'task completion rate' provide a way to objectively evaluate the usability of an evaluated system [3], but fail to offer qualitative insight about the root of potential issues in the user experience [4]. On the other hand, qualitative approaches, such as questionnaires, interviews and video analysis, can provide such qualitative data, but these methods are prone to subjectivity and can be time consuming. More recently, researchers and practitioners have introduced new user experience evaluation approaches using facial expression, speech tone and keystroke analysis [5]. Collecting and analyzing data from users' physiology (e.g., heart rate, respiration, skin conductance) is also a powerful recent usability evaluation method [6–9].

Think-aloud (TA) protocol is a qualitative tool that is used to understand users' behavior while interacting with a system in the context of a usability evaluation study. TA protocol was originally developed to support researchers and practitioners in the domain of cognitive psychology for gaining insight into people's mental processes. Later, it was used to study users' performance in activities such as reading, writing and decision-making in various domains. The HCI field has also adopted the TA protocol, which is on the top of the usability evaluation list for many practitioners [10].

During a TA session, participants are required to verbalize their thoughts about their interaction experience, while they perform tasks on the evaluated system. This method enables evaluators to identify usability issues that need to be resolved in the next system version. Such usability issues may cause activation of users Autonomic Nervous System (ANS), which is known as the "fight or flight" response or stress [10, 11]. Computer users with frequently or daily exposure to stressors are in high risk to confront chronic stress, which may badly affect their health [12]. Apart from health issues, stress may also affect users' performance [13], and its presence in interactive computer environments is typically interpreted as a user experience issue.

According to Nielsen [1], TA is the most valuable single usability engineering method. It is a simple and useful technique for data collection, but it has been criticized [14] for noisy or inaccurate data, due to extra cognitive effort imposed on participants. In [15] two modes of TA application are proposed: "concurrent" and "retrospective". Both protocol modes are widely used by HCI researchers and practitioners. In the concurrent mode, participants are asked to verbalize their interaction experience, while working on the task. One main drawback of this mode is that it may affect the way that participants interact within the task, the time they need to complete the task, and their success in task completion [16]. The specific time cost is referred as reactivity effect.

In the retrospective mode, known as Retrospective Think Aloud (RTA), participants verbalize their interaction experience at the end of a task or a set of tasks. This is often done while viewing a recording of their interaction session. RTA, appears to yield more complex and explanatory data, as the test users who participate in the specific session are not under pressure; instead they are free to think aloud in a natural way [17]. Moreover, since the participants are free to perform the tasks without the need to think aloud, the risk of reactivity is eliminated. However, one of the most important drawbacks of the RTA method is that valuable segments of information may be lost due to participants' memory recall problem, as it has been confirmed by [18, 19]. Furthermore, RTA requires additional time, on top of the user testing session for both the participant and the facilitator.

The effectiveness of these two TA protocol modes in terms of usability issues detected has been examined [20, 21]. However, the effect of TA procedural aspects (i.e., when and how exactly a facilitator intervenes) on the effectiveness of the method remains rather unexplored. Ericsson and Simon [15] showed that application of TA strict guidance (i.e., a facilitator stays in the background just to prompt participants to keep thinking aloud) is very difficult to be applied. Therefore, they propose a free approach with more participant-facilitator interaction than the strict way.

The present study examines how two different treatments of the RTA protocol (a) strict guidance and (b) physiology-supported interventions, affect the number of the user-reported usability issues and users self-reported emotional ratings while experiencing these usability issues. In the strict guidance, condition the facilitator prompted participants to think aloud based on his judgement and experience. In the physiology-supported interventions condition, the same facilitator intervened based on real-time monitoring of user's physiological signals, such as skin conductance. In specific, the research questions investigated by this study are the following:

- RQ1: Is there any effect of the RTA mode on total number of usability issues reported by users in RTA sessions?
- RQ2: Is there any effect of the RTA mode on participants' self-reported ratings for their emotional state during a reported usability issue?
- RQ3: Is there any effect of RTA mode on participants' emotional state during the RTA session, as it is indicated by their physiological signals?

The rest of the paper is structured as follows. Section 2 presents the interaction tasks and the experimental general set-up and protocol, while in Sect. 3, the results from the experiment are presented. Finally, in Sect. 4, conclusions, limitations of the presented work and directions for future research are elaborated.

2 Interaction Scenarios and Experimental Setup

2.1 Scenarios

In this study, participants were asked to perform tasks using the free web-based Orthophotos viewing service[1] offered by the Greek National Cadastre and Mapping Agency (NCMA). In this web application, users can navigate the map of the whole country and perform tasks such as finding a specific place for a set of geographical coordinates, measuring distances on the map, measuring the area of a building etc. This web application was selected because a previous heuristic evaluation study, conducted by three experienced evaluators, showed that it has usability issues.

Participants were asked to use the service in order to perform two tasks. In the first task (see Fig. 1 left), which included two sub-tasks, participants were asked to a) locate a well-known bridge in Patras (i.e., the bridge connecting Rio with Antirrio, known as

[1] http://gis.ktimanet.gr/wms/ktbasemap/default.aspx.

'Charilaos Trikoupis' bridge) and measure the distance between the first and the fourth pillar of this bridge and b) navigate in Patras old harbor and measure the length of the breakwater. In the second task (see Fig. 1 right), which also involved two sub-tasks, participants were asked to (a) locate a popular square (i.e., Georgiou Square) in the Patras city center and measure its inner area as defined by the dotted rectangle shown in the right part of Fig. 1 and (b) to modify the measured area to include all parts of the square as defined by the yellow polygon shown in the right part of Fig. 1.

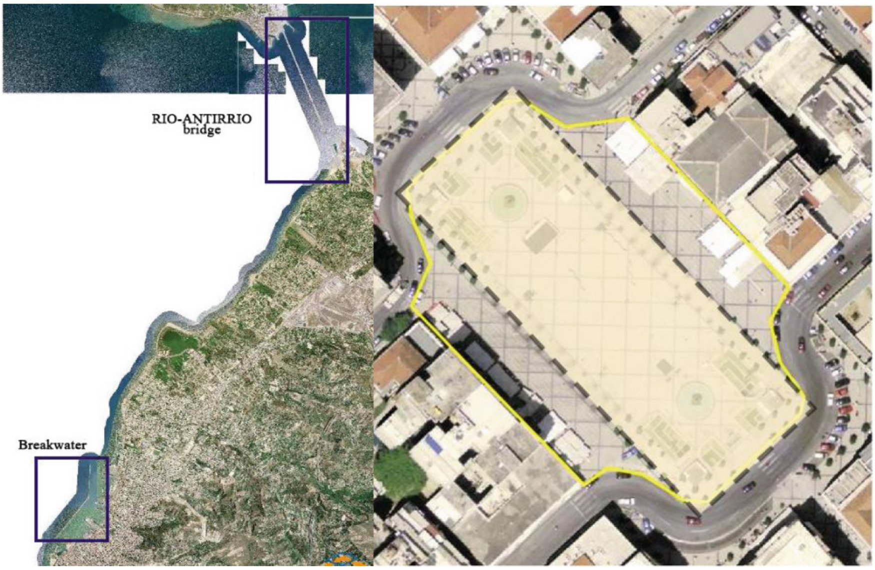


Fig. 1. The Orthophotos viewing service and specific spots used as part of the study tasks given to participants. Left: Participants were asked to find and measure distances (task 1). Right: Participants were asked to find and measure areas (task 2).

None of the study participants had previous experience with the evaluated web application. However, most study participants (24/30) reported that they were rather experienced in map usage and navigation with such applications (e.g., Google Maps). Furthermore, the navigation places were carefully selected to be well-known to participants in an attempt to minimize the effect of spatial knowledge of the area on the interaction experience.

2.2 Experimental Setup

The experiment took place in the facilities of our fully-equipped usability lab. The wireless NeXus-10 physiological platform, along with BioTrace+ interface were used to manage physiological signals recording and real-time monitoring. Three physiological signals were recorded (skin conductance, skin temperature and blood volume pulse) with a sampling rate of 32 Hz. All scenarios were designed to require minimum typing effort in order to minimize participants' hand movements that may affect physiological measurements.

During each experimental session, participants and facilitators were able to communicate through an intercom system. The desktop Tobii-studio recording environment was used to present the interaction scenarios to each participant. Level of room temperature and humidity were continuously monitored to minimize their effect on the collected physiological signals.

A sample of 30 healthy participants (17 males), aged between 18 and 45 (Mean = 32.1, SD = 7.1) was recruited. They were approached from university campus and the place of residence was the single criterion for their selection. Each experimental session lasted approximately 60 min, including short breaks between scenarios. At the end of the experiment, each participant was debriefed about study's purpose and access to their data sources (e.g., eye-activity and physiological signals) was offered as an option to them.

At the beginning of each experimental session, participants were informed that they will be asked to interact with an online map-based service in order to perform two tasks. Next, they completed an appropriate consent form, along with some demographic information. Afterwards, the physiological sensors were placed on participants' non-dominant hand. A short time of approximately five minutes was given to them in order to get used to the sensors' presence, while sensors' transmission quality and participant's body posture in front of the eye-tracker were checked by the experiment facilitators.

Before each task a two minutes relaxing video was presented to participants while their baseline of their physiological signals was recorded. Subsequently, scenarios were presented to participants in a counterbalance mode, in order to remove potential confounds during data analysis phase. At the end of the user testing session, participants answered the Greek version of the standardized 50-item Big Five Trait Test questionnaire[2]. The Google Forms service was used to implement the questionnaire and to collect participants' responses. However, the analysis of both the eye-tracking data and the Big Five ratings are beyond the scope of this paper.

After the user testing sessions, participants were engaged in a RTA session. RTA was applied in two different modes and it was supported by the PhysiOBS tool. PhysiOBS (see Fig. 2) is an innovative tool that effectively combines observation data and self-reported data for continuous emotional states analysis and is delineated in [22]. In this study, it was used to present the video of the user testing session to participants, create Areas of Interest (AOIs) indicating usability issues based on participant's thinking aloud, and assign participant's self-reported ratings to these AOIs.

During each RTA session, participants watched their corresponding interaction session (screen recording) through PhysiOBS. In the strict guidance RTA mode, the facilitator asked from participants to think aloud about their interaction experience and had no further involvement in the process, except reminding them to think aloud in cases of long pauses. In the physiology-supported interventions RTA mode, the facilitator was more engaged in the process. In specific, real-time monitoring of user's physiology (e.g., rising of skin conductance) served as an intervention mechanism for facilitator's actions,

2 http://ipip.ori.org/Greek50-itemBigFiveFactorMarkers.htm.

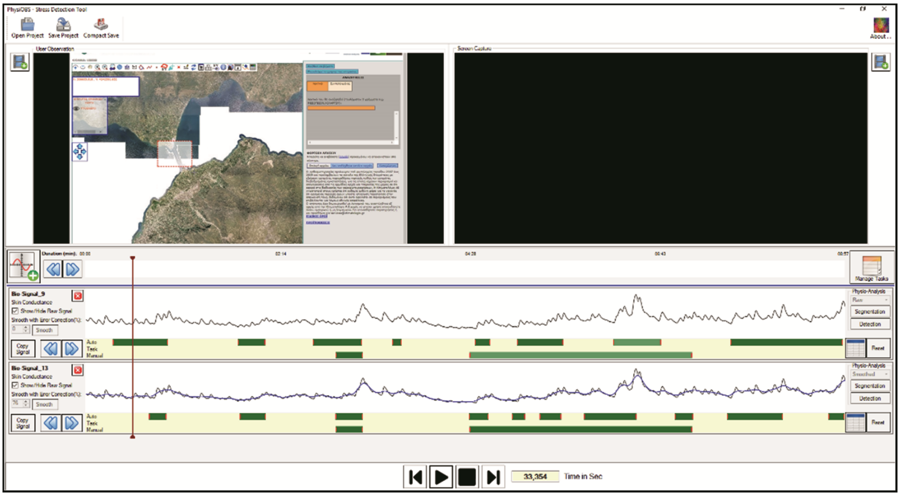


Fig. 2. PhysiOBS: A tool that supports synchronous viewing of multiple user experience data and users' emotional experience evaluation.

such as encouraging participants to think aloud or engaging in brief discussions related to the user-reported issue. The type of RTA session (i.e., strict guidance vs. physiology-supported interventions) was randomly assigned for each participant-task combination.

In addition, the Affect Grid [23] tool was used by participants in order to rate their emotional state in every usability issue they reported. The Affect Grid requires participants to select a point on a 9 × 9 grid that best indicates their emotional state associated with a stimulus, such as a usability issue. Grid's horizontal axis represents the valence (displeasure-pleasure) and the vertical axis the arousal (sleepiness-arousal). For example, if someone feels neutral, then the middle square of the grid (coordinates = 5, 5) is expected to be selected.

3 Analysis and Results

Thirty users participated in a study investigating the effect of two different applications of the RTA protocol on the number of reported usability issues and users self-reported emotional ratings while experiencing these usability issues. All in all, data from twenty-four participants (14 males), aged between 18 and 45 (Mean = 32.3, SD = 7.5) were analyzed. Six cases were excluded from analysis due to missing data (e.g., physiological data recording error). This was a within-subjects study and thus the data analysis was performed on 48 interaction sessions (24 participants × 2 RTA modes). In all subsequent statistical analyses, the effect size r was calculated according to the formulas reported in [24].

3.1 RQ1: RTA Mode and Total Number of User-Reported Usability Issues

Participants reported a total of 115 usability issues: 51 in strict guidance RTA mode and 64 in physiology-supported interventions RTA mode. No grouping was applied to produce a unique list of usability issues.

A two-tailed dependent samples t-test showed a significant difference in the number of usability issues that had been reported between strict guidance (M = 2.08, SD = 1.10) and physiology-supported interventions (M = 2.96, SD = 1.49) mode; t(23) = 2.26, p = 0.033, r = 0.43. This medium-to-large observed effect size [25] demonstrates the importance of the RTA application mode on the effectiveness of the method in identifying usability issues. In specific, participants in the physiology-supported interventions RTA mode reported significantly more usability issues. A parametric test was used, because a Shapiro-Wilk test revealed that the distribution of the differences in the number of usability issues found by the two RTA modes did not deviate significantly from a normal distribution; W(24) = 0.95, p = 0.23.

3.2 RQ2: RTA Mode and VA Ratings for User-Reported Usability Issues

During usability issues reporting, participants were also asked to provide a rating of their emotional state using the valence-arousal space for each usability issue that they reported.

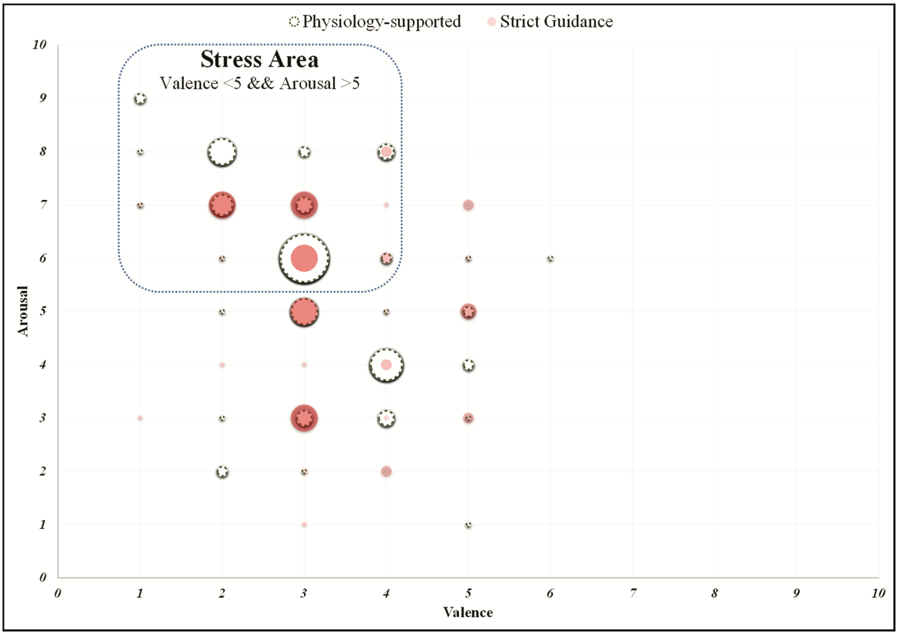


Fig. 3. Valence-Arousal ratings per usability issue for each RTA mode. Bubble size represents the number of ratings for each valence-arousal pair.

Figure 3 illustrates the valence and arousal ratings (N = 115) for each usability issue in each RTA mode. The size of the bubble represents the number of ratings for each valence-arousal pair. Shapiro-Wilk test revealed that emotional ratings were not normally distributed ($p < 0.05$) for both levels of the valence and arousal dependents. A non-parametric test (Mann-Whitney) showed that valence and arousal ratings were not significantly different between RTA modes; valence: $Z = 0.81$, $p = 0.420$, and arousal: $Z = 0.98$, $p = 0.325$.

The dotted frame in Fig. 3 represents usability issues that caused intense emotions (Valence < 5 and Arousal > 5), such as stress [26]. Participants assigned more usability issues (N = 33) in this area during the physiology-supported interventions than during the strict guidance (N = 22) RTA mode. Shapiro-Wilk test revealed that emotional ratings within the stress area were not normally distributed ($p < 0.01$) for both levels of the valence and arousal dependents. A Mann-Whitney test found no effect of RTA mode on valence and arousal ratings in the stress area; valence: $Z = 1.01$, $p = 0.311$, and arousal: $Z = 1.31$, $p = 0.190$.

3.3 RQ3: RTA Mode and Participants' Emotional State During the Thinking Aloud Session

Mean values of participants' physiological signals were recorded during each RTA session and were used as indicators of their emotional state. Table 1 presents descriptive statistics of participants' physiological signals during the two RTA modes.

Table 1. Descriptive statistics of the physiological signals during the two RTA modes. GSR: Galvanic Skin Response, TEMP: Skin Temperature, BVP: Blood Volume Pulse.

Signal	RTA mode	Mean	Median	SD	95% CI
GSR	Strict	4.05	2.35	4.31	[2.23, 5.87]
GSR	Physiology-supported	3.75	2.53	3.16	[2.42, 5.09]
TEMP	Strict	28.43	29.58	5.68	[27.28, 30.83]
TEMP	Physiology-supported	27.96	28.53	5.68	[26.8, 30.36]
BVP	Strict	−22.91	−20.06	12.34	[− 28.12, −17.70]
BVP	Physiology-supported	−23.18	−20.91	12.53	[− 28.5, −17.9]

Two-tailed Wilcoxon signed ranks tests found no significant difference ($p > 0.05$) between the two RTA modes for the recorded signals; skin conductance: $Z = 0.64$, $p = 0.523$, skin temperature: $Z = 0.63$, $p = 0.530$, and blood volume pulse: $Z = 0.16$, $p = 0.875$. Non-parametric tests were used because the assumption of normality was violated for all three recorded signals; skin conductance: $W(24) = 0.35$, $p < 0.001$, skin temperature: $W(24) = 0.78$, $p < 0.001$, and blood volume pulse: $W(24) = 0.88$, $p = 0.010$.

4 Conclusions, Limitations and Future Goals

The aim of this study is to provide usability researchers and practitioners with a better understanding of users' treatment during a RTA session. Think aloud is a popular testing method in collecting usability data. Studies like this one can help evaluators to make more informed decisions about RTA protocol application.

The results of this study demonstrate that the physiology-supported interventions RTA mode significantly outperformed the strict guidance RTA mode in terms of the number of usability issues reported by users. Participants' valence-arousal ratings for the reported usability issues did not differ significantly between the two RTA modes examined in this study. In addition, there was no effect of RTA mode on participants' physiological signals during the RTA sessions. However, in the physiology-supported interventions RTA mode, participants tended to report more stressful usability issues and to have lower mean values for all recorded physiological signals during the RTA sessions.

In sum, the physiology-supported interventions RTA mode seems to be the more appropriate method for evaluators who are interested in detecting more usability issues, rather than the typical strict guidance RTA mode.

As with any research, this study is not without limitations. First, the present study used a within-group design. Hence, no individual differences, such as personality traits and gender, and their possible effects on think-aloud performance was examined. To this end, we are already planning future similar experiments to extend the data collected in this study. In addition, there was one test moderator, and this person was the same between the two conditions (strict guidance and physiology-supported interventions). Although, this approach was chosen to ensure consistency across all users in each condition, it might have affected the results. Future studies need to engage more moderators and investigate this effect, if any. Furthermore, regarding the physiology-supported interventions condition, the interventions were based on real time visual inspection of the physiological signals. The use of an automatic mechanism (a kind of silent alert available only to facilitator) could probably be a parameter for further investigation in an attempt to make more systematic the triggering events of these interventions. Finally, future work also involves designing PhysiOBS-mediated learning activities for instruction of thinking aloud protocols and physiological monitoring of study participants in the context of our previous work [27–31] in HCI education.

References

1. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco (1993)
2. Rubin, J., Chisnell, D., Spool, J.: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, Indianapolis (2008)
3. Seffah, A., Donyaee, M., Kline, R.B., Padda, H.K.: Usability measurement and metrics: a consolidated model. Softw. Qual. J. **14**, 159–178 (2006)
4. Lazar, D.J., Feng, D.J.H., Hochheiser, D.H.: Research Methods in Human-Computer Interaction. Wiley, London (2010)

5. Tsui, W.-H., Lee, P., Hsiao, T.-C.: The effect of emotion on keystroke: an experimental study using facial feedback hypothesis. In: Conference Proceedings of IEEE Engineering in Medicine and Biology Society (EMBC 2013), pp. 2870–2873 (2013)
6. Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M., Karousos, N.: Recognizing emotions in human computer interaction: studying stress using skin conductance. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (eds.) INTERACT 2015. LNCS, vol. 9296, pp. 255–262. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22701-6_18
7. Mandryk, R.L., Atkins, M.S.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int. J. Hum. Comput. Stud. **65**, 329–347 (2007)
8. Liapis, A., Xenos, M.: The physiological measurements as a critical indicator in users' experience evaluation. In: Proceedings of the 17th Panhellenic Conference on Informatics, pp. 258–263. ACM, New York (2013)
9. Georges, V., Courtemanche, F., Sénécal, S., Léger, P.-M., Nacke, L., Pourchon, R.: The adoption of physiological measures as an evaluation tool in UX. In: Nah, F.F.-H., Tan, C.-H. (eds.) HCIBGO 2017. LNCS, vol. 10293, pp. 90–98. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58481-2_8
10. Hernandez, J., Paredes, P., Roseway, A., Czerwinski, M.: Under pressure: sensing stress of computer users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 51–60. ACM, New York (2014)
11. Wilson, G.M., 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 (2000)
12. Anderson, N.B.: Levels of analysis in health science. A framework for integrating sociobehavioral and biomedical research. Ann. N. Y. Acad. Sci. **840**, 563–576 (1998)
13. Diamond, D.M., Campbell, A.M., Park, C.R., Halonen, J., Zoladz, P.R.: The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. Neural Plast. **2007**, 60803 (2007)
14. Nisbett, R., Wilson, T.: Telling more than we can know: verbal reports on mental processes. Psychol. Rev. **84**, 231–259 (1977)
15. Ericsson, K.A.: Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge (1993)
16. Griffiths, M.D.: The role of cognitive bias and skill in fruit machine gambling. Br. J. Psychol. **85**, 351–369 (1994)
17. Eger, N., Ball, L.J., Stevens, R., Dodd, J.: Cueing retrospective verbal reports in usability testing through eye-movement replay. In: Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI…But Not As We Know It, vol. 1, pp. 129–137. British Computer Society, Swinton, UK (2007)
18. Teague, R., De Jesus, K., Ueno, M.N.: Concurrent vs. post-task usability test ratings. In: Extended Abstracts on Human Factors in Computing Systems, CHI 2001, pp. 289–290. ACM, New York (2001)
19. Russo, J.E., Johnson, E.J., Stephens, D.L.: The validity of verbal protocols. Mem. Cogn. **17**, 759–769 (1989)
20. Peute, L.W.P., de Keizer, N.F., Jaspers, M.W.M.: The value of Retrospective and Concurrent Think Aloud in formative usability testing of a physician data query tool. J. Biomed. Inf. **55**, 1–10 (2015)

21. van den Haak, M., De Jong, M., Jan Schellens, P.: Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav. Inf. Technol. **22**, 339–351 (2003)
22. Liapis, A., Karousos, N., Katsanos, C., Xenos, M.: Evaluating user's emotional experience in HCI: the PhysiOBS approach. In: Kurosu, M. (ed.) HCI 2014. LNCS, vol. 8511, pp. 758–767. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07230-2_72
23. Russell, J.A., Weiss, A., Mendelsohn, G.A.: Affect Grid: A single-item scale of pleasure and arousal. J. Pers. Soc. Psychol. **57**, 493–502 (1989)
24. Field, A.: Discovering Statistics Using IBM SPSS Statistics. Sage, Thousand Oaks (2013)
25. Cohen, J.: A power primer. Psychol. Bull. **112**, 155–159 (1992)
26. Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M., Karousos, N.: Subjective assessment of stress in HCI: a study of the valence-arousal scale using skin conductance. In: Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter, pp. 174–177. ACM, New York (2015)
27. Katsanos, C., Tselios, N., Karousos, N., Xenos, M.: Learning web form design by using the KLM form analyzer: a case study. In: Proceedings of the 19th Panhellenic Conference on Informatics, pp. 44–49. ACM, New York (2015)
28. Tselios, N., Altanopoulou, P., Katsanos, C.: Effectiveness of a framed wiki-based learning activity in the context of HCI education. In: 2011 15th Panhellenic Conference on Informatics, pp. 368–372 (2011)
29. Katsanos, C., Tsakoumis, A., Avouris, N.: Web accessibility: design of an educational system to support guidelines learning. In: Proceedings of the 13th Pan-Hellenic Conference on Informatics with International Participation, pp. 155–164, Corfu (2009)
30. Katsanos, C., Tselios, N., Xenos, M.: Tool-mediated HCI modeling instruction in a campus-based software quality course. In: Proceedings of the 20th International Conference on Human-Computer Interaction, Las Vegas (2018, in press)
31. Katsanos, C., Tselios, N., Tsakoumis, A., Avouris, N.: Learning about web accessibility: a project based tool-mediated approach. Educ. Inf. Technol. **17**, 79–94 (2012)

From Design Fiction to Design Fact: Developing Future User Experiences with Proto-Tools

Alessio Malizia[1(✉)], Alan Chamberlain[2], and Ian Willcock[1]

[1] School of Creative Arts, University of Hertfordshire, College ln, Hatfield, UK
{a.malizia,i.willcock}@herts.ac.uk
[2] Mixed Reality Laboratory, School of Computer Science and IT, University of Nottingham, Nottingham, UK
azc@cs.nott.ac.uk

Abstract. Envision a future where the physical and digital become seamlessly intertwined producing unbounded possibilities from interactions with autonomous cars to the advent of smart cities. Tools and methods to explore the design and development of such emergent hybrid spaces will need to be highly inclusive involving engineers, social scientists, up to citizens. Such methods and tools might support the so-called fictitious scenarios in form of digitally augmented physical prototypes set in a near-future reality.

In this paper, we introduce the concept of "proto-tool": multi-purpose digitally augmented artefacts with no constraints and seemingly limitless uses (e.g. virtual and augmented reality devices, digitally augmented surfaces and objects, etc.) that will be modeled on current and fictitious scenarios able to interact both with physical and digital spaces de facto enabling rapid prototyping of Cyber-Physical Interactions.

Keywords: Physical-Cyber systems · Design fictions · Rapid prototyping User experience

1 Introduction

Evolving technologies such as smart cities, humanoid robots, domestic Internet of Things (IoT) and autonomous vehicles are creating a vision of near future environments, i.e. Physical-Cyber Environments [1]. These range from digitally augmented objects (tangible user interfaces) to interactive virtual spaces (MS Hololens) to the Internet of Things vision and Physical-Cyber Environments (P-CE). These emergent technologies together form a paradigm of future interactions where the digital and physical properties of objects are mixed together and merge fictional scenarios with working prototypes.

For instance, conducting user-based research into autonomous vehicles, a product that is yet to properly exist in an accepted day-to-day form, is difficult as it involves researching a 'future' product, which is yet to fully come into being [2]. Controlled experiments and classic research methods in Human-Computer Interaction are considered to have limitations in their application to P-CE suggesting that newer methods for researching user interactions and future technologies are needed [3].

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 159–168, 2018.
https://doi.org/10.1007/978-3-319-91238-7_14

The design of near future products and environments must involve the collaboration and expertise of various disciplines and stakeholders, such as: artists, software engineers, interaction and product designers, business managers and policy makers. Each of them brings their own internalised assumptions and thought processes, making understanding and discussion between the various parties potentially problematic. Tools and methods are needed to aid productive dialogue between those involved in such processes. For instance, Shedroff and Noessel [4] suggest that lessons can be learned from the interfaces written about in science fiction and employed in the development of real world interfaces. They speak of a two-way influence on design, one relating to real world design, influencing hybrid science fiction interfaces and the other based in science fiction influencing real-world interface design by inspiration, expectation, social context and the innovation of new paradigms. Science Fiction Prototyping (SFP) is a method that allows engineers, designers or futurists to think about the technologies they are developing from a human perspective, linking the imaginations of product developers and teams to future users and usage. SFPs are generally short literary works of fictions, which are grounded in scientific facts. The purpose of these stories is to start conversations about the implications, effects or ramifications that technology may have on the future [5].

However, the literature on SFP highlights a number of shortcomings in the use and application of this methodology. In its current format, SFP does not fully support the ideation process with respect to the development of concepts. In part this is due to the individual, exhaustive nature of creating a written SFP. Furthermore, the results are quite limited reducing to a classic wall full of colored sticky notes, few storyboards, a story world and a fancy video at most. Complex environments, such as C-PE need prototypes to be properly explored and render a vision of the future technology that is tangible and accessible to those participating in such design discussions. Low-fidelity prototyping (e.g. paper prototypes, storyboards and narratives) is just a first step towards developing a system and achieving working prototypes can allow users to inspect new forms of interactions; recently, Schmidt [6] noted in a critique to extensive use of low fidelity prototyping to test user experience (sketching, storyboarding, videos, etc.) that «making functional prototypes is a source of inspiration, understanding and reflection».

This work aims to build upon the idea of SFP, to introduce the use of proto-tools – multi-purpose digitally augmented artefacts with no constraints and seemingly limitless uses (e.g. digitally augmented surfaces and objects) – that can aid the creation of future fictional scenarios and corresponding high-fidelity prototypes for exploring the design of interactions in Physical-Cyber Environments.

2 Challenges

A growing challenge to PC-E design practice is the need to bridge the communication gap between various professions, designers, other stakeholders and end user groups involved in the design process. Even within sub departments of organisations people have “unique perspectives” of aims and tasks causing conflict. Another aspect is time to be effective and productive. It can take years to be effectively submerged in a new culture. For example, Myra Strober [7] discovered that at Bio X, an interdisciplinary science centre at Stanford, it took two years of weekly meetings to learn the culture and habits of mind of each other's disciplines. Therefore, time effective solutions need to be explored to begin to address some aspects of developing productive communication between various and diverse collaborators.

For instance, the use of fictional scenario-based design and rapid prototyping methodologies might break the mental image of future scenarios, biased by participants' background. The use of a generative session, or workshop, allows for the understanding of latent and tacit knowledge and while bounded in time by sessions, a fictional design workshop could be complemented by digitally augmented prototypes instead of low-fidelity prototypes (e.g. storyboards). Digitally augmented prototypes can materialise ideas to stimulate participants reflection in action and at the same time create a common language to talk about problems and solutions [8].

The main challenge then shifts to introduce digitally augmented technologies in design fiction workshops that can be used by participants to prototype solutions without requiring any specific technical background, i.e.: designers, stakeholders and end user groups.

3 The Quest for Digitally Augmented Artefacts

Ethnography suggests that collaboration can be enabled by shared representation, but these externalised representations add to cognitive processing [9]. Externalisation of individual's thoughts and ideas via representation in artefacts can aid communication of those thoughts and ideas. The question then becomes: how to inspect others' thoughts or ideas in an effective and productive way to inform the design of future cyber-physical environments.

Our hypothesis is that digitally augmented artefacts spontaneously created by people to support their own vision of such cyber-physical environments can offer a window into their cognitive and creative process. Our hypothesis is consistent with research in distributed cognition and the use of artefacts to externalise cognitive models [10, 11]. Furthermore, the theory of embodied interaction considers the materiality of tools as one of the most critical cognitive resource for human activity [12, 13]. Considering cognitive artefacts as a glimpse on participants' inner design mechanisms has some limitations studied by [14]: especially in terms of artefacts being too bounded to the context or scenario proposed in the design session (context bias). We propose to address such limitation by stimulating rapid prototyping of Cyber-Physical environments through the use of proto-tools allowing the articulation of artefact at "basic level" of generality following Rosch's contribution in prototype theory [15] and Schon's reflection in action [8]. The aim is to use proto-tools to inspect design brainstorming outcomes and reflect on them.

In our vision a proto-tool is a rapid protyping environment used to assign behavior to smart artefacts, i.e. smart objects that offer features like connectivity, sensors, actuators, and embedded software [16]; by exploiting the dual nature of smart objects (physical and digital) and leveraging on human's natural ability of interpreting and manipulating objects in the real world, proto-tools aid participants to render fictional scenarios into functional prototypes to be used as a source of inspiration, understanding and reflection

[6]. The rapid prototyping environments are used to assign specific meaning and behaviour to smart objects without requiring programming expertise but instead providing a technology powerful enough to render a working prototype of the fictional scenarios.

4 Prototyping Cyber-Physical Environments with Proto-Tools

From autonomous vehicles and Internet of Things, to futuristic augmented experiences in museums Physical-Cyber Environments (P-CE) are emerging as a relevant paradigm in future interactions where the digital and physical properties of smart artefacts are mixed together to render fictional scenarios into working prototypes.

For example, in domains like Cultural Heritage (CH), smart artefacts can be installed in museums, archaeological parks and exhibitions to create smart visit experiences, i.e. scenarios where visitors acquire CH content by interacting with the surrounding environment and smart artefacts included in it. Nevertheless, such interactions among people and smart artefacts in an augmented environment are technically challenging for museum curators and other stakeholders involved in design visits. In Desolda et al. [31] researchers introduced a tool called EFESTO-5W for simplifying the creation of Event-Condition-Action (ECA) rules combining smart object events/action (see Fig. 1). Some studies have shown that the EFESTO-5W composition paradigm effectively guides curators and non-technical users in establishing the behavior of multiple smart objects.

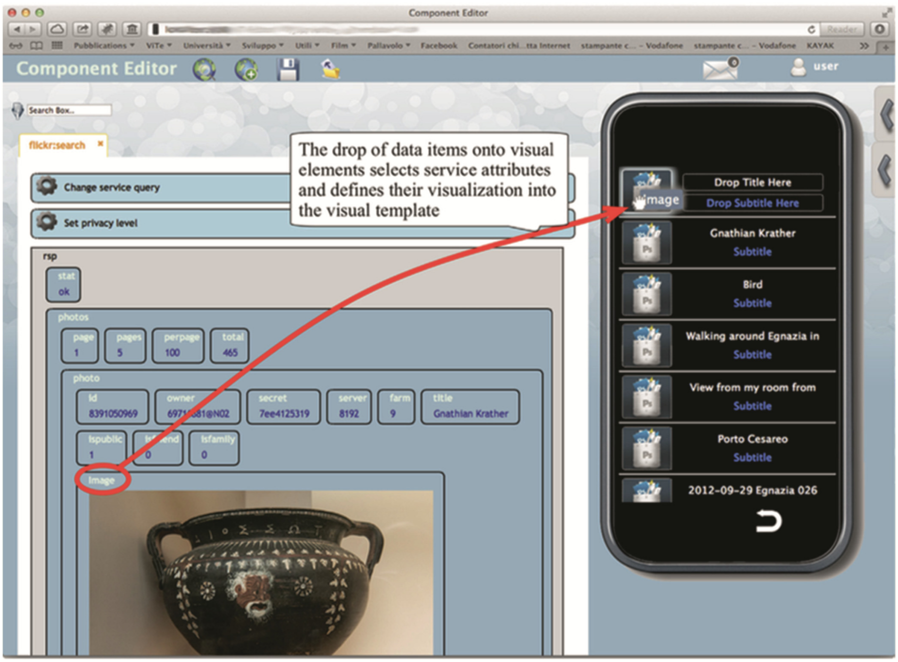


Fig. 1. An example of service attributes assigned to a smartphone for future use by visitors as planned by curators without involving any programming.

In [32] a tangible user interface block-oriented programmable object (TUIBOPO) framework was introduced as an extension of the TUIO protocol providing further interaction capabilities for multi-device environments. The TUIO protocol, as Kaltenbrunner et al. (2015) stated, "is an attempt to provide a general and versatile communication interface between tangible tabletop controller interfaces and underlying application layers. It was designed to meet the needs of tabletop interactive multi-touch surfaces, where the user is able to manipulate a set of objects and draw gestures onto the table surface with the fingertips."

TUIBOPO can be considered a framework for implementing block-oriented programmable objects that simplifies the implementation of such objects with physical and digital properties (Fig. 2).

Fig. 2. Block-oriented programming with tangibles. Users are assigning digital properties and functionalities to a smartphone assembling blocks instead of by programming.

Those are just few examples of the emerging platforms providing capabilities that are compliant with our definition of proto-tools and that can be exploited to design and implement high-fidelity prototypes in fictional design scenarios.

5 From Physical-Cyber Environments Design Fiction to Working Prototypes

The world that we inhabit is full of tangible objects, objects of different weights, shapes, textures, colours, temperatures and even tastes and scents; some are stationary while others are portable. We pick up books, play musical instruments, wear clothes, sit on furniture and engage with all manner of artefacts in our day-to-day existence. With the emergence of smart materials and the IoT [16], we are having to re-think how we engage with a range of smart, interconnected objects, both day-to-day and specialist devices. This reconsideration of purpose is reflected in the emergence of a range of approaches to the design of TUIs.

Developing and optimising tangible interfaces, with their mix of physical active elements combined with variably configured displays, presents significant challenges compared to design processes targeted at traditional, screen-centric keyboard/mouse driven UI systems. Sketching, perhaps the most widely used design tool is obviously used to efficiently communicate concepts to other members of a design team but also, and perhaps more importantly in terms of the current study, is a central methodology in design research. In this role, it can be considered as an early-stage, minimal fidelity, prototyping methodology. "… sketching is thinking. … Its purpose is to test and verify a concept through a communicable representation". [17] This design-research aspect of sketching is problematic in a hybrid digital/physical domain since the elements that make up tangible interface systems tend to require configuration and linking in application-specific ways that mitigate against the flexibility and low investment investigations typical in more conventional, paper-based sketching. Blackwell et al. [18] explicitly follow the sketching paradigm in their 'Tangible-Prototyping Workshop' methodology, but extend the range of sketching resources to include familiar 3D materials such as foam boards, modelling clay, pipe cleaners, string etc. Their aim is to be able to rapidly produce tangible 'sketches' that that provide maximum opportunity for creative exploration and evocative experiences and which support a range of analyses of the resultant 'solid diagram'. Their approach is severely limited in terms of the functionality of the prototypes produced, but a number of other projects such as that of Nam (2005) [19] and 'Sketch-a TUI' [20] have sought to address this problem through approaches based on 'interactive sketching', extensions or augmentations of traditional sketching which seek to retain the spontaneity of drawing while using AR and sensing to implement some of the functionality of reactive objects.

In contrast to the underlying sketch paradigm of these projects, where there are (at least in principle) very few preconceptions about the ontology of the tangible artefact being designed, many projects have attempted to facilitate rapid TUI prototype development by providing collections of selected functional elements which may then be assembled by the design team to construct a prototype. These frameworks or toolkits typically consist of (or enable) a number of physical elements ('iStuff' [21], 'Makey Makey' [22]) or small active networked devices designed to add functionality to physical objects ('Smart-Its' [23], 'CookieFlavour' [24], 'Amerino' [25], 'Kniwwelino [26]) together with a software architecture which handles configuration and intra-object communication in a way that does not require specialist coding skills. An alternative, but related approach seeks to use AR to resolve the difficulties of configuring and connecting the physical elements of a TUI by providing a prototyping framework which is partly [27] or wholly [28] virtual. A third approach, perhaps acknowledging the ludic nature of parts of early-stage design processes as well as the openness and flexibility required of systems designed to stimulate the imagination, uses frameworks such as Lego [29] originally intended for play. While it might be argued that this necessarily reduces the power of the prototyping process, turning it from an exploratory, unbounded process to one where creativity becomes at least to some extent a matter of selection, in practice provided a sufficiently large number of varied elements are included in the toolkit or they are sufficiently generic and open in their potential configurations, results can be useful at least in terms of exploring the materiality of tangible interface elements.

However, such toolkits are at best a collection of casings and interface mechanisms such as buttons, dials, sliders and so on which typically have low fidelity in terms of the physical materiality and haptic identity of tangible interface elements. As the 'TUIkit' project [30] has suggested, evaluating and optimizing material and physical qualities is a hugely important part of the TUI design process, yet, "in the design process of tangible interfaces, little attention has gone to the meaning of material qualities for TUIs, and materials are often chosen more for their availability or technical qualities, rather than for their benefits to the user" and then only comparatively late in the design cycle. So how do we develop design tools to enable us to create high-level cyber-physical systems, that require a tight linkage between the physical and the digitally functional? Developing such tools is a design problem in its own right. In order to think about some challenges that we face, we offer a couple of scenarios, conceptual designs that can aid us in our design-thinking. In our first scenario first we propose a system that is a simple tangible device that we call the Palm Stone (PS). The PS is a small stone like object that sits in the palm of the hand, the user carries around with them and it responds, in different ways to certain activities in the environment. For example our PS gets warm when our son returns home from school, it vibrates gently when our friends are playing an online game that we play together, it vibrates with a certain rhythm for a diary date and it has a range of preferences that can be set to trigger it when certain events, or environmental factors are sensed. Theoretically it appears as though this is a simple case of 'if this then that', but practically making a small prototype of such a system would need us to program the actuators, devise vibration systems, add heating elements, networking protocols would need to be dealt with and the physical form would need to be made. It could be done with a low power Arduino system, soldered together and cased, but imagine if there was a simple system that allowed us to add a series of triggers to a ready made system that would then allow us to quickly prototype and take out the system into 'the wild' [33] gather feedback and refine the system. This would be a much more useful proposition for a range of possible users; from designers to people trying to understand consumer behaviour.

Now imagine if we were able to take such technologies and employ them in domains and places such car interfaces where different people might want different physical interface properties, or want different degrees of haptic feedback, want the interface to sense pollution, or react to voices and so on. Currently having to develop something on this scale would need a whole range of programming, engineering and 3D design skills. Developing proto-tools to help support the development of cyber-physical systems would be a huge leap forwards as a way to quickly create tangible prototypes even about near future technologies.

6 Conclusion

Cyber-Physical environments are becoming a relevant area of research considering the evolution of technology and near-future visions of digitally and physically augmented objects, which can offer unbounded possibilities especially in terms of impact and interaction with the society, In order to design for Cyber-Physical environments and therefore smart objects operating in such scenarios we need to involve a wide range of stakeholders, from engineers and social scientists to citizens and thus we need methods and tools that will allow participation from different group of backgrounds.

Those methods, for instance fictional design scenarios, normally involve low level prototyping in form of storyboards or videos but we think that currently there are more possibility to refine those prototypes and produce higher level artefacts.

We introduce the concept of proto-tools which are an emergent type of environments that can be used to rapid prototype smart objects in Cyber-Physical environments with little or no technical background required. Those proto-tools will provide a way to overcome the limitation of low fidelity prototypes capabilities when rendering fictional scenario environments. For instance, we can imagine a multi-disciplinary group of participants in the automotive sector, like: engineers, policy makers, drivers being able to prototype a dashboard of an autonomous car and render it, instead of with a bunch of sketches and storyboards, with a mixed reality application programmed via a proto-tool where a physical box representing the dashboard frame could be extended with augmented reality projection of controls made interactive by haptic controllers. Such scenario can then be explored thoroughly revealing user experience issues and opportunities at a more complex scale overcoming the limitation of a low fidelity prototype.

References

1. Forshaw, S., Cruickshank, L., Dix, A.: Collaborative communication tools for designing: physical-cyber environments. In: Fourth International Workshop on Physicality (2012)
2. Vavoula, G.N., Sharples, M., Rudman, P.D.: Developing the 'future technology workshop' method. In: Proceedings of the International Workshop on Interaction Design and Children, pp. 28–29, August 2002
3. Bleecker, J.: Design Fiction. A Short Essay on Design, Science, Fact and Fiction. Near Future Laboratory (2009)
4. Shedroff, N., Noessel, C.: Make It So: Interaction Design Lessons from Science Fiction. Rosenfeld Media, Brooklyn (2012)
5. Johnson, B.D.: Science Fiction Prototyping: Designing the Future with Science Fiction, 1st edn. Morgan & Claypool Publishers, San Rafael (2011)
6. Schmidt, A.: Understanding and researching through making: a plea for functional prototypes. Interactions **24**(3), 78–81 (2017)
7. Strober, M.: Interdisciplinary Conversations: Challenging Habits of Thought. Stanford University Press, Stanford (2011)
8. Schon, D.A.: The Reflective Practitioner: How Professionals Think in Action, vol. 5126. Basic books, New York (1984)
9. Lindley, J., Potts, R., Sharma, D.: Operationalizing design fiction with anticipatory ethnography. In: Ethnographic Praxis in Industry Conference (2015)
10. Hutchins, E.: Cognition in the Wild. MIT Press, Cambridge (1995)
11. Norman, D.A.: Cognitive artifacts. In: Carroll, J.M. (ed.) Designing Interaction, pp. 17–38. Cambridge University Press, New York (1991)
12. Dourish, P.: Where the Action Is: The Foundations of Embodied Interaction. MIT Press, Cambridge (2001)
13. McCarthy, J., Wright, P.: Technology as experience. Interactions **11**(5), 42–43 (2004)

14. Collins, A.M., Ferguson, W.: Epistemic forms and epistemic games: structures and strategies to guide inquiry. Educ. Psychol. **28**(1), 25–42 (1993)
15. Rosch, E.: Principles of categorization. In: Rosch, E., Lloyd, B.B. (eds) Cognition and Categorization, pp. 27–48. Lawrence Erlbaum Associates, Publishers, Hillsdale (1978)
16. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)
17. Goldschmidt, G.: The dialectics of sketching. Creat. Res. J. **4**(2), 123–143 (1991)
18. Blackwell, A.F., Edge, D., Dubuc, L., Rode, J.A., Stringer, M., Toye, E.F.: Using solid diagrams for tangible interface prototyping. IEEE Pervasive Comput. **4**(4), 74–77 (2005)
19. Nam, T.-J.: Sketch-based rapid prototyping platform for hardware-software integrated interactive products. In: Extended Abstracts on Human Factors in Computing Systems, CHI 2005, pp. 1689–1692. ACM, New York (2005)
20. Wiethoff, A., Schneider, H., Rohs, M., Butz, A., Greenberg, S.: Sketch-a-TUI: low cost prototyping of tangible interactions using cardboard and conductive ink. In: Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction, pp. 309–312. ACM, New York (2012)
21. Ballagas, R., Ringel, M., Stone, M., Borchers, J.: iStuff: a physical user interface toolkit for ubiquitous computing environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 537–544. ACM, New York (2003)
22. Collective, B.M., Shaw, D.: Makey makey: improvising tangible and nature-based user interfaces. In: Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction, pp. 367–370. ACM, New York (2012)
23. Gellersen, H., Kortuem, G., Schmidt, A., Beigl, M.: Physical prototyping with smart-its. IEEE Pervasive Comput. **3**(3), 74–82 (2004)
24. Kimura, H., Tokunaga, E., Okuda, Y., Nakajima, T.: CookieFlavors: easy building blocks for wireless tangible input. In: Extended Abstracts on Human Factors in Computing Systems, CHI 2006, pp. 965–970 (2006)
25. Kaufmann, B., Buechley, L.: Amarino: a toolkit for the rapid prototyping of mobile ubiquitous computing. In: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 291–298. ACM, New York (2010)
26. Moll, C., Hermen, J., Maquil, V., Rangel, H., Hill, J.-N.: Prototyping connected tangible interactions with Kniwwelino. In: CEUR Workshop Proceedings of the Third European Tangible Interaction Studio (ETIS) (2017)
27. Hornecker, E., Psik, T.: Using ARToolKit markers to build tangible prototypes and simulate other technologies. In: Costabile, M.F., Paternò, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 30–42. Springer, Heidelberg (2005). https://doi.org/10.1007/11555261_6
28. Diewald, S., Roalter, L., Möller, A., Kranz, M.: Simulation of tangible user interfaces with the ROS middleware. In: Adjunct Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction, TEI 2014, vol. 14 (2014)
29. Gohlke, K., Hlatky, M., de Jong, B.: Physical construction toys for rapid sketching of tangible user interfaces. In: Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction, pp. 643–648, New York (2015)
30. Vissers, J., Geerts, D.: TUIkit: evaluating physical and functional experiences of tangible user interface prototypes. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1267–1276. ACM, New York (2015)
31. Desolda, G., Ardito, C., Matera, M.: EFESTO: a platform for the end-user development of interactive workspaces for data exploration. In: Daniel, F., Pautasso, C. (eds.) RMC 2015. CCIS, vol. 591, pp. 63–81. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28727-0_5

32. Malizia, A., Turchi, T., Olsen, K.A.: Block-oriented programming with tangibles: An engaging way to learn computational thinking skills. In: 2017 IEEE Blocks and Beyond Workshop (B&B), Raleigh, NC, pp. 61–64 (2017)
33. Chamberlain, A., Crabtree, A., Rodden, T., Jones, M., Rogers, Y.: Research in the wild: understanding 'in the wild' approaches to design and development. In: Proceedings of the Designing Interactive Systems Conference, DIS 2012, pp. 795–796. ACM, New York (2012)

What's the Impact of Local Cultures on the User Experience of Software Solutions?

Denise Pilar[1](✉) and Cecilia Henriques[2]

[1] SAP Labs Latin America, São Leopoldo, RS, Brazil
denise.pilar@live.com

[2] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, UFSC, Florianópolis, SC, Brazil
ceciliamhenriques@yahoo.com.br

Abstract. Culture has a direct influence on the user experience. Good experiences should consider the context in which users belong and live, whether social, or organizational. For this, the qualitative research plays an essential role, since it is the one that will give the designer the possibility to know the users of the solutions he/she designs, understanding their needs and problems, process and tasks that he/she performs, as well as the famous *'gambiarra'*, a quick fix so frequently seen in computer systems. This article aims at exposing the importance of qualitative research in order to understand the cultural context of users. It is a theoretical article in which the authors use examples to demonstrate the importance of the research, especially the research that is carried out at the users' workplace. Hopefully, it will be able to contribute to future research and reflections in the field, influencing designers and researchers alike, in their work.

Keywords: Cultural context · Culture · User experience

1 Background

About twenty years ago, in 1998, Karen Holtzblatt and Hugh Beyer published their book, Contextual Design: defining customer-centered systems, based on a method they created another ten years back. By then, plenty of computer systems were delivered late and over budget, and did not perform as expected. Unfortunately, it has not changed much, ever since.

The problem with such approach is that it just doesn't work. Today, twenty years later, still many systems are delivered in a far from optimal condition. The costs with rework, change requests, support, and maintenance are still extremely high. There are potentially many reasons why the quality of delivered software is still not good enough and one of the most important ones is that cultural context is almost never considered in the requirements definition.

Work does not happen in a vacuum. Instead, work takes place in a culture, which, according to the authors, defines expectations, desires, policies, values, and the whole approach people take to their work.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 169–177, 2018.
https://doi.org/10.1007/978-3-319-91238-7_15

So, this article aims at exposing the importance of qualitative research in order to understand the cultural context of users. It is a theoretical article in which the authors use examples to demonstrate the importance of the research, especially the research that is carried out at the users' workplace, since it is the one that will give the designer the possibility to know the users of the solutions he/she designs, understanding their needs and problems, process and tasks that he/she performs, as well as the famous '*gambiarra*', a quick fix so frequently seen in computer systems. Hopefully, it will be able to contribute to future research and reflections in the field, influencing designers and researchers alike, in their work.

2 Computer Systems and Culture

Erin Meyer, in her 2014 book, The Culture Map, brings to our attention the most common business communication challenges that arise from cultural differences. Many of the situations raised by such issues can be significantly mitigated if taken into account at requirements definition and design time, and addressed by means of User Experience principles and guidelines.

One of the reasons behind this level of disregard for cultural context is the assumption that all existing differences are due to personality. Actually, before individual differences come into play, people see things through their own cultural lenses, and thus, judge (or misjudge) them accordingly. This is a default and automatic mechanism people acquire from their cultural context. Therefore, it is essential to have an appreciation for cultural context, as well as respect for individual differences.

Jakob Nielsen (2017), in his keynote to the UX Conference last year, pointed out that it is time we made technology work; we need to stop accepting that it always crashes and freezes, and does the wrong thing. He ventured to say that he would like to call for a moratorium on new features for an entire year, or maybe even for two years, where all the programmers in the world would create no new features, not a single one. The programmers would only work on fixing the errors or defects in the features that they had already developed. Based on empirical data, it is known that many of those defects could have been avoided if proper research had been done upfront. In fact, some code might even be unnecessary, since it was written to implement features that were not necessary or wanted, in the first place.

If work takes place in a culture, in order to be successful, a system must fit into their customers' culture. Otherwise, regardless of how well designed and built the system was, or whether it does solve a real problem, it will not succeed. 'Not succeeding' here includes low and lengthy adoption, minimal use, abandonment, high number of incidents, too many unhappy customers, in addition to not selling well.

In order for a system to succeed with a given group of target users, it must not conflict with their self-image. The system must account for the constraints these users are under, and it must not underestimate the values that are important to these users. That would be a guaranteed shortcut to failure.

According to Interaction Design Foundation (2015), the success and failure of a design is not just about one user sitting in front of a screen. It's the result of a broader, more social interaction which means that a designer must have a toolbox full of concepts and methods drawn from the field of Sociology. Simply put, Sociology is the study of social human relationships—and such knowledge is thus necessary to designers.

In a similar way to Psychology, Sociology remains fairly stable because groups of humans still have roughly the same dynamics today as they did a hundred years ago. Therefore, design knowledge based on Sociology is a stable foundation to stand on, even when it seems like the world is moving at an ever-increasing pace.

A person working on a computer in the 1980s used technology that differs largely from the technology used by today workers. Nonetheless, their psychological apparatus are identical, and their needs to get things done in collaboration with their colleagues (i.e. sociologically) are also identical (Interaction Design Foundation 2015). That has not changed much. Both workers still make similar choices when trying to get things done at work.

As Beyer and Holtzblatt (1997) point out, culture influences work by altering the choices that people make. However, the design team that understands these constraints is fully equipped to build systems to account for them.

You may be asking yourself by now, if culture is in fact invisible, how can one get to know it, and build a system that accounts for it. Fortunately, like with many other intangible aspects, it can be deducted from indicators that are observable, that can be seen and heard when the team takes the opportunity to join target users in their environment and spend time with them.

"Culture is to us like water is to a fish - pervasive and inescapable, yet invisible and intangible. Cultural context is the mindset that people operate within and that plays a part in everything they do. Issues of cultural context are hard to see because they are not concrete and they are not technical. They are generally not represented in an artifact, written on a wall, or observable in a single action. Instead they are revealed in the language people use to talk about their own job or their relationships with other groups. They are implied by recurring patterns of behavior, nonverbal communications, and attitudes. They are suggested by how people decorate and the posters on their walls.

The cultural context includes the formal and informal policy of an organization, the business climate created by competitors and by the nature of the business, government requirements, the decor of the site, the self-image of the people doing the work, and the feelings and fears created by the people or groups in the organization" (Beyer and Holtzblatt 1997) (Fig. 1).

The indicators are the tone of the place, the policies, and organizational influences. In the following paragraphs we will examine the indicators, individually.

The first indicator is the tone of the place. If the place looks clean and organized, it tells you that this customer values cleanliness and organization, and thus it is unlikely that they will accept a system that looks messy and cumbersome.

Second, the policies, people do follow policies at work – both formal and informal, even if these are unwritten. Policies can be unveiled through the words people use and relationships they maintain. Most importantly, the words people use will show the policies they care about, which, in other words means, it points out to which problems should be solved first. Policies that are followed let us know how the work is done, what the

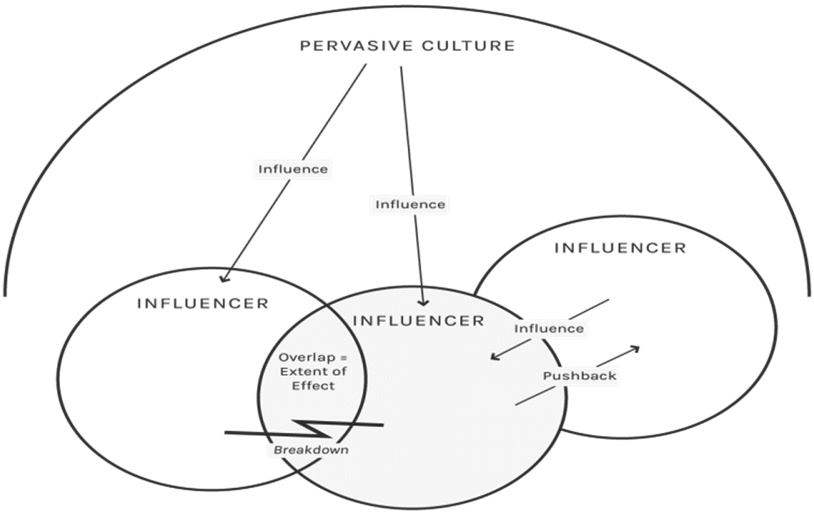


Fig. 1. Cultural model

power relations are like, how flexible processes are or are not, or what kinds of communication are used.

The third indicator is organizational influence, which relates to who or what group is the source of conflicts and irritation in the workplace. When listening to people talking about other people in the workplace, it becomes evident which colleagues, departments, or job functions they trust and/or the ones they would rather avoid. It can be clear the potential workflows that will work.

In order to visualize the data observable and thus, make culture tangible, Holtzblatt and Beyer (1997) created *Cultural Models*. In a Cultural Model, the influencers (those that affect or constrain work) are shown as *bubbles*. *Influencers* can be internal (colleagues, formal groups, "management") or external (customers, vendors, competitors, government). When bubbles overlap, it shows how much the work is affected by this influencer. *Arrows* represent the direction of the influence and the extent to which an arrow enters a bubble shows how pervasive it is. Breakdowns or problems that are particularly harmful are represented in the model as *lightning bolts*.

The influences that tend to be more relevant to design are: standards and policy (e.g. allowed software, gifts, security procedures), power formal and informal (e.g. the weight of hierarchy, knowledge experts influence, family or affective relationships outside work), values of a company or team (e.g. conservative vs innovative, social responsibility), a group's own sense of identity (e.g. 'we don't do process', 'in this team we have a barbecue on the last Thursday of the month') and people's emotions about what they do (positive or negative, e.g. fear of losing the job, professional pride, gossip).

All of the aforementioned elements, as relevant influences on design, can be captured or better understood through qualitative research, especially through observation in the workplace. Interviews are also useful, however, people do not always say what they really think or do. The practice in user research shows that, when interviewed (especially if outside their workplace), people tend to tell what they consider to be an ideal process for accomplishing their daily tasks rather than describing what they actually do, shortcuts they use, and the *gambiarras*[1], which are nothing more than improvised solutions, with whatever means are at hand. *Gambiarras*, at first, are created as quick fix and a firm intention to be temporary, but that is often forgotten, and they become permanent.

Gambiarras are an example of cultural practice that is worth exploring a bit further, since it so ingrained in the Brazilian culture, that it interferes with work in all its forms. For example, considering only the Information Technology field, we can find *gambiarras* both in software and in hardware. Hardware *gambiarras* are those that use components that are not suitable for equipment, such as desk fans to replace computer fans, or using cables, wires and pins of other equipment. Software *gambiarras* are those in which programmers seek to solve organizational and technological problems, such as the absence of qualified professionals in the team, poorly designed systems, lack of time, managerial pressure, low budget, short deadlines, inadequate hiring, low involvement of professionals, etc. with seemingly miraculous solutions, that are to be used temporarily and that, for whatever reason, end up becoming part of the permanent solution. And that means that more effort will be required from the users, in the long term.

Hence, we must not expect to find information about the cultural context of a company in its organizational chart, instead we need to collect qualitative data and find a way to make it explicit. By using Beyer and Holtzblatt's cultural model based on observation, we can model what people think but don't say, what people do rather automatically, and thus, do not say either.

The field research techniques used in Contextual Design are not very different from the techniques used overall in User Research, which stem from Ethnography, although with a higher degree of structure and constraints, so that it can fit the business constraints of Software companies. The cultural model, specifically, can be a very powerful tool to product teams, by providing a solid foundation to build their products upon.

Notwithstanding, in order to capture cultural context, we must embody the traveler metaphor (Kvale 2007), where the Researcher is a traveler on a journey to a distant country in a quest for a tale to tell back home. The Researcher-traveler must walk along with the target users, asking questions and encouraging them to tell their own stories of their world. The Researcher can then transform those stories into cultural models that can be understood and considered by the product teams.

[1] *gambiarra* means an "improvised solution to solve a problem or a need" (PRIBERAM, 2018). It is so common that on Wikipedia there is a page describing the term in several areas. According to the authors, "in programming, *gambiarra* is a palliative (and creative) way of solving a problem or correcting a system in an inefficient, inelegant or incomprehensible way, but that nevertheless works. In Brazil, *gambiarra* in computer programming is also referred to as POG (Programming Oriented to Gambiarra), alluding to the concept of object-oriented programming" (Wikipedia, 2018).

3 Qualitative Research and Cultural Context

Qualitative research is based on text and image data, it has unique steps in data analysis and draws on different investigation strategies (Creswell 2010). Contextual Design, in turn, is an approach to collect data in the context of users, which uses analysis and interpretation of data in an organized way. The purpose of contextual design is to understand the goals, desires and motivations of users, which can only be verified in the field and by asking and observing users (Holtzblatt and Beyer 2013).

Holtzblatt and Beyer (2013) proposed five models of focus group and participant observation to carry out field research, namely:

- “Flow model, which is used to capture communication and coordination to complete a task. This model reveals formal and informal work groups, critical communication points and formal and informal division of responsibilities;
- “Cultural Model, which is used to show the cultural policies that involve the execution of work, explaining what strategies users choose in order to deal with everyday situations;
- “Sequential model, used to show in detail the sequence of steps, strategies and difficulties of the users when trying to complete each step of a work;
- “Physical model, which is used to understand what physical environments prevent or assist in completing a task;
- “Artifacts model, which is used to make explicit the artifacts that are created and used to complete a task.

The contribution of these models in designing a good solution is to bring the users’ mindset and their perspective to the Design and Development teams. It is an efficient way to bring to attention issues and structure of the users’ context, that might otherwise not be perceived, and thus, lead to future confusion, errors, or even more impactful consequences.

Differently from quantitative research, which focuses on numerical data and can be characterized as a linear series of steps moving from theory to conclusions (Bryman 2004), in User Experience, best represented by Mike Kuniavsky, qualitative research is concerned with the natural environment, it has the researcher as a fundamental instrument, it uses multiple sources of data, such as observation, interviews, documents, and audiovisual material. It then performs inductive data analysis, while considering the meanings given by the participants to the study (Creswell 2010).

According to Creswell (2010), when using qualitative research, the researcher can study individuals, explore processes, activities and events, or learn about the broad behavior of individuals or groups with relation to their own culture and how they share it with outsiders. The latter is essential for design and for ensuring a user experience that is appropriate to the context of use, since values, origin, gender, history, interpersonal and intrapersonal relations, socioeconomic status, among others, can shape the users’ relationships and experience.

The quality of the qualitative research is guaranteed by the questions asked (Creswell 2010), as well as the collection of other data and artifacts, such as spreadsheets, notes, memoirs, process maps, training manuals, and even objects that are used to solve one problem or another in the workplace, like for instance the use of a stapler by a user tired of pressing the "enter" key on his keyboard when the system often opens a (or several) confirmation window, as you can see in Fig. 2.

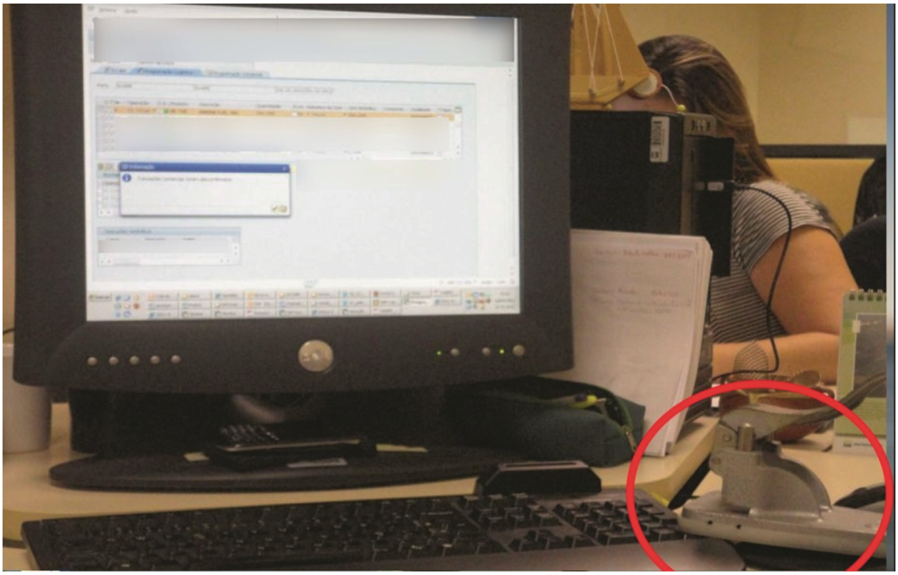


Fig. 2. Use of a stapler by a user to trick the computer and press the "enter" key on your keyboard when the system often opens a confirmation window. Authors' collection.

This type of research seeks to extract meaning from the collected data, considering, mainly, the context and the individuals' current and previous experiences. That is to say, while quantitative research explains the phenomena in the light of hypotheses and theories, the qualitative research understands them from the reflection upon the data collected. Thus, population, location, and the historical moment of the research are considered as fundamental elements that will influence the results, which means that there is no measurement, but rather interpretation; that there is understanding and not explanation; that results cannot be generalized to other contexts, and that the application of the same research in other contexts requires adaptation of the same (Creswell 2010).

For example, in 2015, at a Brazilian bank, when the customer wanted to pay the credit card invoice at the ATM, the machine requested that the customer insert the card and. When the user selects the option '*payment > credit card*' in the menu, she sees the following message: "*Please enter the credit card number or, if you prefer, insert/remove the credit card in the card reader to capture the card number. Do you confirm the transaction?*", as shown below (Fig. 3). However, unless the user knows by heart all sixteen digits of their credit card number – which is unlikely – she is in a deadlock. The message asking the customer to insert and remove the card to read the data does not make sense to the user because the card was already inserted in the machine, but since the system did not recognize the card, of for whatever reason, did not read the card number, it now forces the user to remove the card, copy the numbers to a piece of paper, which would then restart the operation.

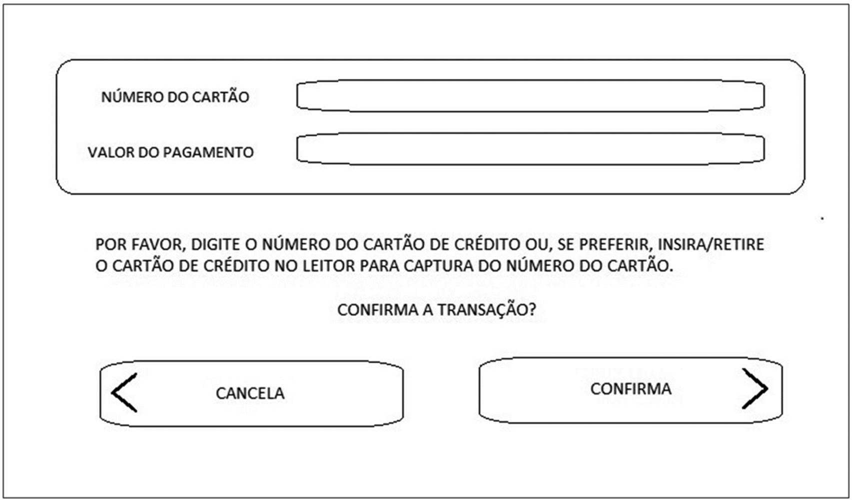


Fig. 3. ATM screen of a Brazilian bank. Authors' collection.

This episode was experienced by the authors in two Brazilian cities and when narrated in a course taught in Brazil, a student, who was an IT professional from the aforementioned bank, was surprised to know that this was happening at ATMs of different cities. According to the student, this had never been explicitly considered, and it was likely that the bank was not aware of the fact. The student took note of the fact details and volunteered to report it the responsible area in the bank.

What does this experience tell us? That the bank did not conduct research with its users? That they did not consider the context of use? That the process was not mapped? That the task was not previously described? For these questions, perhaps the most appropriate answer is "yes and no". It is possible that the bank carried out surveys, considering the clients' profiles and the context of use, in their own way. They might even have had the process mapped. But it is also possible the design was not validated or usability tested after the system went live. So, when the new demands were not met, they brought up not only the non-attendance of users' needs, but also the presence of possible system bugs.

There might also be an indirect impact on other users, other customers who want to use the ATM machine, people bothered with the queue that forms due to the delay in finishing the transaction, and so on. That means, a single, simple example might indirectly create a lot of problems, that could have been avoided if only the bank had done their research and validation.

In view of the above, considering the context of the users, especially when the researcher makes observations in the workplace, one can learn (Spool 2007): (1) processes and terminology: what users actually do and what terminology they use, as they do it, subtleties of their work which they do not realize; (2) context: what external forces influence the design, if the needs change in extreme situations, details of the environment that the user does not know how to describe; (3) similarities and differences: by observing several participants, it is possible to identify more critical needs, which are common to most.

4 Final Thoughts

Currently accepted in several knowledge areas, the qualitative approach lends itself to the study of opinions, values or beliefs of a particular group, showing itself to be very useful in design because it allows a broader, more complete understanding of reality, enabling the generation of more detailed information about human experiences.

Back to the *gambiarras* example, it seems clear that they are bad for the user experience. Once a provisional and apparent solution is in place, it remains for a long time, since it gives the organization the false impression that the problem is no longer an issue. To the organization, there doesn't exist errors or failures that would justify the investment of time and resources. Analyzing and understanding how, when, where and why *gambiarras* happen, can enable us to promote a good user experience, by highlighting to us mainly what not to do for different cultural contexts.

Qualitative research, especially interviews and observation in the workplace, are essential to avoid the generalization and replication, so common in the business software realm. It is from the research with users, especially the one held in the workplace that the design can understand the context of use, the processes, the accomplishment of the task, the needs, the problems and the pains of the user, within their cultural context.

References

Beyer, H., Holtzblatt, K.: Contextual Design: Defining Customer-Centered Systems. Morgan Kaufmann Publishers Inc., San Francisco (1997)

Bryman, A.: Social Research Methods. Oxford University Press, Oxford (2004)

Creswell, J.W.: Projeto de pesquisa: Métodos qualitativo, quantitativo e misto. Artmed, Porto Alegre (2010)

"gambiarra", in Dicionário Priberam da Língua Portuguesa [em linha], 2008–2013. https://www.priberam.pt/dlpo/gambiarra

Holtzblatt, K., Beyer, H.: Contextual Design (2013). https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/contextual-design

Interaction Design Foundation. The 3 Components of Courses from the Interaction Design Foundation (2015). https://www.interaction-design.org/literature/article/is-old-news-bad-news

Kvale, S.: Doing Interviews. SAGE Publications, London (2007). Print

Spool, J.M.: Field Studies: The Best Tool to Discover User Needs (2007). https://articles.uie.com/field_studies/

Nielsen, J.: 10 UX Challenges for the Next 25 Years (Jakob Nielsen Keynote) (2017). https://www.nngroup.com/videos/10-ux-challenges-next-25-years-jakob-nielsen-keynote/. Accessed 2 Feb 2018

How Is User Research Essential in Making e-Government Accessible to All?

Emmanuelle Savarit(✉)

Analyse Concept Ltd., London, UK
e.savarit@analyse-concept.com

Abstract. It becomes important that today's governments include user research in the development of their digital platforms. Taking into consideration users, their needs and the context, has to be at the centre of their approach. User research uses several research methods coming from ergonomics, HCI, human behaviour, cognitive sciences and human interaction. The evidences provided from user research enable to improve and make digital platform accessible to all. This paper will start with an account of User research, its origins, its benefits, and when it should be applied in the product development process. A case study "the Universal Credit Digital Service" (UK) will be presented. 23 participants took part in face-to-face user testing sessions using mobile devices (iPhone, Samsung and iPad). The full user journey from registration to the claim submission was tested. Every step of the journey was analysed screen by screen. Issues in terms of usability, layout and content that could have impacts on the general user journey were identified and will be discussed.

Keywords: e-Government · User research · Citizen · End users
Government portal · User testing

1 Introduction

E-government emerged in late 1990 with the democratisation of the internet [1]. A shared aim was to create a self-service online portal/web base service for citizens that is easy to use and meets government benefits [2].

Lacking citizen experience seems to be a common practice in e-government platform development. We could identify that there is a little coverage of citizen and user satisfaction in the literature between 2000 and 2012.

We can see a recurrence that the development of a digital solution was very often led by policy and technical rather than by user needs. There is some clear evidence that what is missing is the monitoring of citizen experience in the use of government services [3].

It becomes important that today's organisations include user research [4]. At the end of the day, if a government wants to move digital, they must take into consideration their users, their needs and context.

Even if user research is not a new concept, some countries such as France are not yet embedding it within their e-government service development. This becomes obvious, as one can see the difference in terms of user journey, user experience and

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 178–197, 2018.
https://doi.org/10.1007/978-3-319-91238-7_16

user satisfaction. Even if they try to make some efforts, the user experience seems to be "*la cinquième roue du carrosse*", the fifth wheel, which means that user research is generally seen as an optional extra. Bearing in mind that in some countries decision makers have no idea what user research is, let alone its benefits.

For more than a decade, the United Kingdom started to put in place a digital transformation agenda, with the aim to develop an online solution that will benefit the citizens and the government. Their main objective was to reduce costs and provide a self-service online journey. Nevertheless, it was reported that there was a "lack of focus on citizen during the development" [2]. Furthermore, the development of their digital solution was also led by policy and technical rather than by user needs.

One of the major issues was that that they did not involve user researchers at first. They seem to commission external agencies that were conducting Qualitative research (focus groups, interview and field work). Once the report was ready, designers, business analyst and IT team put in place a solution, based on their understanding of the report. Departments were working in isolation, even if we could anticipate some overlap in terms of processes and citizen needs. This led to isolated platform/websites that are not always easy to use.

Since then the Government Digital Service (GDS) department unit part of the Cabinet Office started introducing user research and user needs in their digital strategies. Standardisation of the processes, introduction of multi-disciplinary teams working following agile methodology became a must. GDS emphasised the importance of conducting user research [5].

Scrum teams were set up and 25 exemplars [6] were put in place, introducing user research through the whole product development.

The outcome was that GDS created guidelines for best practices. Even if they are not perfect, as they lack academic background. The guidelines are a very good starting point in the necessity of embedding users through the e-government product/service development.

Every end-user-facing service must follow the guidelines and meet the 26 service standards – that have now been reduced to 18 [7]. Every service that involves citizen-facing must go through the GDS assessment at the different stages of the product development, generally at the end of Alpha and Beta phase. Lack of user research leads to failure of the assessment; in extreme cases the service is switched off.

1.1 What Is User Research, Where Does It Come from, and What Is the Role of User Researchers?

User research uses scientific (qualitative and qualitative) research methods. To understand and evaluate user behaviour, needs, motivation while using a digital product.

User research provides strong evidence on how users react and might interact with a product or service.

The aim of conducting user research is to meet user needs, reduce risk and generate revenue or lower cost. It can also help stakeholders make better decisions, which will result in a more successful product or service.

User research (UR) incorporates usability, human and computer interaction, social interaction, psychology, ergonomics, anthropology etc. UR is becoming a discipline on its own and is now integrated into the agile product development circle [8].

User research is more than usability. User research helps to understand who are the users and what are their needs. It puts the user at the centre of the investigation. UR looks at the full user journey across multiple digital tools (apps, software or digital products). It looks at how the user interacts with the whole ecosystem (devices, apps, software, site, social media, environment etc.).

User research provides valuable information collected using research methods coming from ergonomics, HCI, human behaviour, cognitive sciences and human interaction. These findings are evidence-based. They help web designers, software developers, engineers, business owners, and other stakeholders create better (digital) products; that should in principle be usable by all.

1.2 Why Is User Research Essential in the Development of a Government Platform?

Citizens want to have access to everything at their fingers tips, in no time, through their mobile, laptop, TV etc.

This digital transition transformation affects all organisations, dematerialisation, data management, optimisation of processes, online sales, marketing, cost reduction, or staff reduction where new profiles are required.

Differentiating Client Needs to User Needs: Client/stakeholder needs will be related to the business needs (we want to reduce the number of emails, we want to communicate directly on the intranet, we want to be able to make the financial transaction on our site, we want as a business to get X or Y etc.)

User needs relates to how users will interact with the tools.

Doing User Research Will

- Clarify business needs
- Identify the requirements
- Draw an account of the users by creating the persona
- Test the concept with real users
- Make the recommendation to the UX designers
- Check with the IT what is possible to be done technically.
- Test prototype before development
- Evaluate the functionalities
- Evaluate the architecture and layout of the product
- Evaluate content and terminology
- Evaluate how the users behave and what their needs are

Advantages
User research enables you to work agilely and improve the product through the product development.

Making changes before development is far less expensive than doing it once the product is developed.

Doing user research will limit the risk of failure. It will increase the chance of success and meet user expectation.

The user researchers are objective; they have several research methodologies to capture evidence that will help the business, the IT, the design to make the right decision.

User research can intervene when you want to change the design of your site or if you want to add a functionality.

e-Government Users and Their Needs

Government services are more complex than other services as there are constrained by policies, and there is no limitation in term of users, as it could be any citizen.

Therefore, it is essential to understand who the users are, how they will interact with the service, as well as translate the policy jargon to user-friendly language.

1.3 When Doing User Research?

Involving users throughout the product development is essential, it should happen as early as possible and through every phase (discovery, Alpha, Beta and Live).

Discovery: A short phase, in which you start researching the needs of your service's users, find out what you should be measuring and explore technological or policy-related constraints. (4–8 weeks)

Alpha: A short phase in which you prototype solutions for your users' needs.

You'll be testing with a small group of users or stakeholders, and getting early feedback about the design of the service. (6–12 weeks)

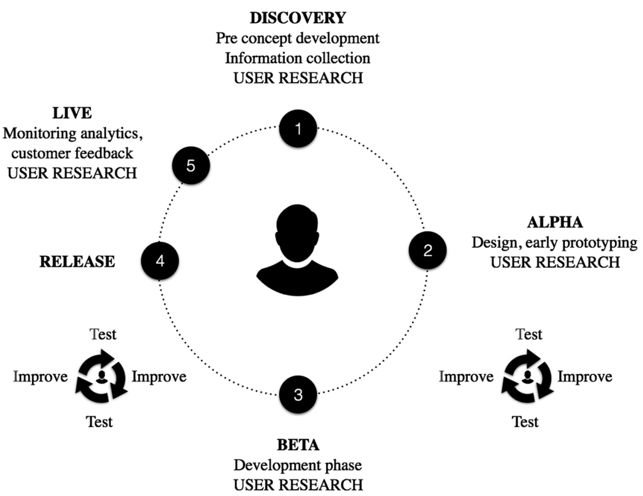


Fig. 1. Integration of user research to the product development circle Source – Savarit 2015 [9]

Beta: You're developing against the demands of a live environment, understanding how to build and scale while meeting user needs. You'll also be releasing a version to test in public. (12–24 weeks)

Live: The work doesn't stop once your service is live. You'll be iteratively improving your service, reacting to new user needs and demands and meeting targets set during its development (until retirement).

2 Case Study: Universal Credit Digital Service

The Universal Credit Digital Service is the biggest European e-government platform and aims to be used by 10 million citizens.

Based on a new policy, the UK Government wanted to create a single platform with the fusion of 6 benefits.

- Income-based Jobseeker's Allowance
- Income-related Employment and Support Allowance
- Income Support
- Working Tax Credit
- Child Tax Credit
- Housing Benefit

The objectives of the Universal Credit were to make a single monthly payment to the citizen, reduce fraud, put a cap on benefits, have monthly follow-ups with the citizen in search of employment, and make it usable for 10 million users [10].

2.1 Background UCDS

The project started 12 months before we started implementing user research. At that point, UX designers were just creating screens based on Policy requirements, or based on research reports which were done by external agencies.

The project was development and policy led without or with very little UX/User Research.

The UX team was composed of:

- 1 content designer
- 2 designers to prepare the screens
- 2 civil servants (1 part-time not on site) both acting as user researchers without previous research qualifications or training.

The complexity of the project was that it involved so many end users. It was shaped by two large policy documents, involved the interaction with other services such as HMRC, as well as many job centre agents and other back-office civil servants.

The basic architecture of the platform was developed offering a hybrid offline and online journey. The registration was online but ID verification was offline, as claimants had to come to the job centre to show a proof of identity.

User research was not recognised as an important part of the product development, for policy people, product owner and/or delivery managers.

It was clear that Stakeholders did not see the importance of spending time with the users despite the GDS recommendation.

Stakeholders did not realise that citizens on benefits may not be good with computers, they may not have internet, or may not have a computer etc. and some may fall in the Assisted Digital categories [11].

Furthermore, stakeholders did not realise that users were not limited to claimants. Users were also every civil servant in the back office that needed to interact with the platform.

The policy was not very popular with the general opinion. The press never missed any occasion to write negative comments. Therefore, poor usability and user experience could have a negative effect on the policy, government, and of course the next election.

Making the product usable for all the users was a challenge but overall was a necessity. Every section of the UCDS platform were broken down in features and prioritized.

Every section of the portal went through the following process:

– *Mini discovery*: to understand the policy requirement, user profile/persona, user needs etc., the business analyst and user researcher have to gather the information.
– *Screen or prototype creation:* based on the BA, UR, designers and content designers who worked on the screens.
– *User testing: user researchers tested the prototype.*
– *Analysis:* feedback and recommendation by the user researcher.
– *Prototype updating:* content and design.
– *Retest prototype:* user testing sessions until good enough to go for development.
– *Validation:* of the prototype, screen user researchers and BA.
– *Preparation of screens:* for development designers.
– *Development.*

Once the every features were developed, we tested the user journey on mobile devices. Traditionally government focused on non mobile platform service, but today it is important to make mobile e-government service in order to increase the access public services to all the citizen [12]. GDS took the direction to develop every platform that will be on the Gov.uk as mobile first [13].

2.2 Mobile and Tablet Testing

20% of consumers check their phone more than 50 times a day [14] Over **30% of UK** adults look at their phone within 5 min of walking. The average instant message (IM) user sends **over 55 IMS** a day. 4G subscriber's numbers expected to exceed 10 million by the end of 2015 [14]. This increased the importance in creating a mobile responsive portal (Fig. 3).

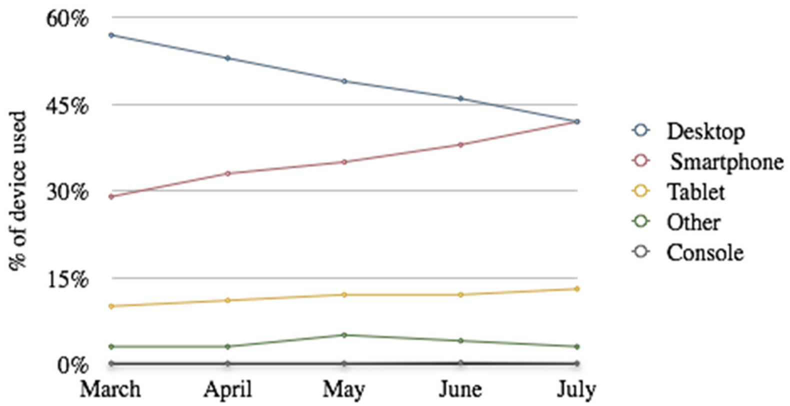


Fig. 2. Percentage of devices used to access UCDS March 2015-July 2015 - Source Savarit (2015) [15]

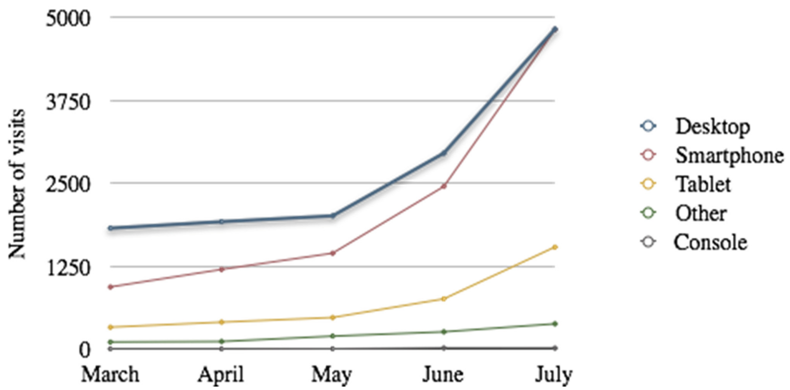


Fig. 3. Number of visits to UCDS per devices March 2015-July 2015 - Source Savarit, (2015) [15]

<u>Web analytics UCDS March – July 2015</u>

The analytics in Figs. 1 and 2 show clear evidence that mobile devices became the primary choice used by citizens to access UCDS. We kept moderating the analytics and projected that the pivot point was going to happen sooner than later.

We decided to carry user testing on the full end-to-end UCDS journey. This is to be done on mobile devices (iPhone, Android and Tablet), aiming to get the maximum amount of feedback to improve the mobile journey, to be easy and satisfactory.

2.3 Approach

Face-to-face user testing sessions were undertaken across 3 days on the: 7th July, 21st July and 4th August 2015 at the Experience Lab facilities, in Holborn.

The sessions lasted between 45–60 min and were based on:

1. Semi-structured/informal interview conducted to keep the session on track
2. Tasks performed based on a script prepared by the researcher
3. We used the UX external demo on iPad2, iPhone 5S and Samsung galaxy S5

2.4 Participants

All participants were 1st time users of the service and should be eligible to the universal credit.

They were recruited by a recruitment agency based on a screener that the researcher prepared. The details of participants are presented in Table 1.

Table 1. Participants information

Session	1	2	3
Devices	iPad	iPhone	Samsung galaxy
Participants	8	7	8
Age - average	33	35	33
IT skills	Average to proficient	Average to proficient	Average to proficient
Gender	6 F & 2 M	5 F & 2 M	4 F & 4 M
Benefits	HB[1], WTC[2], JSA[3], ESA[4]	HB, WTC, JSA, ESA	HB, WTC, JSA, ESA

[1]HB: Housing Benefit
[2]WTC: Working Tax Credit
[3]JSA: Job Seeker Allowance
[4]ESA: Employment Support Allowance

Participants were briefed verbally at the beginning of the session, asking if they were happy to take part in the research, explained their right to withdraw and asked to sign the consent form. Ethical guidance was followed.

At the end of the session, participants were debriefed and given some incentives for their participation.

2.5 Method

All sessions were video recorded and projected in the viewing room. Screen capture as well as the face of the participant were recorded. Notes were taken during the sessions by the researcher as well as by the observers (designers, business analysts etc.), which were in the viewing room following the live session. The notes were summarized in an Excel spreadsheet and the video files were reviewed to confirm what had been identified during the session. Despite the fact that reviewing the video is time-consuming, it enables us to verify if the phenomenon occurred and that it is not influenced by individual perception. Thematic analysis has been used to analyse the outcome of the sessions, with data coded and organised into categories. Similarities in themes could be identified across participants, across sections of the task and across user testing sessions.

2.6 Materials

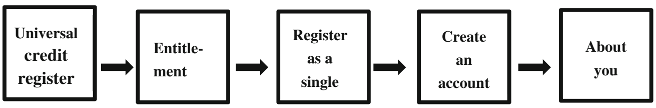
- Discussion guide
- Consent form
- Tablet
- iPhone
- Samsung
- Camera
- Video recording
- SUS
- Excel document for the analysis

2.7 Findings

The sessions confirmed, that users did not have a clear understanding of what universal credit was.

Quote 1: General understanding of the universal credit

Entitlement. To make their claims the users initially had to check if they were entitled to apply to UCDS. At this stage, only a few postcodes were taking part in the private beta. Every claimant had to go through the postcode check which we called the entitlement. Once the user was eligible to use the service, they had to register.

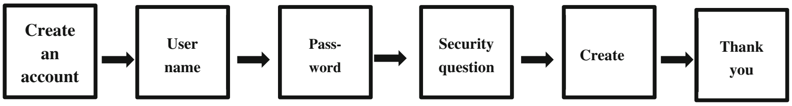


User journey 1: Entitlement, register and create account

Participants did not experience any issues in checking if they were eligible to use the service. They used their postcode and all participants managed to check their eligibility autonomously.

After, they had to select if they were doing a single claim or a couple claim. All participants taking part in the research were doing a single claim.

Create an Account



User journey 2: Create an account

Participants were asked to create a user name, password and to answer two security questions.

Password. At this stage, participants find some **difficulties to move beyond** the username/password.

The user testing on mobile devices brought even more evidences towards participants who were struggling in the creation of their account. It was also more prominent to create an account on the smartphone (iPhone and Android) than on the iPad.

Most of the users needed between 3 to 5 attempts to go through the "create an account" process. Participants that were not computer savvy did not manage to go through it and required the help of the moderator.

Furthermore, despite the instructions available to the users, and the colour coding which was in place, they struggled to go through the account creation process.

The findings from the user testing show that participants were not reading the instructions (symbols, letters, upper/lower cases and number) that were required.

We had evidence from previous user testing on desktop, that users had difficulties to create an account. The reason was due to the complexity of the password requirement. Even if we made the recommendation to simplify the password requirement, the security service declined our recommendation.

On mobile devices, participants did not see the errors symbols in red, even if it was more prominent on the smartphone than on desktop.

Therefore, we had to find other solutions to facilitate the journey of the password creation.

Security Questions. To create an account, once the user managed to create a password successfully, two security questions were presented to the users to be able to retrieve their account in case they lose their login details.

We identified different types of issues:

Content Issue. Based on the reaction from the users during the testings, it seems that two security questions was too much.

Furthermore, the content of the security question seems to be a bit outdated as some questions did not apply to the type of users (e.g. what was your first car? - some users may never have owned a car).

Usability Issue. iPhone users could not see the totality of security questions, as they did not fit the carousel preventing the user to read the questions. Users did not switch their phone instinctively on the horizontal mode to see the questions. Even if few managed to turn their phone, the content still did not fit the screen. Furthermore, the position of the carrousel at the bottom of the screen was not visible to the user. Many people do not see the carousel.

On the other devices (iPad & Samsung S5) we did not identify any issues.

Quote 2: Security questions

Once the participants completed created a username, password, selected the two security questions and their answer, they successfully created their account. This was taking certain users up to 5–10 min.

About You

User journey 3: About you

This section was for the user to confirm their email address. The main issue was that the users did not read the content and the instructions.

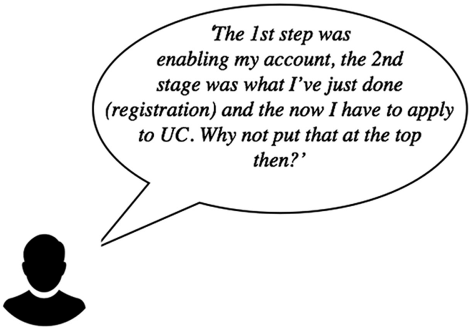
Users just realised that they have made a mistake once they have submitted the page 'apply for UCDS' in the user journey 3.

Confirm My Email. The confirmation of the email was straight forward. Users were expecting to get an email confirmation in their mailbox. No issue (content or usability) was identified, the layout on the mobile device was responsive.

Home Address. Some users mentioned that they had already entered their postcode when they went through the entitlement screen, and questioned why they were asked again. They expected their postcode to be kept in the system. No other issues were identified.

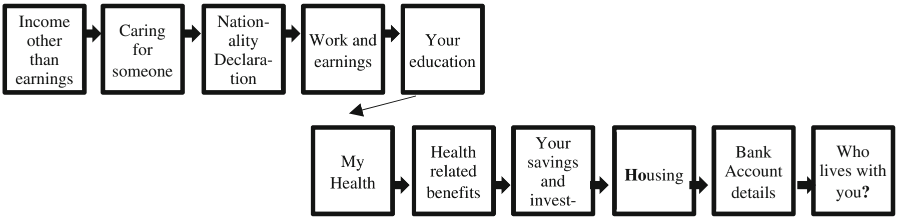
Quote 3: Home address and postcode

Apply For Universal Credit. The process was straightforward and the users managed autonomously to go through the process. The only issue identified was that the user did not know at which stage of the journey they were. They had no visibility of how far they were in the process of making their claim. Furthermore, they make the remark that they have no visibility of how long it was going to take them.



Quote 4: Information about the process

To Do List. Once the user clicked on to apply for universal credit, they landed on a screen with a list that "we" call the "to do list". This covered the following sub-Section

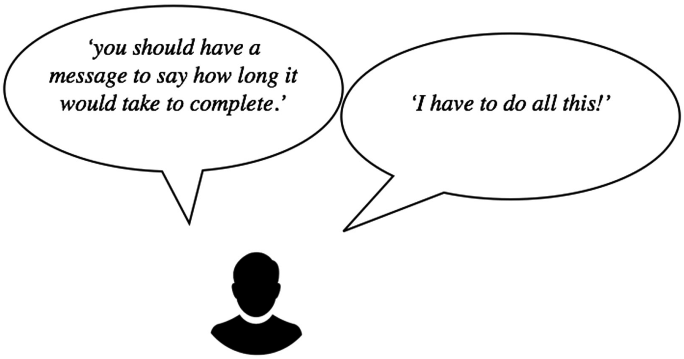


User journey 4: To do list

1. Income other than earnings
2. Caring for someone
3. Nationality declaration
4. Work and earnings
5. Your education
6. My health
7. Health-related benefits
8. Saving and investment
9. Housing
10. Bank account details
11. Who lives with you?

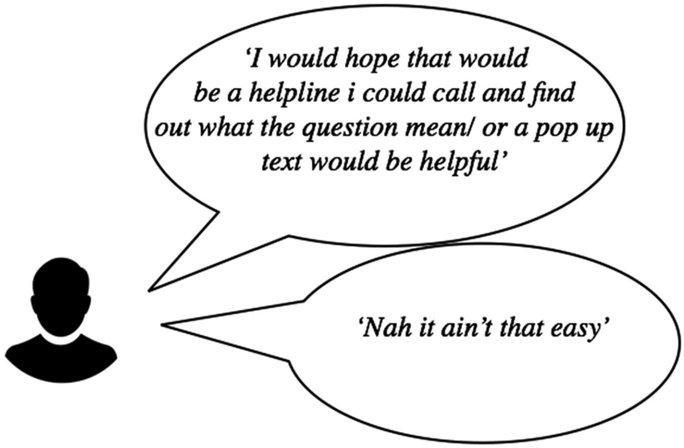
We could identify a mixed reaction from the users. All of them were quite overwhelmed by the length of the "to do list"; most of them made faces or commented on the list. We previously had this comment on desktop user testing, it was also more visible on smaller devices as the user had to scroll down. They all understood that they had to complete each section. Despite this fact, they believed that some of the sections were not relevant to them.

The order of the "to-do list" was not perceived as logical. Some users started with the one that seemed easiest to complete (e.g. housing, earning was easy whereas saving and investment tended to be kept for the end).



Quote 5: Time to apply

Income Other Than Earning
Users did not understand the meaning of the terminology "Income other than earning". They asked the moderator what it meant, we asked them what they will be doing if they were on their own to complete their claim? Most of them respond that they will have contacted the helpline to get more explanation. This comprehension issue was clearly stopping the users from completing their initial application.



Quote 6: Income other than earning

Caring for Someone
This referred to the people who were caring for someone who had a disability. This question was generally for people that were getting or eligible for the carer allowance. Some users thought that this section also was for parents that had children. They had to click on the link to see what this section was about.

Nationality Declaration
No issue identified

Work and Earnings
Content issues: There is no clear understanding from the user of what work and earnings meant (e.g. if they had to include benefits, maintenance etc.)

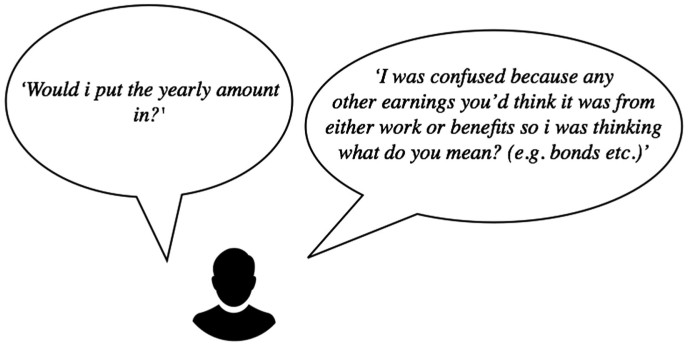
It was also not clear whether they had to put in their monthly or annual income.

There was also an issue for those on '0' hour contracts (e.g. they don't know until the end of the month how much they're going to make)

Once again they did not read the instructions before entering the numbers.

Layout issues: The box on the screen to enter the number was divided into two boxes, one for the pound and the other for the pence. Users failed to recognise in which box they had to enter the numbers of their earnings.

The errors message was only activated once they have submitted the screen.



Quote 7: Work and earnings

Your Education
We did not identify any usability issues for this section. However, at first, the users did not understand what this referred to (e.g. do they need to provide their education history).

Quote 8: Education

My Health
We did not identify any issues with this section

Health Benefits
We did not identify any issues with this section

Saving and Investment
Most of the applicants did not think this section applied to them. The ones who had a bit of savings were worried that if they declared their savings, they will not qualify for their UCDS benefit. Information of the cap would be needed to reassure users.

Housing
This section was quite long for the users and several questions were asked. We did not identify any functionality or usability issues; nevertheless, the content was an issue.

Content issues: Users did not understand which figure they had to put in the box. Many users were on housing benefits (Housing benefits were at that time a separate benefit), which very often was paid directly to the landlord or the housing association. The users did not know if the amount required was the full amount of their rent or if it was the amount not covered by the housing benefit (the difference between the rent and the housing benefit that they were currently paying.)

We came across the fact that many users that get their housing benefits paid directly to their landlord or housing association had no idea of how much their rent was.

The section related to the **'frequency'** was not clear to the users. In the UK, rents could be paid: monthly, weekly, biweekly or every 4 weeks. However, users did not understand that some people may have different rent payment frequency. Some instruction, such as how often do you pay your rent, could have been added above the options to make it clearer to the users.

The section related to the **'service charge'**, was also unclear to the users, many users did not know what this section was referring to.

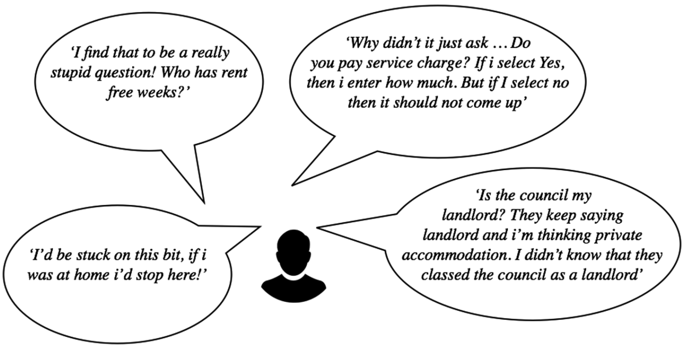
The section "free weeks" was also not very clear to users that were not entitled to this type of benefit.

The section that asked to provide the address of the landlord was not clear to the users that did not have a private landlord. More information and guidance were needed.

There was some clear evidence that the content on this section was not clear and confusing to the users. Most of them completed the form but we could question whether the answers were accurate.

Most of the users will have contacted the helpdesk or even some of them will go directly to the job centre to get some support.

This overall section was perceived as being too long and congested on smartphones. However, on the iPad, the general overview was positive probably due to the screen dimension.



Quote 9: Housing

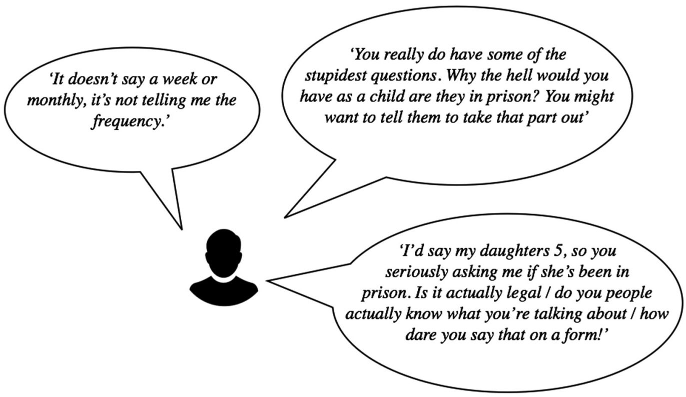
Bank Account Details
There were no usability/content issues in this section

Who Lives With You?
We identified two key issues in this section:

Content issue: The users had some difficulties in entering the childcare costs, there was no clarity about which amount and the frequency of the childcare cost to be entered (e.g. weekly, monthly, etc.).

The question ‘Had your child been in prison?’ caused a lot of animosity amongst users who clearly showed their displeasure during the session and recommended that this were to be removed! This is a policy requirement; nevertheless, how many claimants have children who have been in prison? A more discrete way may need to be considered.

In addition, this section was also perceived to be too long (e.g. lot of content on a small screen) to get through on a smartphone.



Quote 10: Who live with you?

Confirm Details
There were no usability issues in this section. The feedback was more positive than on the on Desktop

Submit Claim
There were no usability issues in this section.

SUS Results. The System Usability Scale (SUS) provides **a high-level satisfaction** score of usability of a site, application or any technological item.

SUS is a simple, ten-item scale giving a global view of subjective assessments of usability.

The SUS scale is generally used after the respondent has had an opportunity to use the system being evaluated, but before any debriefing or discussion takes place.

A SUS score has a range of 0 to 100. It is not a percentage.

A score above 68 is above average and a score below 68 is below average Table 2.

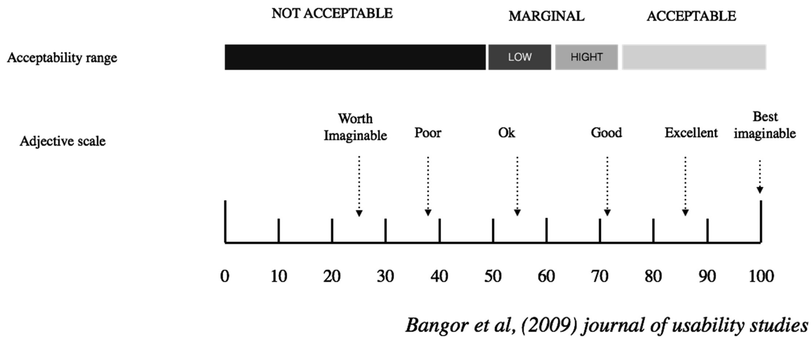


Fig. 4. System Utilisability Scale (SUS) interpretation

Table 2. System Utilisability Score score per devices

	iPad	iPhone	Samsung
SUS score	85.6	77.9	79.4

The overall the score across the three devices was of **80.9**, which is acceptable between good to excellent.

We could see that the iPad had a higher score and that the iPhone the lower one. Nevertheless they are all scoring in the acceptable range, between good to excellent in reference to the Fig. 4.

3 Conclusion

The overall user journey on mobile devices was positive. We could identify some usability issues at the beginning (create account, registration). Password and security questions needed to be improved, otherwise this could have a major impact on the online user journey, which will increase the number of calls to the helpline or visits to the job centres.

Most issues identified were related to content understanding, policy jargon that needed to be translated to a more accessible terminology (e.g. housing, child care cost etc.). Furthermore, making clearer and sign posted guidance would help and facilitate the overall user experience.

We could also identify that the general user experience was more satisfactory on iPad, than on Samsung, the iPhone was not as good, due to the security questions.The screen size could also have caused some problems (e.g. amount of content displayed on each screen.)

Despite those issues, the general experience on mobile was positive, this was important as the number of citizen using mobile devices to access the UCDS platform keep increasing based on the analytics. The user testing on mobile devices showed how users interacted with the platform and their experience in-situ provided feedback that permitted to improve the overall user experience.

The fact that user research was involved at every steps of the product development enabled to fix issues on the spot every time we identified them. In this case study, we tested the whole user journey on mobile devices. The findings show that some issues in terms of content were still present, and work still needed to be done to make the portal easy to use. We could also conclude that the fact that we have introduced user research and user testing systematically for every feature has reduced the number of issues. By involving users and taking a user centred approach, we optimised product development, reduced high risk issues, and constantly evaluated functionalities and content.

Further research should take place with people which could be qualified as Assisted Digital (people with very little or no knowledge of computer, or people with cognitive limitation), as an e-government platform must be accessible and usable to all the citizen.

References

1. Grölund, Å., Horan, T.A.: Introducing e-Gov: history: definitions and issues. Commun. AIS **15**, 713–729 (2004)
2. Hamilton, F., Pavan, P., McHale, K.: Designing usable e-government services for the citizen. Int. J. Publ. Inf. syst. **3**, 159–167 (2011). http://www.ijpis.net/issues/no3_2011/ijpis_no3_2011_p6.pdf. Accessed 7 Jan 2018
3. Petorius: categorisation of digital government services informed by user research. In: Proceedings of the 17th European Conference on Digital Government, At Lisbon, Portugal (2017). https://www.researchgate.net/publication/318113702_Categorisation_of_Digital_Government_Services_informed_by_User_Research. Accessed 7 Jan 2018
4. Portigal: Interviewing user. How to uncovercoppelling insights. Rosenfeld media (2013)
5. Government Digital Service – gov.uk: User research (2016)a. https://www.gov.uk/service-manual/user-research. Accessed 10 Jan 2018
6. Government Digital Service – gov.uk: 25 exemplars (2013). https://www.gov.uk/government/publications/gds-transformation-programme-2013-to-2015/digital-transformation-exemplar-services. Accessed 10 Jan 2018
7. Government Digital Service – gov.uk: Service manual (2016)b. https://www.gov.uk/service-manual/service-standard. Accessed 10 Jan 2018
8. Savarit, E.: (in progress) The A to Z of User research; a practicle guide to incorporate user research to your product development
9. Savarit, E.: Get Your Research Skills Up to Speed: An Introduction to User Research. UX, Cambridge (2015)
10. Department work and pension – gov.uk (2010). https://www.gov.uk/government/publications/universal-credit-welfare-that-works. Accessed 15 Jan 2018
11. Government Digital Service – gov.uk: c Assisted digital (2010). https://www.gov.uk/service-manual/helping-people-to-use-your-service/assisted-digital-support-introduction. Accessed 19 Jan 2018

12. Hung, S.-H., Chang, C.-M., Kuo, S.-R.: User acceptance of mobile e-government services: an empirical study. Gov. Inf. Quater. **30**(1), 34 (2013). https://doi.org/10.1016/j.giq.2012.07.008
13. Government Digital Service – gov.uk: designing for different devices (2012). https://gds.blog.gov.uk/2012/11/02/designing-for-different-devices/. Accessed 19 Jan 2018
14. Deloitte: global mobile consumer trends, 2nd (edn.) Mobile continues its global reach into all aspects of consumers' live Mobile Consumer report (2017). https://www2.deloitte.com/global/en/pages/technology-media-and-telecommunications/articles/gx-global-mobile-consumer-trends.html. Accessed 12 Jan 2018
15. Savarit, E.: Mobile devices Universal Credit Digital Service: User testing report. Department Work and Pension (2015). Accessed 6 Jan 2018
16. Government Digital Service – gov.uk: Assisted digital (2010). https://www.gov.uk/service-manual/helping-people-to-use-your-service/assisted-digital-support-introduction. Accessed 10 Jan 2018
17. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. **4**(3), 114–123 (2009). http://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/. Accessed 08 Jan 2018

Check for updates

Addressing IoT: Towards Material-Centered Interaction Design

Mikael Wiberg(✉)

Department of Informatics, Umeå University, 901 87 Umeå, Sweden
mikael.wiberg@umu.se

Abstract. This paper takes a point of departure in how IoT - the Internet of Things - is increasingly described as the next step forward for digitalization. As a background to this trend I describe how a great number of applied research projects and development efforts has been conducted to address various specific needs. Further, I argue in this paper that there is still a lack of a stable knowledge base – including developed theories and methods - for working across physical and digital materials in the design of IoT solutions. Motivated by this identified lack of methods this paper presents a theoretical and empirical ground for the development of a material-centered approach to the design of IoT systems. The proposed method is focused on material interactions as an approach for working across physical and digital materials in design. In more particular terms this paper (1) describes how this proposed method adds to this current body of research in HCI, (2) it presents a model for doing material-centered interaction design, and (3) it outlines some methodological implications for the development of a method for the design of IoT systems. Finally, this paper introduces an empirical case to serve as a demonstration of the need for such methods in practice as to address IoT, and as to push the design of IoT systems forward.

Keywords: IoT · Internet of Things · Material-centered interaction design

1 Introduction

IoT - the Internet of Things - is increasingly described as the next step forward for digitalization in our society. IoT has recently been described as a driver for sustainability - including the development of smart homes and smart cities, for e-health solutions, and for learning. Further on, it has been proposed as a solution for more energy efficient transportation solutions, including logistics. In short, IoT has been proposed as a solution to a number of societal challenges.

In reviewing the current initiatives taken in this area it is not an understatement to say that the current expectations are high on these IoT systems to address a wide range of societal challenges. For instance, and in terms of how IoT systems have been proposed as a driver for sustainability we can notice this orientation under the emergence of IoT enabled areas such as smart homes, smart cities, and smart transportation solutions) (Firner et al. 2011; Pan et al. 2013). Further, we notice initiatives taken on using IoT to address health problems (Savola et al. 2012; Garcia, et al. 2017; Chishiro et al. 2017)

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 198–207, 2018.
https://doi.org/10.1007/978-3-319-91238-7_17

and as an emergent opportunity in the areas of Financial technologies and Cyber Security to utilize IoT technologies in the design of smart systems – ranging from NFC and RFID solutions for mobile payments to finger print based encryption for mobile devices.

Here it should be noted that all of these IoT solutions build on "*material interactions*" as the central interaction model, and that these solution are all dependent on the following three factors: (1) tight integration of computing and networking with physical materials and objects (Want 2015; Romano 2017), (2) alignment to people's needs (Pignotti et al. 2014), and (3) the development of methods and approaches (Wiberg 2013; Karana et al. 2016; Garbajosa et al. 2017) for working across digital and physical materials in interaction design projects (Sulistyo 2013).

However, and despite the current efforts made in this area to address societal needs, and despite these identified dependencies we still lack a stable knowledge base concerning interaction design built on material interactions – in particular in terms of (1) *understanding user needs* and (2) *validated design methods*. The consequence is that resource consuming trial-and-error approaches are currently applied. Furthermore, the overwhelming risk is that the current expectations on IoT to solve societal needs might not be fulfilled.

Motivated by these identified needs the aim of this paper is to propose a design method for doing *material-centered interaction design* that acknowledge and correspond to user needs while being relevant for the design of IoT systems that heavily depend on tight integration of computing and networking with physical materials and objects.

2 Proposing a Method for Material-Centered Design

With a point of departure taken in this background this paper proposes a method focused on *material interactions* as a unique approach for working across physical and digital materials in design.

Material interactions is a new approach to interaction design that stretches across physical and digital materials and approaches to the design of interactive systems and networked products (See e.g.: Wiberg 2018; Wiberg et al. (2013a); Wiberg et al. (2012; 2013)).

In the construction of the method proposed in this paper we take stock in the increasing interest in our community to address interaction design through a material lens. This includes for instance Jenkins (2015) approach to prototyping material interactions for IoT - the internet of things, material programming (Vallgårda et al. 2016), and material probes (Jung and Stolterman 2011). These approaches belong to a growing strand of research in interaction design where new design approaches to the materiality of IoT systems are currently being explored - including the work by Berzowska (2012) on approaches to programming materiality and the work by Karana et al. (2016) on craft-based approaches to the tuning of materials.

In this paper we describe (1) how our proposed method adds to this current body of research in HCI, and (2) we present a model for doing material-centered interaction design, and (3) we outline the implications from this proposed method for the design of

IoT systems. Finally, we conclude the paper with a practical case that illustrate the need for such methods in practice followed by a draft of this method and suggestions about three cornerstones of importance for the further development of this method to address and scaffold the design process of developing new IoT systems.

3 Theoretical Ground for a Material-Centered Approach

In proposing a Material-Centered Approach to Interaction Design in general, and in particular in relation to the development of IoT systems we ground this proposed method in the growing body of research on *material interactions* in the area of HCI/interaction design research (See e.g.: Wiberg 2018; Wiberg et al. (2013a); Wiberg et al. (2012)) and current efforts made in interaction design research on addressing interaction design from the viewpoint of its materiality.

This notion of 'materiality' is a growing theoretical perspective in interaction design research. As formulated by Wiberg et al. (2013b) in their paper "Materiality matters – Experience Materials" it denotes a new perspective that enable design across physical and digital design, and it has its roots in Ishii's pioneering research on tangible interaction design (see e.g. Ishii and Ulmer 1997 and the most recent Ishii et al. 2012). As further suggested by Wiberg (2016; 2018) this approach is now expanding into new hybrid forms of digital products, including smart watches, smart cars, and the Internet of things – and it signals a trend toward combining digital and analog materials in design. As interaction with these new hybrid forms is increasingly mediated through physical materials interaction design is increasingly a material concern. One could even argue that the "material turn" in human-computer interaction has moved beyond a representation-driven paradigm (Robles and Wiberg 2010; Wiberg and Robles 2010), and in relation to this Wiberg (2018) has recently suggested that this idea of "*material-centered interaction design*" might work as a new approach to interaction design and its materials. This approach embraces a view of interaction design as a practice of imagining and designing interaction through material manifestations. Further, a material-centered approach to interaction design enables a fundamental design method for working across digital, physical, and even immaterial materials in interaction design projects.

This proposed method also takes an explicit point of departure in a set of related and recent research approaches to material-centered interaction design. This includes e.g. Giaccardi and Karana's (2015) approach to understand *material experiences,* Jenkins and Bogost's (2014) approach to *prototyping material interactions* for the internet of things, *material programming* (Vallgårda et al. 2016), and *material probes* (Jung and Stolterman 2011). These approaches belong to a growing strand of research in interaction design where new design approaches to the materiality of interactive systems are currently being explored – see e.g. Wiberg 2018; Wiberg et al. 2012 including aspects of the *form* (Jung and Stolterman 2011), and *agency* (Tholander et al. 2012), of interactive materials and systems and how they are *experienced* (Pignotti, et al. 2014). In terms of additional methods developed for working with this materiality this perspective is for instance reflected in the work by Berzowska (2012) on approaches to *programming* *materiality* and the work by Karana et al. (2016) on craft-based approaches to the *tuning of materials.*

While there is this growing body of research on material-centered approaches to interaction design there is also an even more stable research area on user-centered interaction design that we can turn to here as to address IoT in relation to both user needs and computing in the form of material configurations. In fact, it should be highlighted here that over the last 30 years the methods for studying user behaviors and to work with users in interaction design projects has been constantly refined and developed. Today there are specific methods for doing user studies (ranging from ethnographic observational studies), to more controlled experiments (c.f. Nielsen et al. 1993, Karat 1997, Dix et al. 1998, Virzi 1997, Wiberg 2003). Further, if reviewing the basic knowledge base for working across physical and digital materials in design of interactive systems we need to acknowledge that the area of industrial design serves as a stable knowledge base related to the design of *things* (including product design) and, in similar terms, we have a stable knowledge base that can inform the design of digital services.

To summarize, there is at the current moment (1) a new knowledge base forming around the tight integration of physical and digital materials in design, and (2) there exist a well-established body of research and methods for doing user-centered design. However, less is know about how to integrate these two strands. Accordingly, and in line with the research question formulated in this paper we need to develop new design methods for doing material-centered interaction design that acknowledge and correspond to user needs. In this paper we suggest that this is key for the further development of solutions that is built around material interactions as a central design principle in interactive systems design in general, and in particular when it comes to the design of IoT systems.

4 Exemplifying a Material-Centered Approach to IxD

I have now motivated the proposed material-centered approach from a theoretical viewpoint, but before moving forward to a draft where we present the cornerstones of one such method or approach I will now also illustrate how the experimentation with the integration of physical and digital material play out in practice.

So, as to illustrate this focus on material interactions and what it can look like when doing material-centered interaction design I will now in the following section present an example from our ongoing collaboration with the IT-company KnowIT in Sweden.

Here, this particular example illustrates material-centered interaction design in the context of an ongoing IoT project at KnowIT in Umeå. In this particular project the design challenge was to design an interactive system for a smart parking service. The problem addressed concerned how to design an interactive system that automatically can detect if cars occupy the parking lots, and to present this data to the driver who wants to park his/her car. Some commercial solutions are already in place for some parking garages that builds on counting cars and displaying the number of available slots on a screen, but more lightweight solutions that builds on data from the cars actually parked in each parking lot are still not available. In relation to this design challenge the IT-company KnowIT is currently exploring different solutions to this problem trough a trail-and-error approach.

Their first approach, Fig. 1(A, B and C) is to use ultrasonic sensors for object detection as to determine if a car is parked in a particular parking lot (1A). If that is the case the sensor detects that and the data is transmitted via a LoRa-network (1B), to a server and then a script sends a signal to an arduino board to turn a lamp on in the window (1C) so that a driver can see if the parking lot in the basement is occupied or not. Here, a combination of sensors, networks, server tech, and a lamp is used for the system architecture, and the position of the car serves as input to this system. Accordingly, this solution demonstrates design across digital materials and physical objects (in this case the position of the car), to offer a new service to the user, i.e. to meet a particular user need. However, this first approach has its limitations. It needs sensors installed for every parking lot, and it demand one representation/parking lot (in this case a lamp2).

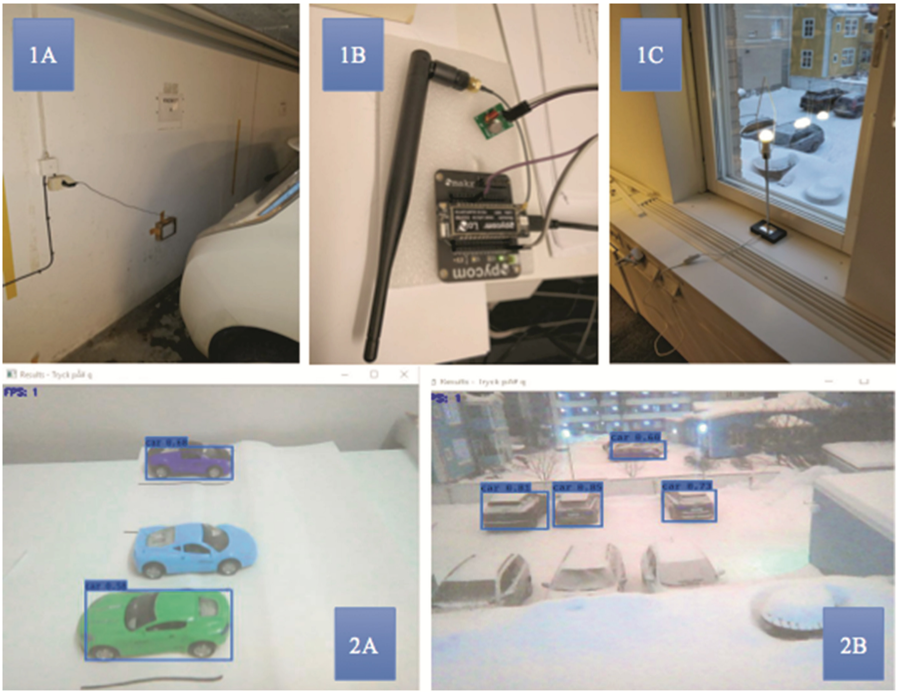


Fig. 1. Two examples of design explorations in the area of material interactions at KnowIT.

To overcome these problems KnowIT is now exploring an alternative solution (2A and 2B). For this second solution KnowIT is experimenting with computer vision. By using a camera and the YOLO v2 framework for object detection the camera can read the whole parking lot and look for parked cars in the parking spaces. To test this solution they first did a small-scale test with a number of toy cars (2A), before going for a full-scale implementation (2B). This small-scale design allowed to test the system by moving the cars around, whereas the full-scale test allow for testing the solution in different weather conditions (e.g. during a full snow fall as in Fig. 2B).

Again, this second solution also builds on the tight integration of physical and digital materials, but a completely different material configuration (in this case a combination of camera tech, computer vision for object detection, and the positions of the cars parked as input to the interactive system). As such this short example illustrates how different material configurations can enable a particular digital service. Further, it illustrate how different material configurations might solve a particular problem, in different ways, but also with different associated pros and cons – an aspect that again is a call for more systematic approaches and methods for design of interactive systems that depend on the interplay between physical and digital materials.

5 Discussion – Towards a Material-Centered IxD Method

So what can be learnt from this particular case in relation to the development of a method for doing material-centered interaction design in the context of IoT systems?

Well, first of all this practical case illustrate that it is possible to come up with different designs in relation to the same design challenge, second that each solution builds upon different material configurations, and finally, that each solution integrate the physical and the digital in different ways. For instance, in the first case the integration between the physical and the digital is made by using a proximity sensor to measure if a (physical) car is near the (digital) sensor, whereas in the second solution it is a (digital) camera that uses the YOLO v2 framework for object detection as to detect if a (physical) car occupies a parking lot or not.

On a more general level, and if now starting to draft the skeleton of a method for doing material-centered interaction design we can notice how a material-centered approach for sure has materials as a central concern. From a design viewpoint it is about understanding how different materials can be combined in the design of new interactive solutions, and that in return demands an understanding of what materials are available in the first place, second how different material properties can be used in the design, and further to have an understanding of how different material properties can be set in relation to each other in the design of interactive systems. In relation to IoT systems design I would say that this calls for a skill that I would like to label as a "material sensitivity", i.e. an ability to carefully consider how different materials could be brought together in the design of an interactive system.

Further, and if now moving towards an understanding of how this might be done, that is to turn this into a method and approach, there are again a couple of things to consider. If truly subscribing to a material-centered approach, then such design becomes an activity of carefully shifting between imagining what different materials might add to the design, and really hands-on explorations of how such combinations of materials actually work in the design. In short, a dialectic process that is constantly shifting between reflecting upon, and trying out, different material solutions. As such, the material-centered approach becomes a craft-based practice that is both about intellectually exploring and imaging computing in material form, as much as it is about crafting, experimenting and building new solutions with the materials at hand.

In moving forward I would now like to take three quotes from the newly released book “The Materiality of Interaction” (Wiberg 2018) to discuss three cornerstones of importance for the establishment of a material-centered approach to interaction design.

In doing so I first start with the following quote that is about how a material-centered approach recognizes *the multitude of materials available* for the design of interactive systems:

“If computing is no longer limited to one single substrate (digital materials), and if the set of available materials (digital, physical, and smart) is growing at a rapid pace, then the biggest challenge is not finding ways to manifest interaction in material form; instead, the challenge is navigating this landscape of available materials and devising a method and an approach for doing this job”. (Wiberg 2018)

Further, the following quote illustrates that among these different materials we should make no categorical distinction between different matters. In short, there is no point in making a separation between physical and digital materials!

“It is important to remember is that a material-centered approach to interaction design does push for looking at interaction design through a material lens. In doing so it makes no ontological or metaphysical distinction between digital and physical materials” (Wiberg 2018).

Finally, the third quote that I want to introduce here illustrates “the interaction first principle” which is a central cornerstone for a material-centered approach to interaction design. It is central as it illustrate how a material-centered approach to interaction design at the same time can be established with user needs, i.e. the interaction being supported, as its main and foremost concern:

“The interaction-first principle is about conceptually defining the mode and form of the interaction being supported. It is about defining who the user is and how he or she will interact with the interactive system”. (Wiberg 2018)

With these three cornerstones as a point of departure, in the theoretical grounding in material interactions, in the current literature on “materiality” and how it plays out in the area of interaction design, and with this practical case as introduced in the paper as an empirical illustration I suggest that a draft of a method for doing material-centered interaction design can include the following three components and main activities:

(1) **Exploring and defining** the form of interaction being designed
(2) **Exploring and evaluating** the range of possible materials that can be used for designing the interaction
(3) **Working iteratively** between the integration of different materials in the design of the interaction, while recurrently revisiting and if necessary revising the initial idea of the interaction.

Further, I should here also underscore that these three parts of one such method works well together with the “interaction-first principle” as to do material-centered design in close relation to user needs.

6 Conclusion

In this paper I have taken a point of departure in how IoT (the Internet of Things) is increasingly described as the next step forward for digitalization.

As a background I have described how a great number of applied research projects and development efforts has been conducted to address various specific needs, and I have argued that there is still a lack of a stable knowledge base – including developed theories and methods - for working across physical and digital materials in the design of IoT solutions.

Motivated by this identified lack of methods I have in this paper presented a theoretical and empirical ground for the development of a material-centered approach to the design of IoT systems. The proposed method is focused on *material interactions* as an approach for working across physical and digital materials in design.

In more particular terms I have in this paper (1) described how this proposed method adds to this current body of research in HCI, (2) I have presented a model for doing material-centered interaction design, and (3) I have outlined some methodological implications for the development of a method for the design of IoT systems.

Finally, I have introduced an empirical case to serve as a demonstration of the need for such methods in practice as to address IoT, and as to push the design of IoT systems forward.

References

Berzowska, J.: Programming materiality. In: TEI 2012: Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction. ACM (2012)

Chishiro, H., Tsuchiya, Y., Chubachi, Y., Bakar, M., Silva, L.: Global PBL for environmental IoT. In: ICEEG 2017: Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government. ACM Press (2017)

Dix, A., Finlay, J., Abowd, G., Beale, R.: Human-Computer Interaction. Prentice Hall, Europe (1998)

Firner, B., Moore, R., Howard, R., Martin, R., Zhang, Y.: Smart buildings, sensor networks, and the Internet of Things. In: SenSys 2011: Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems. ACM Press (2011)

Garbajosa, J., Magnusson, M., Wang, X.: Generating innovations for the internet of things: agility and speed. In: XP 2017: Proceedings of the XP2017 Scientific Workshops. ACM (2017)

Garcia, C., Fernandes, P., Davet, P., Lopes, J., Yamin, A., Geyer, C.: A proposal based on IoT for social inclusion of people with visual impairment. In: 2017: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web. ACM Press (2017)

Giaccardi, E., Karana, E.: Foundations of material experiences: an approach to HCI. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015), Seoul, Republic of Korea, 18–23 April 2015, pp. 2447–2456. ACM, New York (2015)

Ishii, H., Lakatos, D., Bonanni, L., Labrune, J.B.: Radical atoms: beyond tangible bits, toward transformable materials. ACM Interact. **19**(1), 38–51 (2012)

Ishii, H., Ullmer, B.: Tangible bits: toward seamless integration of interfaces between people, atoms, and bits. In: Proceedings of CHI 1997, pp. 234–241. ACM Press, New York (1997)

Jenkins, T.: Designing the “Things” of the IoT. In: TEI 2015: Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (2015)
Jung, H., Stolterman, E.: Form and materiality in interaction design: a new approach to HCI. In: Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI 2011), pp. 399–408. ACM, New York (2011)
Karana, E., Giaccardi, E., Stamhuis, N., Goossensen, J.: The tuning of materials: a designer’s journey. In: DIS 2016: Proceedings of the 2016 ACM Conference on Designing Interactive Systems. ACM (2016)
Karat, J.: User-centered software evaluation methodologies. In: Helander, M., Landauer, T.K., Prabhu, P. (eds.) Handbook of Human-Computer Interaction, pp. 698–704. Elsevier Science B.V., New York (1997)
Nielsen, J.: Usability Engineering. Academic Press, London (1993)
Pan, D., Hang-yat Lam, A., Wang, D.: Carrying my environment with me in IoT-enhanced smart buildings. In: MobiSys 2013: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services. ACM Press (2013)
Pignotti, E., Beran, S., Edwards, P.: What does this device do? In: URB-IOT 2014: Proceedings of the First International Conference on IoT in Urban Space. ICST (2014)
Robles, E., Wiberg, M.: Texturing the ‘Material Turn’ in interaction design. In: Proceedings of TEI 2011, Fifth International Conference on Tangible, Embedded, and Embodied Interaction, pp. 137–144. ACM Press, New York (2010)
Romano, B.: Managing the Internet of Things. In: SIGCSE 2017: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education. ACM Press (2017)
Savola, R., Abie, H., Sihvonen, M.: Towards metrics-driven adaptive security management in e-health IoT applications. In: BodyNets 2012: Proceedings of the 7th International Conference on Body Area Networks. ICST (2012)
Sulistyo, S.: Software development methods in the Internet of Things. In: Mustofa, K., Neuhold, E.J., Tjoa, A.M., Weippl, E., You, I. (eds.) ICT-EurAsia 2013. LNCS, vol. 7804, pp. 50–59. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36818-9_6
Tholander, J., Normark, M., Rossitto, C.: Understanding agency in interaction design materials. In: CHI 2012: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2012)
Virzi, R.A.: Usability inspection methods. In: Helander, M., Landauer, T.K., Prabhu, P. (eds.) Handbook of Human - Computer Interaction, pp. 705–715. Elsevier Science B.V. (1997)
Vallgårda, A., Boer, L., Tsaknaki, V., Svanes, D.: Material programming: a new interaction design practice. In: Proceedings of DIS 2016. ACM Press, New York (2016)
Want, R.: The physical web. In: IoT-Sys 2015: Proceedings of the 2015 Workshop on IoT challenges in Mobile and Industrial Systems. ACM Press (2015)
Wiberg, C.: A Measure of Fun. Extending the scope of web usability. Thesis, Department of Informatics. Umeå University (2003)
Wiberg, M.: The Materiality of Interaction - Notes on the Materials of Interaction Design. MIT Press (2018)
Wiberg, M.: Interaction, new materials and computing: beyond the disappearing computer, towards material interactions. Mater. Des. **90**, 1200–1206 (2016)
Wiberg, M., Kaye, J., Thomas, P.: PUC theme Issue: Material Interactions, Personal and Ubiquitous Computing. Springer, UK (2013a)
Wiberg, M., Ishii, H., Rosner, D., Vallgårda, A., Dourish, P., Sundström, P., Kerridge, T., Rolston, M.: Materiality Matters – Experience Materials. ACM Interactions, March/April 2013 (2013b)
Wiberg, M.: Methodology for Materiality: Interaction Design Research Through a Material Lens, Personal and Ubiquitous Computing. Springer, UK (2013)

Wiberg, M., Ishii, H., Rosner, D., Vallgårda, A., Dourish, P., Sundström, P., Kerridge, T., Rolston, M.: Material interactions – from atoms and bits to entangled practices. In: Panel at CHI 2012, in Proceedings of CHI 2012, New York, NY, USA, pp. 1147–1150 (2012)
Wiberg, M., Robles, E.: Computational compositions: aesthetics, materials, and interaction design. Int. J. Des. **4**(2), 65–76 (2010)

Check for updates

Evaluation of Visualization Heuristics

Ryan Williams(✉), Jean Scholtz, Leslie M. Blaha, Lyndsey Franklin, and Zhuanyi Huang

Pacific Northwest National Laboratory, Richland, WA 99354, USA
{ryan.williams,jean.scholtz,leslie.blaha,lyndsey.franklin, zhuanyi.huang}@pnnl.gov
https://www.pnnl.gov/

Abstract. Multiple sets of heuristic have been developed and studied in the Human Computer Interaction (HCI) domain as a method for fast, lightweight evaluations for usability problems. However, none of the heuristics have been adopted by the information visualization or the visual analytics communities. Our literature review looked at heuristic sets developed by Nielsen and Molich [7] and Forsell and Johansson [1] to understand how these heuristics were developed and their intended applications. We also reviewed heuristic studies conducted by Hearst and colleagues [2] and Väätäjä and colleagues [10] to determine how individuals apply heuristics to evaluating visualization systems. While each study noted potential issues with the heuristic descriptions and the evaluator's familiarity with the heuristics, no direct connections were made. Our research looks to understand how individuals with domain expertise in information visualization and visual analytics could use heuristics to discover usability problems and evaluate visualizations. By empirically evaluating visualization heuristics, we can identify the key ways that these heuristics can be used to inform the visual analytics design process. Further, they may help to identify usability problems that are and are not task specific. We hope to use this process to also identify missing heuristics that may apply to designs for different analytic purposes.

Keywords: Heuristics · Visualization · Heuristic evaluation
User study · Visual analytics · Information visualization

1 Introduction

Heuristics have been used since the early 1990's in the Human Computer Interaction (HCI) domain, for evaluating user interfaces. Heuristics, as described by Nielsen, are more rules of thumb rather then specific usability guides [6]. Heuristic evaluations are an approach used to discover usability problems, most commonly in software user interfaces [4]. Heuristics establish a common language around a prescribed definition to focus the evaluation of a user interface. These ten heuristics developed by Nielsen [6] have been used by the HCI community.

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

© This is a U.S. government work and its text is not subject to copyright protection in the United States; however, its text may be subject to foreign copyright protection 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 208–224, 2018.
https://doi.org/10.1007/978-3-319-91238-7_18

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

Heuristic evaluations are an efficient method of discovering usability problems. They can be applied early on in the design process of a software user interface to discover potential usability issues before code is written. Forsell and Johansson [1] describe heuristic evaluations as a discount method that is easy to learn and apply during all phases of development. Alternatives to heuristic evaluations are eliciting qualitative opinions about the usefulness of a particular software user interface. This method is less useful because finding a consensuses on usability problems with unstructured opinions can be difficult.

At this point, neither the information visualization (InfoVis) community nor the visual analytics community has adopted a common set of heuristics to use early on in the design process. However, Forsell and Johansson [1] used the same methodology as [8] and developed a set of 10 heuristics for information visualization. Forsell and Johansson designed and executed a study to determine which heuristics of a set of 63 published heuristics could best explain a collection of 74 usability problems. Their study asked 6 participants with expertise in InfoVis and/or HCI to rate how well each heuristic explained a usability problem. The resulting set of 10 heuristics (Table 1) had the best explanatory coverage out of all possible heuristic combinations [1]. Explanatory coverage, as described by Nielsen [5], is a rating of a heuristic that explains a major part of a usability problem. The resulting set of ten heuristics from Forsell and Johansson cover the broadest set of their 74 usability problems.

While these heuristics have not been commonly used for evaluations in the InfoVis community, several studies have been conducted using these heuristics in an attempt to understand how they could be applied. Hearst and colleagues [2] found that using the heuristics for visualizations along with asking questions about the data resulted in complementary results as the questions about the data could compensate for heuristics that were difficult to understand and apply.

Foresell and Johannson noted that future work needed to be done to investigate any issues in applying their heuristics to find and explain usability problems [1]. Väätäjä and colleagues [10] found that heuristics related to interaction, veracity, and aesthetics needed to be added to the Forsell and Johansson set based on participant feedback captured during their study. Participants noted limitations in the heuristics to account for visualization libraries that are not flexible that could potentially leave out important information. The study also noted that if issues beyond basic usability issues are of interest, training domain experts, with a good understanding of the data and information system being analyzed, to carry out the heuristics evaluation would likely provide insightful feedback [10].

Our study wants to determine more specifically what was difficult about using these heuristics and how domain experience in creating and evaluating visualizations influences the applied use of heuristics for discovering usability problems in visualizations. We conducted a controlled experiment to determine what factors might impact an evaluators' capability to conduct a heuristic evaluation of a static information visualization. In particular, we wanted to understand how their familiarity with the visualizations and their experience level in conducting heuristic evaluations impacted their evaluations of visualizations using the Forsell and Johannnson heuristic set.

2 Study Overview

The goal of our research in heuristic evaluations for visualizations is to determine how a heuristic set can be used in the InfoVis community much in the same way Nielsen's heuristics have been used in the HCI community. Heuristics provide a fundamental benefit of discovering and communicating usability problems early in the development process. With a better understanding of which heuristics are best suited for visualizations and how heuristics can be applied to visualization evaluations, the InfoVis community can move closer to adopting a common set of heuristics to use to evaluate visualizations.

We conducted a controlled experiment to determine what factors might impact an evaluator's capability to conduct a heuristic evaluation of a static information visualization. We asked that each participant have experience in user experience design or research, user interface development and/or HCI research. In particular, we wanted to understand how their familiarity with the visualizations and the participant's experience level in conducting heuristic evaluations impacted their evaluations of visualizations.

We investigated the following three hypotheses about the usefulness of heuristics for finding usability problems:

Hypothesis 1. Participants would find the heuristics that involved interactive features less useful when given only static visualizations.
Hypothesis 2. Participants would find visualizations that they were less familiar with more difficult to evaluate using heuristics.
Hypothesis 3. Participants who were less familiar with heuristic evaluation would have a more difficult time using the heuristics overall.

2.1 Methods

Participants. Ten domain experts in the visualization field participated in the study. Each self-reported experience with one of more of the following: user experience design or research, user interface development, and HCI research.

The age range of the participants are: 4 ages 24–29, 2 ages 30–25, 3 ages 36–41, and 1 age 48–53. Because of the reliance on color visualizations in this study, all participants were screened for normal vision. All participants demonstrated normal (2 participants) or corrected-to-normal (6 glasses, 2 contacts) vision by achieving acuity of at least 20/20 using the Snellen eye chart [9]. Participants successfully completed the Ishihara color plates [3] with 100% accuracy indicating normal color vision. All volunteers consented to participate in accordance with the policies of the Institutional Review Board at the Pacific Northwest National Laboratory. Nine participants completed the evaluation for all visualizations; one participant evaluated only four of the five visualizations due to time constraints.

Visualization Heuristics. Participants were provided with a fixed set of heuristics with which they evaluated five common visualizations. We used the set of ten heuristics proposed by Forsell and Johansson [1], listed in Table 1.

Visualizations. Five common visualizations were selected for the study: Scatter Plot, Sunburst, Tree Map, Parallel Set, and Area Graph. To keep the visualized data a constant factor, all visualizations depicted data from the the VAST 2008 Mini Challenge Two: *Migrant Boats*[1]. The 2008 VAST Mini Challenge 2 comprises records dealing with the mass movement of persons departing a fictitious island to the United States during a two-year period. The resulting visualizations are illustrated in Fig. 1.

Usability Problems Ground Truth. Visualizations were used herein with default settings and layout options. Although we did not "plant" usability issues in the visualization, we did not adjust them to ensure that any usability problems were omitted. Two of the authors, JS and LF, who were quite familiar with conducting heuristic evaluations, performed a "ground truth" evaluation of the different visualizations. In this process, at least one usability problem was identified for each heuristic that the experts deemed relevant to the visualization. We do not claim that all of the usability problems were found in this set, but we felt that it was a good start towards predicting the usability problems that would be identified by our participants. The "ground truth" set of usability problems are given here for each visualization.

Scatter Plot

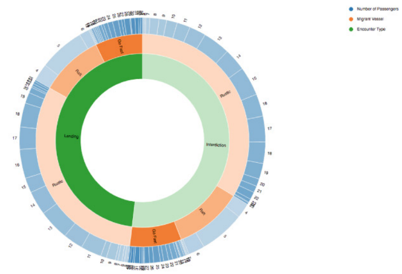
H1 Occlusion: Difficult to see individual points
H2 Date axis compressed and difficult to read
H5 Difficult to compare dot size from different months

[1] See https://www.cs.umd.edu/hcil/VASTchallenge08/ for full Mini-Challenge details and data.

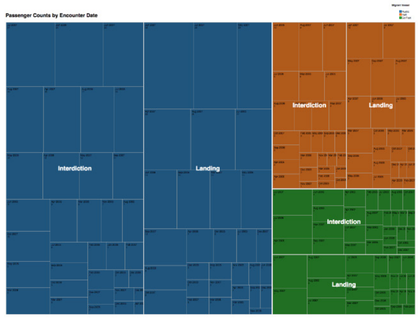
Table 1. Forsell and Johansson (2010) heuristics leveraged in our present study

Number	Name	Description
H1	Spatial organization	Is related to the overall layout of a visual representation, which comprises analyzing how easy is to locate an information element on the display and to be aware of the overall distribution of information elements in the representation
H2	Information coding	Concerns the mapping of data elements to visual elements, as well as use of additional symbols or realistic characteristics that can be used either for building alternative representations (like groups of elements in clustered representations) or to aid in the perception of information elements
H3	Orientation and Help	This refers to functions that provide support for the user to control level of details, redo/undo of user actions and representation of additional information (for example, the path a user follows while navigating in a complex structure)
H4	Data set reduction	Features such as filtering allows reduction of information shown at a certain moment, leading more rapidly to adjustment of the focus of interest, and clustering allows the representation of a subset of data elements by means of special symbols, while pruning simply cuts off irrelevant information for the understanding a visual representation
H5	Recognition rather than recall	Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate
H6	Remove the extraneous	Is a principle that pushes the graphic designer to present the largest amount of data with the least amount of ink. Extra ink can be a distraction and take the eye away from seeing the data or making comparisons
H7	Prompting	Refers to the means available in order to guide the users towards making specific actions whether these are data entry or other tasks. This criterion also refers to all the means that help users to know the alternatives when several actions are possible depending on the contexts. Prompting also concerns status information, that is information about the actual state or context of the system, as well as information concerning help facilities and their accessibility
H8	Minimal actions	These refer to workload with respect to the number of actions necessary to accomplish a goal or a task. It is here a matter of limiting as much as possible the steps users must go through
H9	Consistency	Refers to the way interface design choices (codes, naming, formats, procedures, etc.) are maintained in similar contexts, and are different when applied to different contexts
H10	Flexibility	Is reflected in the number of possible ways of achieving a given goal. It refers to the means available to customization in order to take into account working strategies, habits, and task requirements

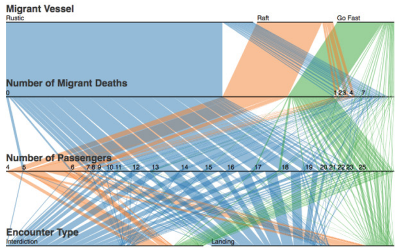
(V1) Scatter Plot. Analytical question: Which vessel type by the highest death per passenger ratio?



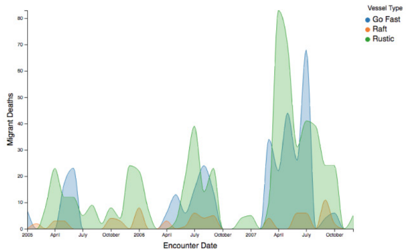
(V2) Sunburst. Analytical question: Which vessel type had the most cumulative passengers?



(V3) Tree Map. Analytical question: Which vessel type had the most cumulative passengers?



(V4) Parallel Sets. Analytical question: Which vessel type by the highest death per passenger ratio?



(V5) Area Plot. Analytical question: Which vessel type had the highest cumulative deaths?

Fig. 1. The five visualizations used in the present study. All depict the VAST 2008 Mini Challenge Two: *Migrant Boats* data.

H5 Difficult to verify that 24 is the largest number of deaths? (There seem to be sizes that do not correspond to the 0, 1, 12, 24 in the legend)
H2, H8 Scales do not help answer the question directly (which vessel type has the highest death per passenger ratio), requires mental math

Sunburst

H1 Vessel types are not contiguous which makes it difficult to count passengers (add from two parts in the display, both landing + interdiction)
H1 Numbers occlude, are difficult to read in Go Fast vessel type
H9 Blue is inconsistent within a vessel type and the variation does not mean anything (is it that the number of passengers $\sim$ opacity?)

Tree Map

H1 Labels occlude each other a little bit
H3 Passenger count not labeled
H2 Shape of vessel of same passenger count is not consistent (comparing squares and rectangles)
H8 Have to add to get total passengers

Parallel Sets

H3] Number of deaths impossible to read (only labels 0, 1, 2, 3, 4, and 7, have to guess at the others)
H3 Number of passengers impossible to read (there are not enough labels at far right end)
H1 Connecting Death to Passengers requires following a thin line
H8 Still need mental math to answer ratio question
H1 Line crossings difficult to follow

Area Plot

H8, H6 Requires mental math to get cumulative deaths (sum area under curve)
H6 Labels are not precise enough for exact math
H2 Continuous lines and area makes it difficult to know how to do the addition (i.e., flat portion of raft for April 2005, how many 4 s do we add?)
H2 Viz better suited for relative comparison than precise reading

Procedure. Following informed consent and vision screening, participants were asked to complete a demographic survey in which they self-rated the following: (1) their experience using each visualization type; (2) their experience conducting heuristic evaluations for visualization; (3) the frequency at which they use each visualization type to extract information, and (4) their experience developing software to create each visualization type. Ratings were reported on a 6-point

Likert scale with a rating of 1 being very infrequently/inexperienced and 6 being very frequently/experienced.

The experience level of participants conducting heuristic evaluations were: 2 participants reported experienced, 4 participants reported somewhat experienced, and 4 participants reported that they were inexperienced as recorded in Table 2.

Table 2. Participant visualization experience levels

Visualization	Unfamiliar	Novice	Moderate	Expert	Advanced
Tree map	-	2	4	-	4
Parallel sets	3	3	3	-	1
Gantt chart	1	3	3	1	2
Sankey diagram	4	3	1	-	2
Donut chart	2	3	2	-	3
Stack bar	-	1	4	2	3
Heat map	-	1	3	1	5
Scatter plot	-	1	3	1	5
Sunburst	3	3	2	-	2
Time series	2	-	2	1	5

This assessment was not used to down select the from the 10 visualizations to the 5 used in our study but merely to understand how participants rated their experience using the visualizations listed. We inadvertently did not collect this information about the area graph though we felt that participants were most likely reasonably familiar with this one as well.

Participants were given a suggested analytical question related to each visualization to keep in mind as they applied the 10 heuristics. The analytical question for each visualization is given in the relevant figure captions in Fig. 1.

Participants were shown the visualizations in a random order. For each visualization, participants were asked to apply each heuristic, in the order listed in Table 1. For each heuristic, within a given visualization, participants were asked the following:

- Using this heuristic, do you see the potential for a usability problem with the visualization? [yes/no]
- Please describe the usability problem that you found. [free response]
- How well does this heuristic capture the usability problem that you identified? [6-point Likert scale]
- How relevant is this heuristic for evaluating this visualization? [6-point Likert scale]
- How confident do you feel in your ability to apply this heuristic to evaluate this visualization? [6-point Likert scale]

Following the visualization evaluations, participants provided qualitative feedback on the usability issues of each visualization. They rated the clarity of each heuristic description on a 6-point Likert scale, and were asked to provide (free response) feedback on which parts of the each description were easy to understand and use and which were difficult to understand and use.

3 Results

3.1 Usefulness of Heuristics

Hypothesis 1 posited that participants would find the heuristics that involved interactive features less useful when given only static visualizations. All the visualizations in this study were static with no interactions possible in the interfaces provided. The heuristics that refer to interactions are H3, H7, H8, H10; the heuristics that do not refer to interactions are H1, H2, H4, H5, H6, and H9. We operationalize useful in this context as relevance and confidence. Thus, we predict that, across all visualizations, for the set of heuristics referring to interactions, relevance and confidence ratings will be lower than for the heuristics that do not refer to interactions.

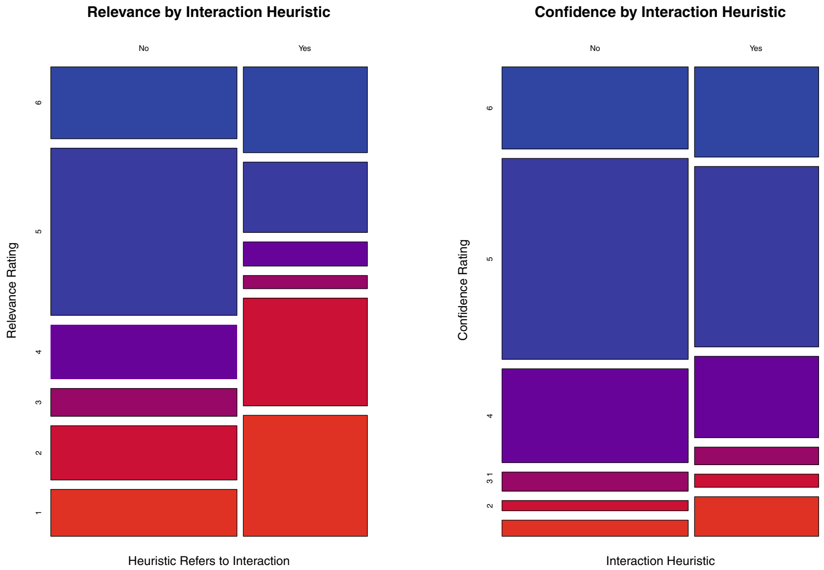


Fig. 2. Relevance (left) and confidence (right) ratings for the heuristics referring to interaction (Yes column) versus the those for heuristics no referring to interaction. These plots depict the data associated with evaluation hypothesis 1.

We compared relevance ratings between the two groups with a linear mixed ordinal regression model with a logit link function, using the fixed factor interaction heuristic (2 levels: yes, no) and the random factor observer. The Type I error rate was $\alpha = .05$. The relevance ratings for heuristics referring to interactions were significantly lower than the relevance ratings for heuristics not referring to interactions (Parameter $\beta_{\text{Yes}} = -1.10$, $z = -5.95$, $p < .001$). The relevance ratings are shown in Fig. 2.

Confidence ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor interaction heuristic (2 levels: yes, no) and the random factor observer. The Type I error rate was $\alpha = .05$. The confidence ratings for heuristics referring to interactions were significantly lower than the confidence ratings for heuristics not referring to interactions (Parameter $\beta_{\text{Yes}} = -0.2$, $z = -123.4$, $p < .001$). The confidence ratings are shown in Fig. 2.

Hypothesis 2 posited that participants would find visualizations with which they were less familiar more difficult to evaluate using heuristics. According to their self-reported ratings, participants were less familiar with the Sunburst (V2) and Parallel Set (V4) visualizations. They were more familiar with the Scatter Plot (V1), Tree Map (V3), and Area Plot (V5) visualizations as recorded in Table 2. We operationalized difficult to be confidence ratings in using the heuristics. The prediction is that participants will report lower confidence across all heuristics between more and less familiar visualizations.

Confidence ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor familiarity with visualizations (2 levels: low, high) and the random factor observer. The Type I error rate was $\alpha = .05$. There was no difference in confidence ratings between familiarity groups (Parameter $\beta_{\text{Low}} = -0.08$, $z = -0.51$, $p = .61$).

Hypothesis 3 posited that participants who were less familiar with heuristic evaluation would have a more difficult time using the heuristics overall. We split participants into two groups, with low familiarity with heuristic evaluation defined as those participants who self-reported their experience with heuristic evaluations as level 1–3, and the high familiarity participants defined as those who self-reported their experience as level 4–6. We operationalized difficult to mean confidence in using the heuristics. We predicted that more difficulty would be lower confidence in the participants reporting less experience with heuristic evaluations (Fig. 3).

Confidence ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor familiarity with heuristic evaluations (2 levels: low, high) and the random factor observer. The Type I error rate was $\alpha = .05$. There confidence ratings were significantly lower for participants with less experience with heuristic evaluations (Parameter $\beta_{\text{Low}} = -1.3973$, $z = -7.09$, $p < .001$).

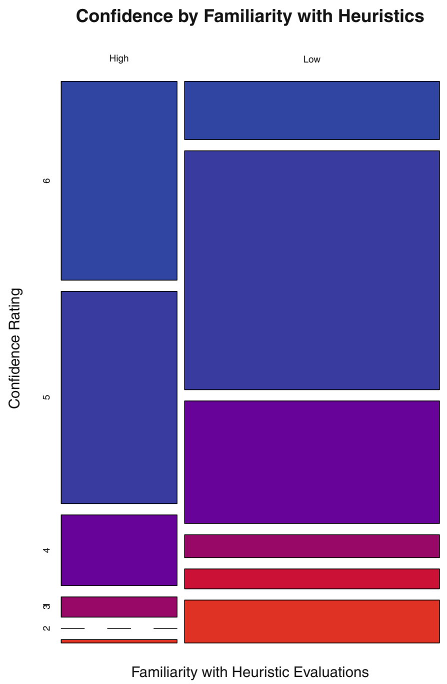


Fig. 3. Confidence ratings for the participants with low familiarity with heuristic evaluation (Low column) versus the those with high familiarity with heuristic evaluation (High column). Plot depicts data associated with evaluation hypothesis 3.

3.2 Usability Problems Reported by Participants

Table 3 shows the ground truth usability problems that were reported by our participants, regardless of the heuristic participants attributed them to. In a number of instances, the description of the problem provided by a participant was vague, and at times they described several things that would fit under different heuristics. At times the usability problem they found did not fit with the heuristic to which they attributed it. Several participants reported "failed" if the heuristic required an interaction to determine if there was a usability issue. Therefore, we created confusion matrices for each of the five visualizations to tease out the usability problems that were actually attributable to the specific heuristics. Tables 4, 5, 6, 7 and 8 contain the confusion matrices for each of the five visualizations. The first column notes the expert ground truth, with a star denoting if the experts found a usability problem related to that particular heuristic. The subsequent columns show whether each participant found a usability problem using the heuristics listed there.

Table 3. Ground truth usability issues described by participants

Visualization	Issue	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Scatter plot	Occlusion (H1)	X	X	-	-	X	X	X	X	X	-
	Compressed axis (H2)	-	X	-	-	-	X	X	-	-	-
	Dot compare (H5)	X	-	-	-	X	-	X	-	-	-
	Dot size (H5)	-	X	-	-	X	X	X	X	-	-
	Indirect scales (H8, H2)	-	-	-	-	X	-	-	-	-	-
Sunburst	Non-contiguous vessel types (H1)	X	-	X	X	X	X	-	-	-	X
	Occlusion (H1)	-	X	-	-	-	X	X	X	-	X
	Inconsistent color (H9)	-	-	-	X	X	X	X	X	-	X
Tree map	Occlusion (H1)	X	-	X	X	-	-	-	-	-	X
	Missing labels (H3)	-	-	-	-	-	-	-	X	-	-
	Mental addition (H8)	-	X	-	X	X	-	X	X	X	X
	Inconsistent shapes (H9)	-	-	-	X	-	X	X	X	-	-
Parallel	Labels unreadable (H3)	-	-	X	X	X	X	X	X	X	-
	Line crossings (H1)	X	X	X	X	X	X	-	X	X	-
	Line following (H1)	X	X	X	X	X	-	-	X	-	-
	Missing labels (H3)	-	-	X	-	X	X	X	-	X	-
	Mental math (H8)	-	X	-	X	X	-	-	-	-	-
Area	Continuous lines vs addition (H2)	-	-	X	X	X	-	X	-	-	X
	Vis-task mismatch (H2)	-	-	X	X	X	-	-	-	-	X
	Labels not precise (H6)	-	X	-	-	-	-	X	-	-	X
	Imprecise labels (H6)	-	X	-	-	-	-	X	-	-	X
	Mental math (H8)	-	-	-	-	X	-	-	-	-	-

In Table 4, we can see that 4 participants found a usability problem that was comparable to the one found in our expert evaluation for heuristic 1. Only one participant found usability problems attributed to heuristic 2. No participants found usability problems attributable to heuristics 3 and 4. Six participants found 1 or 2 usability problems attributable to heuristic 5. One participant found a usability problem using heuristic 8. There were a total of 7 usability problems described that did not fit in any of the heuristics but were reported by participants (total of the row labeled “other”). If you compare the results in Tables 3 and 4 you see that in Table 3, in the Scatter plot section, participants 5, 7, 8 and 9 reported a usability problem attributable to heuristic 1. However, the description they provided did not match with that heuristic. In the confusion matrix, such a description would be attributed to another heuristic or “other”. Looking at the various confusion matrices, it seems that more participants were able to find usability problems that were explained by heuristics 1 and 2.

We also have a number of problems found by participants classified as “other”. These include participant observations about the lack of interaction or controls for the visualization (“There are no user controls to filter, reduce, or

Table 4. Confusion matrix for the Scatter plot visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	5	1	-	1	-	2	-	-	-	-
EH2*	-	2	-	-	-	-	-	-	-	-
EH3	-	-	-	-	-	-	-	-	-	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5*	1	2	2	-	1	1	-	-	1	-
EH6	-	-	-	-	-	-	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	-	1	-	-	-	-	-	-	-	-
EH9	-	-	-	-	-	-	-	-	-	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	-	1	1	-	1	1	-	2	1
Failed	-	-	1	-	-	-	1	1	-	1
No response	4	5	6	8	9	6	8	9	7	8

Table 5. Confusion matrix for the Sunburst visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	6	2	-	-	2	3	-	-	-	-
EH2	1	-	-	-	-	-	-	-	-	-
EH3	2	-	1	-	1	-	-	-	-	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5	-	-	-	-	-	-	-	-	-	-
EH6	-	-	-	-	-	-	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8	-	-	-	-	-	-	-	-	-	-
EH9*	-	3	-	-	-	1	-	-	5	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	-	3	2	-	1	2	2	-	2
Failed	-	-	-	1	-	-	1	1	-	1
No response	3	5	6	7	7	5	7	7	5	7

re-order the data."). Vague color critiques ("Lots of color and text.") that did not describe a specific problem also fell into this category. It may be that additional heuristics are needed or if current heuristic descriptions need to be improved. Color in particular may be deserving of its own heuristic, given the number of time participants described it across all visualization and heuristics.

Table 6. Confusion matrix for the Tree map visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	1	1	-	-	-	1	-	-	-	-
EH2	-	-	-	-	-	-	-	-	-	-
EH3*	-	-	2	-	2	4	-	-	1	-
EH4	-	-	-	1	-	-	-	-	-	-
EH5	1	2	2	-	1	1	-	-	1	-
EH6	-	-	-	-	-	-	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	6	1	-	1	1	-	-	1	-	-
EH9*	3	1	-	-	-	-	-	-	2	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	3	2	3	2	1	3	1	2	-
Failed	-	-	-	-	-	-	1	1	-	-
No response	1	4	6	5	5	4	4	7	5	7

Table 7. Confusion matrix for the parallel coordinates visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	5	1	1	1	1	3	-	-	-	-
EH2	-	-	-	-	-	-	-	-	-	-
EH3*	1	1	2	-	-	-	-	-	1	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5	-	-	-	-	-	-	-	-	-	-
EH6	-	-	-	-	-	1	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	-	1	-	-	-	-	-	-	-	-
EH9	-	-	-	-	-	-	-	-	-	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	1	3	2	2	1	2	1	-	-	2
Failed	-	-	-	-	1	-	1	1	-	-
No response	3	4	5	7	7	4	8	9	9	8

The experts who built the ground truth set of usability problems also helped to code participant responses. In a number of cases, participants described a usability problem using what the experts felt was the appropriate heuristic. However, the problem was not a usability problem that was called out in the ground truth set of usability problems. This suggests that there may be some subjectivity with regards to what is considered a "problem" and what is not, and

Table 8. Confusion matrix for the area visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1	-	-	-	-	-	-	-	-	-	-
EH2*	3	1	-	1	1	2	1	1	2	-
EH3	1	-	-	-	-	-	-	-	1	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5	-	-	-	-	-	-	-	-	-	-
EH6*	1	-	-	-	-	1	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	-	-	-	-	-	-	-	-	-	-
EH9	-	-	-	-	-	-	-	-	1	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	4	4	1	2	1	1	-	-	1
Failed	-	-	-	-	-	-	1	1	-	1
No response	5	4	6	8	7	6	7	8	6	8

just how much of a problem something must be before a visualization begins to suffer. Minor issues to one person may in fact be usability breakdowns to another.

Looking across Tables 4, 5, 6, 7 and 8, it is clear that within each visualization, all participants reported similar numbers of usability problems. This illustrates that participants found similar numbers of issues, regardless of their confidence or ratings of relevance of the heuristics. But given that a number of the issues were found in the "other" category, we examined the ratings participants gave for the clarity of the heuristics themselves. We note that participants 1–3 did not complete the clarity ratings, so the analysis was conducted on the remaining seven participants. We split participants into two groups, with low familiarity with heuristic evaluation defined as those participants who self-reported their experience with heuristic evaluations as level 1–3, and the high familiarity participants defined as those who self-reported their experience as level 4–6. This is because we found that familiarity with heuristic evaluations influenced ratings of relevance.

Clarity ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor familiarity with heuristic evaluations (2 levels: low, high) and the random factor observer. The Type I error rate was $\alpha = .05$. There was not a significant difference in clarity ratings between the participant groups (Parameter $\beta_{\text{Low}} = -1.01$, $z = -1.74$, $p = .08$).

4 Conclusion

Our study shows that having familiarly with heuristic evaluations increases the confidence in the ability to apply heuristics for evaluating visualizations. This suggests that those who are asked to conduct a heuristic evaluation should first have some experience or training in using heuristics to evaluate visualizations. A training tool could be developed to help those novice to heuristic evaluations learn best practices in heuristic application for visualization evaluations. Or the evaluators who will be conducting the heuristic reviews could use the visualizations in this paper to see what usability issues they find and compare them with the ground truth. Another idea would be to hold some workshops at various information visualization and visual analytics conferences to introduce the technique of heuristic evaluation to the communities.

Our data suggests that conducting heuristic evaluations for visualizations does not require any experience using the visualization itself. This is actually a good outcome as heuristic evaluations in information visualization and visual analytics will certainly involve novel visualizations.

While the heuristic clarity ratings did not differ significantly based on the participant's experience with heuristic evaluations, we did find that participants had difficulty attributing a usability problem to a specific heuristic. In particular, participants had difficulty attributing usability issues relating to color to a heuristic. This could be addressed by creating a heuristic that specifically addresses color usability issues. Or through training that demonstrates how evaluators can operationalize usability problems so that developers can understand how to fix the usability issue. Future studies could help determine specifically what evaluators found difficult with attributing a usability problem to a heuristic by asking the participants to describe the found usability problems aloud.

We definitely recommend that visualizations on which a heuristic evaluation is going to be conducted include the necessary information so that interactions can be properly evaluated. To simulate the interaction of a visualization, interactive interface elements such as buttons, icons or menus could be displayed. Another static image could show the result of the different interactions allowing evaluators to assess heuristics concerned with interactions and processes.

Our study has shown that while some further investigations into heuristics for visualizations and visual analytics should definitely be done at this time, using these heuristics for information visualization and visual analytics should definitely help researchers and developers to produce early visualizations with fewer usability issues. We encourage the communities to use the heuristic review technique using this set of heuristics to see if their visualizations are improved.

Acknowledgments. The research described in this document was sponsored by the U.S. Department of Energy through the Pacific Northwest National Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Forsell, C., Johansson, J.: An heuristic set for evaluation in information visualization. In: Proceedings of the International Conference on Advanced Visual Interfaces, pp. 199–206. ACM (2010)
2. Hearst, M.A., Laskowski, P., Silva, L.: Evaluating information visualization via the interplay of heuristic evaluation and question-based scoring. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5028–5033. ACM (2016)
3. Ishihara, S.: Tests for Colour-Blindness. Hongo Harukicho, Handaya (1917)
4. Nielsen, J.: Finding usability problems through heuristic evaluation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 373–380. ACM (1992)
5. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 152–158. ACM (1994)
6. Nielsen, J.: 10 usability heuristics for user interface design. https://www.nngroup.com/articles/ten-usability-heuristics/
7. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 249–256. ACM (1990)
8. Plaisant, C., Grinstein, G., Scholtz, J., Whiting, M., O'Connell, T., Laskowski, S., Chien, L., Tat, A., Wright, W., Görg, C., et al.: Evaluating visual analytics at the 2007 vast symposium contest. IEEE Comput. Graph. Appl. **28**(2) (2008)
9. Snellen, H.: Test-types for the Determination of the Acuteness of Vision. PW van de Weijer, Utrecht (1862)
10. Väätäjä, H., Varsaluoma, J., Heimonen, T., Tiitinen, K., Hakulinen, J., Turunen, M., Nieminen, H., Ihantola, P.: Information visualization heuristics in practical expert evaluation. In: Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization, pp. 36–43. ACM (2016)

The Temporal-Spatial Metaphorical Expression Difference in Spatial Schema Design

Weiqi Zheng[1,2], Ye Liu[1,2](✉), and Xiaolan Fu[1,2]

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

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. The temporal-spatial metaphor is a conceptual metaphor which can help construct the concept of time through spatial experience. The motion direction in space can be used to represent the motion direction of time. When comprehending the relative motion between time and people, two kinds of time movement perspectives are often used in the frame of temporal-spatial metaphor, namely the ego-moving perspective and the time-moving perspective. The present study focused on revealing the kinds of spatial schemas that could represent time movement perspectives and explored the feasibility of these spatial schemas under different situations, which will be helpful in depicting the concept of time in a human-computer interaction environment. We conducted two experiments to explore to what extent vertical and horizontal spatial schemas could represent the two perspectives of time movement. The results showed that the foreshortening vertical spatial schema could be adopted to represent both the ego-moving and the time-moving perspectives, and the horizontal spatial schema is more appropriate to describe the ego-moving perspective. The findings have implications for computer spatial schema design in time representation.

Keywords: Spatial schema · Temporal-spatial metaphor · Ego-moving Time-moving

1 Introduction

Metaphors are widely used in our daily life as a rhetorical device. Psycholinguistics suggest that metaphors are also a way of thinking to allow people to draw on concrete and familiar knowledge to reason about abstract concepts [1]. And even if metaphors are not used in sentences apparently, people could still rely on metaphors to help thinking [2]. In the domain of computer design, metaphors exhibit these properties in interfaces, they are used to produce figurative interfaces which involve a transfer of meaning from the vehicle to the topic [3]. And metaphors occur throughout the interfaces we use and design just as they invisibly permeate in our daily speech. For example, when a child learns to move a document icon from one folder to another in computer, the object and container we use in real life is actually a metaphor that will help the child to comprehend

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 225–234, 2018.
https://doi.org/10.1007/978-3-319-91238-7_19

the function of folder in computer system. When working with interface metaphors, it's essential for designers to think about which metaphors are proper and can be friendly understood by users in various operating environment.

Among all the abstract topics, time is one of the primary concepts to represent. Time is invisible and intangible, yet humans are capable of feeling time in certain ways. The temporal-spatial metaphor is a conceptual metaphor which can help construct the concept of time through spatial experience [4]. In addition, there is a broad consistency and also some differences of temporal-spatial metaphor expressions in various language and cultures [5, 6]. Therefore, it's important to investigate the figurative interfaces of temporal-spatial metaphor in human-computer interaction design, which used in different conditions and cultures. Usually, the motion direction in space can be used to represent the motion direction of time. And there are three basic spatial axes in temporal-spatial metaphor expressions: Front-back axis, up-down axis and left-right axis. The front-back axis and up-down axis are commonly used to represent the order of time, for example, Mandarin speakers use the spatial words "qian" ("front") and "hou" ("back"), as well as "shang" ("up") and "xia" ("down") to express the occurrence sequence [7]. Although left-right axis is not often used in speech to represent the temporal order, it can be found in some timing device like calendars and workflow charts [8].

When comprehending the relative motion between time and people in temporal-spatial metaphor, two kinds of time movement perspectives are often used in the front-back axis, namely the ego-moving perspective and the time-moving perspective [1, 4]. From the ego-moving perspective, time is considered static while people are in motion, and people are facing and approaching the future as in "We are coming up on Christmas". From the time-moving perspective, people are considered static while time is in motion, and the future is approaching or upon us as in "Christmas is coming up". Previous psycholinguistic studies have verified the psychological reality of the two perspectives of temporal-spatial metaphors [9, 10]. However, there is no systematic studies concerning the spatial schema which can represent the temporal-spatial metaphors so far. In two-dimensional space, designers always use perspective plan method to depict the distance relationship between objects. Thus, representing ego and time moving perspectives in front-back axis can rely on up-down vertical schema and left-right horizontal schema.

The present study focused on revealing the kinds of spatial schemas that could represent time movement perspectives and explored the feasibility of these spatial schemas under different situations, which will be helpful in depicting the concept of time in a human-computer interaction environment. We conducted two experiments to explore to what extent vertical and horizontal spatial schemas could represent the two perspectives of time movement. We designed a question as the reference standard to examine the validity of those schemas. The answer for the standard question comprised of a visualized picture and a descriptive sentence which could directly express the meaning of ego-moving or time-moving perspective, which was widely examined to be valid to measure perspectives of time movement in previous studies [11–13]. The visualized picture was referenced from the schematic of ego/time moving schema used to organize events in time in Boroditsky [11] 's study. The descriptive sentences are matched with different time movement perspectives. Under the ego-moving perspective schema, the sentence is "I'm approaching the future", and under the time-moving perspective schema, the sentence is "The future is approaching me".

In experiment 1, we tested the validity of vertical spatial schema. To examine the foreshortening effect of vertical schema, we compared the foreshortening schema with a schema which contained same-size objects aligned in the up-down axis. Our assumption for experiment 1 was that the foreshortening schema would be more consistent with the standard question than the same-size vertical schema. In experiment 2, we tested the validity of horizontal spatial schema. People have no obvious preference when perceive the moving direction of elements in vertical spatial schema. However, the moving direction of elements in horizontal spatial schema was affected much more by cultural customs, such as our writing and reading direction, because most of the modern people write and read from left to right. Previous studies suggested that emotion could influence the choice of time movement perspectives. Emotion with approach motivation, such as happiness and anger, might promote the choice of the ego-moving perspective, while emotion with avoidance motivation, such as anxiety and sadness, might promote the choice of the time-moving perspective [12–16]. Therefore, we tried to evoke participants' mood with one certain type of emotion of happiness, anger, anxiety or sadness. And the consistency and difference of answers between the horizontal spatial picture and the standard question would be examined. Our assumption was that in approach-motivation emotion groups (happiness and anger), the spatial picture was more consistent with the standard question than in avoidance-motivation emotion groups (anxiety and sadness).

2 Experiment 1

2.1 Participants

Participants consisted of a total of 30 undergraduate and graduate students (14 males and 16 females), ranging from 19 to 26 years of age (M = 22.7 years, SD = 1.9 years). All participants were native Chinese speakers and unaware of the purpose of the experiment.

2.2 Stimuli and Procedure

Three spatial pictures were designed as the testing stimuli (See Fig. 1): (1) Three foreshortening flying saucers aligned vertically (Fig. 1A; The flying saucer in the bottom was the largest and the one in the topside was the smallest. The picture is the ambiguous spatial target used in Boroditsky [11]'s study.); (2) Five foreshortening balls aligned vertically (Fig. 1B; The ball in the bottom was the largest and the one in the topside was the smallest.); (3) Five rectangles aligned vertically in the same size representing five soldiers in line (Fig. 1C). These three spatial pictures and the corresponding descriptions constituted the spatial schema questions.

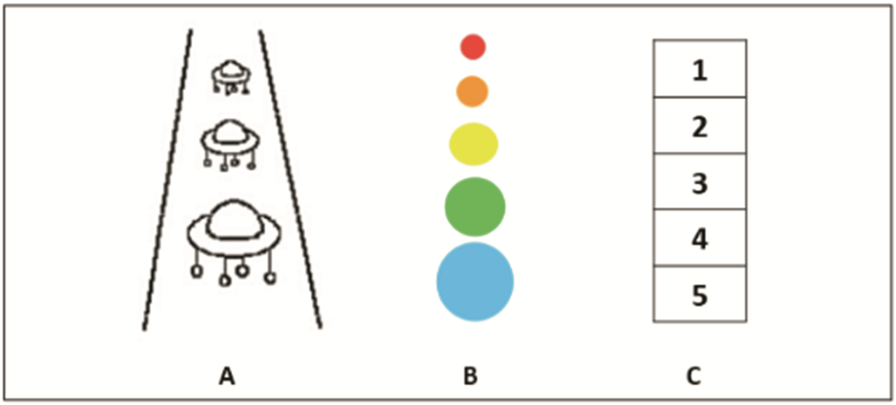


Fig. 1. Three vertical spatial pictures in Experiment 1 (*Note. Picture A is the ambiguous spatial target used in Boroditsky* [11]*'s study)*

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 were asked to answer all the questions (Three spatial schema questions and the standard question). To avoid the order effect, the spatial schema questions and the standard question were presented randomly. For the standard question, the specific question which appeared above the schema picture was "I will attend an activity held by my school recently. Which picture can express my feeling now more properly?" The answer options were under the question. Participants were required to choose one answer from the two pictures. Answer A represents the ego-moving perspective and answer B represents the time-moving perspective.

For the vertical spatial schema A (Foreshortening flying saucers), the specific question which appeared above the picture was "There are three flying saucers arranged in a line, the biggest one is the nearest one to us, if you move one step of the middle flying saucer, which place the middle one will be in?". If participants choose the nearest place, it means they choose the time-moving perspective, if they choose the farthest place, it means they choose the ego-moving perspective. For the vertical spatial schema B (Foreshortening balls), the specific question was "There are five balls arranged in a line, the biggest one is the nearest one to us, if you move two steps of the middle ball, which place the middle one will be in?". If participants choose the blue ball's place, it means they choose the time-moving perspective, and the red ball's place represents the ego-moving perspective. For the vertical spatial schema C (Same-sized rectangles), the specific question was "There are five soldiers arranged in a line, number five is the nearest to us, if number three moves two steps, which place he will be in?" If participants choose number five, it means they choose the time-moving perspective, and number one's place represents the ego-moving perspective.

2.3 Results

All the data were analyzed using SPSS 18.0. Firstly, for each participant, we calculated whether the answers were consistent between every vertical spatial schema question and the standard question (If both of the answers represent the same perspective, they are regarded as consistent and labelled as 1, if not, they are labelled as 0). Then we summed the number of participants for each question that labelled 1, and created a 2×2 contingency table between each vertical spatial schema and the standard question for ego-moving and time moving perspectives(All the contingency tables are concluded in Table 1). And we used the Kappa consistency test and the paired Chi-square test (McNemar's test) to examine the consistency and difference of answers between the vertical spatial schema questions and the standard question.

Table 1. The participant numbers for consistent answers between vertical spatial schema and the standard question

		Schema A		Schema B		Schema C	
		Ego	Time	Ego	Time	Ego	Time
Standard question	Ego	13	6	10	2	7	12
	Time	4	7	9	9	3	8

Note. Ego represents the ego-moving perspective, time represents the time-moving perspective.

The results are as follows: (1) For the consistency and difference between Schema A (Foreshortening flying saucers) and the standard question, the Kappa value is 0.309, and the p value of McNemar's test is $0.754 > 0.05$; (2) For the consistency and difference between Schema B (Foreshortening balls) and the standard question, the Kappa value is 0.304, and the p value of McNemar's test is $0.065 > 0.05$; (3) For the consistency and difference between Schema C (Same-sized rectangles) and the standard question, the Kappa value is 0.082, and the p value of McNemar's test is $0.035 < 0.05$.

2.4 Discussion

Results of experiment 1 showed that the consistency between foreshortening vertical schema and the standard question was higher than the same-sized vertical schema. Accordingly, the difference between foreshortening vertical schema and the standard question was not significant, in contrast, the difference between same-sized vertical schema and the standard question was significant. Thus it indicated that the foreshortening design was crucial to characterize the time movement perspective in temporal-spatial metaphor expression.

3 Experiment 2

3.1 Participants

Participants consisted of a total of 136 undergraduate and graduate students (48 males and 88 females), ranging from 19 to 27 years of age (M = 22.6 years, SD = 1.7 years). All participants were native Chinese speakers and unaware of the purpose of the experiment. They were different students from the participants in Experiment 1. They were randomly assigned to four emotion groups (happy, angry, anxious and sad), with 34 participants in each group.

3.2 Stimuli and Procedure

The standard question mentioned above and a horizontal spatial picture (Fig. 2; five same-size matchstick men lined in the horizontal direction. The picture is referenced from Richmond [17] 's study, the Figure b in spatial group in Appendix.) were used as the two questions to distinguish participants' preferences of time movement perspectives. The specific question which appeared above the schema picture was "There are five matchstick men lined in the horizontal direction, if number three was moved two steps, which matchstick's place the middle one will be in?".

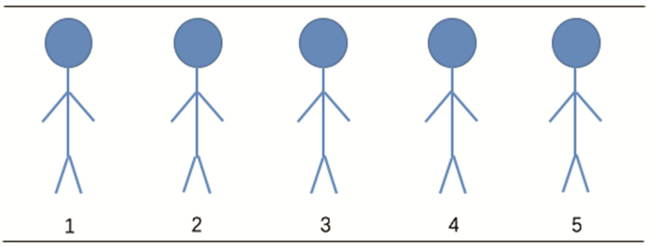


Fig. 2. The horizontal spatial picture in Experiment 2

All the participants were recruited to finish the experiment in a paper-and-pencil questionnaire in a lab. After completing necessary demographic information, participants rated their current emotional states of happiness, anger, anxiousness and sadness on four nine-point Likert scales ranging from 0 (not at all) to 8 (to the extreme). They were then asked to recall and write down their own experience with one certain type of emotion of happiness, anger, anxiety or sadness. They were encouraged to write down their experience in detail within 20 min. Next, participants need to rate their feelings again on the four scales as before. Finally, all the participants were required to answer the standard question and the horizontal spatial schema question.

3.3 Results

All the data were analyzed using SPSS 18.0. Firstly, for each participant, we calculated whether the answers were consistent between the horizontal spatial schema question and the standard question in every emotion group (If both of the answers represent the same perspective, they are regarded as consistent and labelled as 1, if not, they are labelled as 0). Then we summed the number of participants that labelled 1 and created a 2 × 2 contingency table between the horizontal spatial schema and the standard question for ego-moving and time moving perspectives in each emotion group (All the contingency tables are concluded in Table 2). And we used the Kappa consistency test and the paired Chi-square test (McNemar's test) to examine the consistency and difference of answers between the horizontal spatial schema questions and the standard question.

Table 2. The participant numbers for consistent answers between the horizontal spatial schema and the standard question

		Happy		Angry		Anxiety		Sad	
		Ego	Time	Ego	Time	Ego	Time	Ego	Time
Standard question	Ego	19	2	13	5	9	7	8	3
	Time	7	6	9	7	16	2	17	6

Note. Ego represents the ego-moving perspective, time represents the time-moving perspective.

The results are as follows: (1) In the happy group, for the consistency and difference between the horizontal spatial schema and the standard question, the Kappa value is 0.395, and the p value of McNemar's test is $0.18 > 0.05$; (2) In the angry group, for the consistency and difference between the horizontal spatial schema and the standard question, the Kappa value is 0.162, and the p value of McNemar's test is $0.424 > 0.05$; (3) In the anxious group, for the consistency and difference between the horizontal schema and the standard question, the Kappa value is −0.316, and the p value of McNemar's test is $0.093 > 0.05$; (4) In the sad group, for the consistency and the difference between the horizontal spatial schema and the standard question, the Kappa value is -0.009, and the p value of McNemar's test is $0.003 < 0.01$.

3.4 Discussion

Results of experiment 2 showed that the consistency between the horizontal schema and the standard question was different under different emotional states. In approach-motivation emotion groups, the horizontal spatial picture was consistent with the standard question in medium degree, while in avoidance-motivation emotion groups, there was no significant consistency between the horizontal spatial picture and the standard question. Previous studies implied that approach-motivation emotion would promote the choice of ego-moving perspective [12–15]. Therefore, the results indicated that the horizontal spatial schema could represent the ego-moving perspective better than the time-moving perspective.

4 General Discussion

Results from the two experiments demonstrated that the consistency between foreshortening vertical spatial schema and the standard question was much higher than the non-foreshortening schema. The horizontal spatial schema is more appropriate to describe the ego-moving perspective. The findings have implications for computer spatial schema design in time representation.

4.1 The Vertical Spatial Schema

As the findings turn out, the foreshortening design in the vertical direction is essential to express the meaning of ego-moving or time-moving perspective. Because the ego or time moving perspective reflects the relative motion between time and people in essence. It's difficult to portray motion in the front-back axis from the human-centered angle in two-dimension space. The foreshortening design takes advantage of the perspective plan method and can help comprehending the front-back relationship in two-dimension space. And except for foreshortening objects aligned in up-down axis as the stimuli showed, foreshortening road schema might have similar effect to represent ego or time moving perspective.

4.2 The Horizontal Spatial Schema

In experiment 2, we used another idea to test the validity of the horizontal spatial schema which contains the manipulation of different emotion. No obvious foreshortening effect can be created to portray three-dimension scenes in left-right axis, so we don't need to investigate this. The motion direction is also very important to distinguish the time movement perspective. Previous studies provided the idea that emotion plays a role in choosing time movement perspective [12–15]. Specifically, emotion with approach-motivation emotion promote the adoption of ego-moving perspective, while avoidance-motivation emotion promote the adoption of time-moving perspective. In the current study, only in approach-motivation emotion group, there is significant consistency between the horizontal spatial schema and the standard question. The approach motivation direction in left-right axis is the same as in ego-moving perspective. The findings indicated that the horizontal spatial schema is much more appropriate in ego-moving perspective representation.

4.3 The Limitation of the Study

Metaphors are ubiquity in daily speech and interface design. Metaphors serve as natural models and allow us to take our knowledge of familiar objects and events to give structure to abstract, less well understood concepts [18]. The choices of different interface metaphors can provide users with different models of the system. If the interface metaphor is not appropriate, it might cause some problems. Thus coming up with a good metaphor is essential to an easy-to-use human interface.

Our study provided two kinds of spatial schema to represent the time movement perspective, but we did not compare the effect of them directly. In future study, we will focus on the feasibility of vertical and horizontal spatial schema in real application.

In sum, the present study provided the first behavioral evidence about the representation of temporal-spatial metaphors from different spatial schemas. This result gives the valuable guidance of time representation in human-computer interaction.

Acknowledgements. This research was supported by the grant of the National Natural Science Foundation of China (Grant No. 61632004, 61773379, and 61375009) and the grant of the German Research Foundation (DFG) and the National Natural Science Foundation of China in project Crossmodal Learning (NSFC61621136008/DFG TRR-169).

References

1. Lakoff, G., Johnson, M.: Metaphors we live by. University of Chicago press, Chicago (2008)
2. Landau, M.J., Meier, B.P., Keefer, L.A.: A metaphor-enriched social cognition. Psychol. Bull. **136**(6), 1045–1067 (2010)
3. Gillan, D. J., Bias, R. G.: Use and abuse of metaphor in human-computer interaction. In: 1994 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Humans, Information and Technology, vol. 2, pp. 1434–1439. IEEE (1994)
4. Clark, H.H.: Space, time, semantics, and the child. In: T., Moore. (ed.) Cognitive Development and Acquisition of Language, pp. 27–63. Academic Press, New York (1973)
5. Casasanto, D.: Who's afraid of the big bad Whorf? Crosslinguistic differences in temporal language and thought. Lang. Learn. **58**(s1), 63–79 (2008)
6. Chen, C.: A contrastive study of time as space metaphor in English and Chinese. Theory Pract. Lang. Stud. **4**(1), 129–136 (2014)
7. Boroditsky, L., Fuhrman, O., McCormick, K.: Do English and Mandarin speakers think about time differently? Cognition **118**(1), 123–129 (2011)
8. Santiago, J., Lupiáñez, J., Pérez, E., Funes, M.J.: Time (also) flies from left to right. Psychon. Bull. Rev. **14**(3), 512–516 (2007)
9. Gentner, D., Imai, M., Boroditsky, L.: As time goes by: evidence for two systems in processing space → time metaphors. Lang. Cogn. Proc. **17**(5), 537–565 (2002)
10. McGlone, M.S., Harding, J.L.: Back (or forward?) to the future: The role of perspective in temporal language comprehension. J. Exp. Psychol. Learn. Mem. Cogn. **24**(5), 1211–1223 (1998)
11. Boroditsky, L.: Metaphoric structuring: understanding time through spatial metaphors. Cognition **75**(1), 1–28 (2000)
12. Hauser, D.J., Carter, M.S., Meier, B.P.: Mellow Monday and furious Friday: the approach-related link between anger and time representation. Cogn. Emot. **23**(6), 1166–1180 (2009)
13. Margolies, S.O., Crawford, L.E.: Event valence and spatial metaphors of time. Cogn. Emot. **22**(7), 1401–1414 (2008)
14. Lee, A., Ji, L.J.: Moving away from a bad past and toward a good future: feelings influence the metaphorical understanding of time. J. Exp. Psychol. Gen. **143**(1), 21–26 (2014)
15. Ruscher, J.B.: Moving forward: the effect of spatiotemporal metaphors on perceptions about grief. Soc. Psychol. **42**(3), 225–230 (2011)

16. Gable, P., Harmon-Jones, E.: The motivational dimensional model of affect: implications for breadth of attention, memory, and cognitive categorisation. Cogn. Emot. **24**(2), 322–337 (2010)
17. Richmond, J., Wilson, J.C., Zinken, J.: A feeling for the future: how does agency in time metaphors relate to feelings? Eur. J. Soc. Psychol. **42**(7), 813–823 (2012)
18. Erickson, T.D.: Working with interface metaphors. In: Human-Computer Interaction, pp. 147–151. Morgan Kaufmann Publishers Inc. (1995)

Perception and Psychological Issues in HCI

"I'm not Stupid" - Attitudes Towards Adaptation Among People with Dyslexia

Gerd Berget[1] and Siri Fagernes[2](✉)

[1] Department of Archivistics, Library and Information Sciences, Faculty of Social Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway
gerd.berget@oslomet.no

[2] Mobile Technology Lab, Department of Technology, Westerdals Oslo School of Arts, Communication and Technology, Oslo, Norway
siri.fagernes@westerdals.no

Abstract. A significant portion of the population have dyslexia, which is commonly associated with reading and writing difficulties. In the context of developing materials well-suited for users with reading disorders, one solution has been to develop materials especially targeted at dyslexic users. However, how are the attitudes among users with dyslexia towards adaptation? In this paper, we report the findings from qualitative interviews with 20 adults with dyslexia. The main finding was that they were sceptical towards adapted products, among others because it made them "feel stupid" and because the adapted format affected the reading experience negatively. In this paper we argue to instead work within the universal design paradigm, trying to develop products and services usable by all people, thus reducing the need for particular user groups to utilise "special solutions".

Keywords: Dyslexia · Adaptation · Universal design

1 Introduction

Dyslexia is a widespread cognitive impairment, which is prevalent in 5–10% of every population [1]. Most dyslexic people experience reading and writing difficulties, and a majority of dyslexics also exhibit reduced concentration [2], impaired short-term memory [3] and poor rapid naming skills [4]. Consequently, many activities in everyday life may be cognitively challenging for dyslexic users. Examples of potential difficulties range from reading text [5] to web navigation [6] and searching for online information [7].

A variety of adapted solutions have been developed, among others high-content/low-skills books, that are especially made for people with reading disabilities such as dyslexia. In Norway, an organization called Books for everyone produces a variety of adapted books for different user groups, including books with easy language and a particular typography with the purpose of making books accessible for people with reading impairments. In some of these books, sentences are actively split over several lines, which has resulted in a layout with quite short line lengths. These books have a

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 237–247, 2018.
https://doi.org/10.1007/978-3-319-91238-7_20

layout that differs from "regular books", and consequently may look "adapted". Such a practise is the starting point for the interview data presented in this paper.

People with dyslexia are commonly reported to experience low self-esteem [8], which is often a result of negative experiences in school- or work-related settings [9]. In this study we investigate the attitudes among dyslexic users towards adapted products, with the aim of exploring whether such solutions are purposeful or regarded as stigmatising. These findings have implications for researchers and practitioners working with dyslexic users in general, and adapted texts in particular.

The research questions that form the basis for this paper is: *What are the attitudes among people with dyslexia towards adapted texts?* The motivation behind this study is to better understand how to accommodate the needs of dyslexic users, while at the same time maintaining a respectful and purposeful approach when designing inclusive user interfaces or other services or products.

2 Background

Previous studies have addressed the needs of dyslexic users in HCI-related contexts, with the aim to improve accessibility. Such studies have typically either investigated a potentially problematic task, for instance [6, 10] or suggested or evaluated guidelines, for example [11]. Several researchers have investigated how to present readable text for dyslexic users. Most of these studies have focused on digital texts and devices, with a limited focus on printed books. One of the reasons for this perspective may be that it has been suggested that digital materials are most suitable for dyslexic users [12].

Rello and Baeza-Yates [13] claim that the presentation of text is especially important for people with dyslexia. For instance, Rello and Baeza-Yates [13] concluded that the font styles sans serif, monospaces and roman font improved performance during screen reading. In contrast, proportional fonts, serif and italics affected reading performance negatively. O'Brien et al. [14] found that larger fonts might increase reading speed of dyslexic people. This finding was in accordance with Rello et al. [15], who reported that font size has a significant impact on the understandability and readability of a screen text, suggesting using 18-point font. In contrast, British Dyslexia Association [16] recommends 12–14 point, but emphasises that some dyslexics may require even bigger font.

Letter spacing has also been addressed, and according to Rello et al. [15], line spacing has no effect on readability for dyslexics when reading on-screen texts. However, letter spacing seems to be important. For instance, Zorzi et al. [17] reported that extra-large letter spacing benefited dyslexic readers, a finding which is in accordance with Marinus et al. [18].

Regarding line lengths, there is also an emphasis on digital texts, and the results in previous are so far contradictory. Schneps et al. [12] studied e-readers in a PAD and POD condition and found that it was most efficient to read on small digital devices such as PODs. The PAD condition displayed on average 67.2 characters per line, while the POD condition had 12.7 characters. Schneps et al. [19] concluded that shorter lines might be beneficial. This finding is in contrast to the guidelines from British Dyslexia

Association [16], which recommends 60–70 characters per line. Moreover, Rello and Baeza-Yates [20] also investigated column widths in digital texts and found that these had no significant effect on the readability for dyslexics, who expressed no preferences regarding line lengths. In contrast, control (non-dyslexic) users found line lengths of 44 characters to be most readable. Further, Rello and Baeza-Yates [20] reported that some of the dyslexic users actually preferred the wide columns because the overall impression of the text was that it seemed shorter and was consequently more motivating to read. This study suggests that layout may affect the motivation for a person to read, and that such design may be particularly important for dyslexic readers.

In a review of accessible typology for dyslexic users, Jackson [21] reported that most of this research has been conducted on participants under 18 and often for younger children and emphasises the need to investigate these issues in the adult dyslexic population as well.

Previous attempts to make products and services for users with impairments were within the context of accessible design, where adaptation for users "with special needs" was quite common [22]. However, this approach is now increasingly being replaced by the universal design philosophy. The purpose of universal design is to develop products or services that are applicable by all people, despite differences in age, gender, cultural background or functional levels [23]. The objective of universal design is to design one solution suitable for everyone, thus removing the need for especially adapted solutions for particular user groups [22], such as dyslexics.

One example of accessibility in practice is the high content/low skills-books produced by Books for Everyone with short line lengths. The purpose behind these books is to make reading more accessible and less demanding for people with reading difficulties. Such books were therefore used as a starting point for a reading experiment and the qualitative interviews with adults with dyslexia presented in this paper.

The need to not feel different may be closely related to low self-esteem, which is reported to be a key issue in people with dyslexia, both adults and children [8, 24, 25], and especially among females [26].

Adaptation in general and adapted books in particular are topics that are rarely addressed in the research literature. In a study of children with dyslexia, Thiessen and Dyson [27] found that dyslexics preferred books that resembled books that their peers were reading over books that were considered easier to read by typographic convention. However, little is known about the attitudes among adult dyslexics towards adaptations.

3 Method

This study was a within-subjects design, with no control group. The participants were 20 adults with dyslexia, recruited through social media and the organisations Dyslexia Norway and Books for Everyone. A total of 11 participants were female (55%), while 9 were male (45%). Average age was 26.2 years, with the youngest participant being 18 years and the oldest 40 years. The following inclusion criteria was applied; the participants needed to be above 18 years, have a dyslexia diagnosis, and not wear glasses, since these would interfere with the eye-tracking glasses used in the reading experiments. A total of 22 participants conducted the tests, but two were excluded due to too high scores on the dyslexia screening test, indicating too proficient reading skills.

3.1 Procedure

The session started with providing general information about the study, signing of consent forms and registering participant data before several tests were conducted (Fig. 1).

Each participant was screened for dyslexia. A Norwegian Word Chain Test [28] was applied, which is a commonly used test for such purposes, and has been applied in other studies for example [29–31]. In this test, a low score is regarded as indicative of dyslexia, and adults who score below 43 points are recommended to conduct further diagnostic tests. Using this screening removed the need for access to sensitive medical diagnostic papers, while at the same time confirming the diagnosis and providing more details on the severity of the dyslexia in each participant.

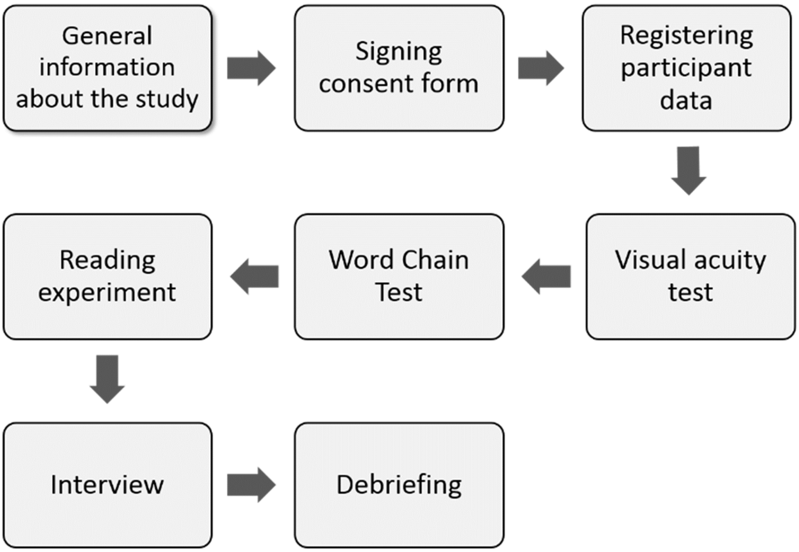


Fig. 1. Experimental procedure

A Landolt C visual acuity test was applied with a distance of 40 cm for short vision according to the European standard [32], to ensure that reduced or blurred vision would not affect the results of the reading tasks. The participants had at least a visual acuity of 0.8 with both eyes open, results that are in accordance with the limits of normal visual acuity [33]. A short interview was also conducted before the actual reading experiment started.

The participants were asked to read 12 different texts from three books, four from each book. Each text was only read once and presented in one of four different layouts (Fig. 2). The participants were presented with the different books and conditions in a randomised order. Participants were given one of 6 premade folders, where all the books were presented in different orders, and with a variation of which pages that were presented in each condition. The texts were inputted in a hard folder, to resemble the feeling of holding and reading a physical book as closely as possible.

The reading session was immediately followed by an interview regarding what characteristics of a text they considered most important in terms of readability. Moreover, participants were showed excerpts from the adapted books they had read in the different conditions and asked about their attitudes towards such texts in particular and adaptations in general. They were also asked to rank the different conditions from 1–4, where 1 was the most favoured format and 4 the least preferred layout. The participants were given a gift card of 500 NOK after completing the session.

The project was approved and ethically screened by Norwegian Centre for Research Data. At the beginning of the session, all participants signed consent forms, and were informed that they could withdraw from the study at any time without any negative consequences. No participants were related to the researchers.

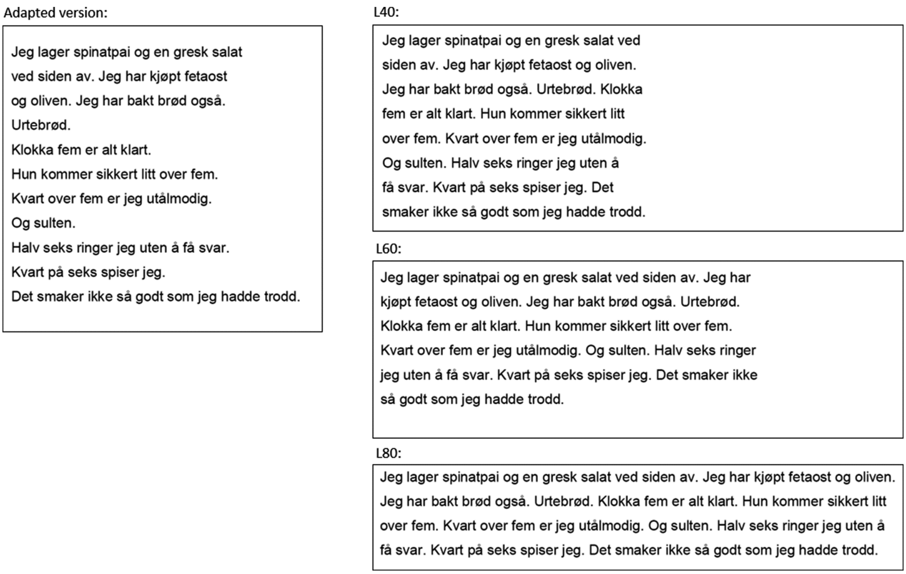


Fig. 2. Page 17 from Lindkvist (2013) in the four different formats. Reproduced with permission by author and publisher.

3.2 Materials

Three adapted books in Norwegian (Table 1) were used as a starting point for the reading experiment. The adapted books were formatted with sentences that were broken over several lines, with the motivation to reduce the line lengths.

During the reading experiments, participants were given four texts from each book. The pages were given in an unedited, original form (with typically short line lengths), and conditions with 40, 60 and 80 characters per line, including white space (see Fig. 2). No words were split, and all the texts from each book had the same amount of words on each page. The texts were printed on a pearl white, 130-g paper, in Arial font size 14 points, and with a left-justified text and ragged right edge.

Table 1. Adapted books included in the study.

Author(s)	Title [Translated title]	Year
E. Lindkvist	Ditt røde hår, Unn [Your red hair, Unn]	2013
H. Hagerup & K. Roskifte	Barnet mitt [My child]	2015
V. Salinas	Og [And]	2016

4 Results

Each participant was asked to rank the four different formats according to their personal preferences. Each participant hence gave the score 1–4 to each of the formats; 1 indicating their favourite, and 4 if it was the least favourable format.

The wider formats (L60 and L80) were significantly more popular than the narrower formats (L40 and the adapted format). As many as 18 out of 20 listed one of the widest formats as their favourite (See Fig. 3).

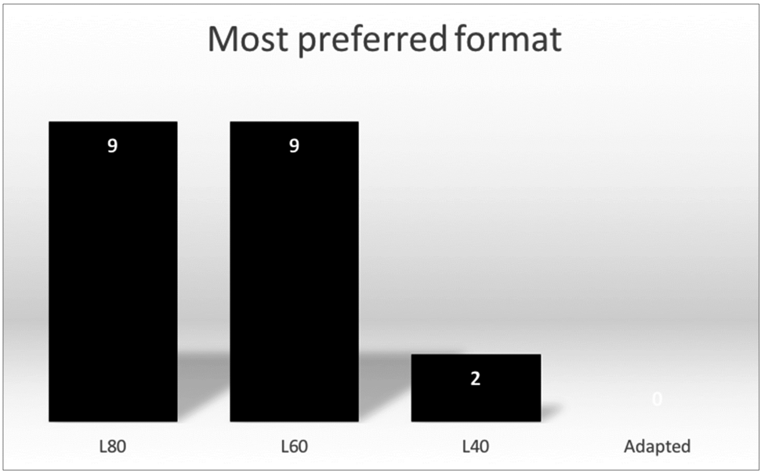


Fig. 3. Preferred format, in number of participants.

Only 2 participants preferred the shorter format (L40), and nobody listed the adapted format as their favourite. A total of 12 participants regarded the original, adapted format as their least favourable reading format (Fig. 4).

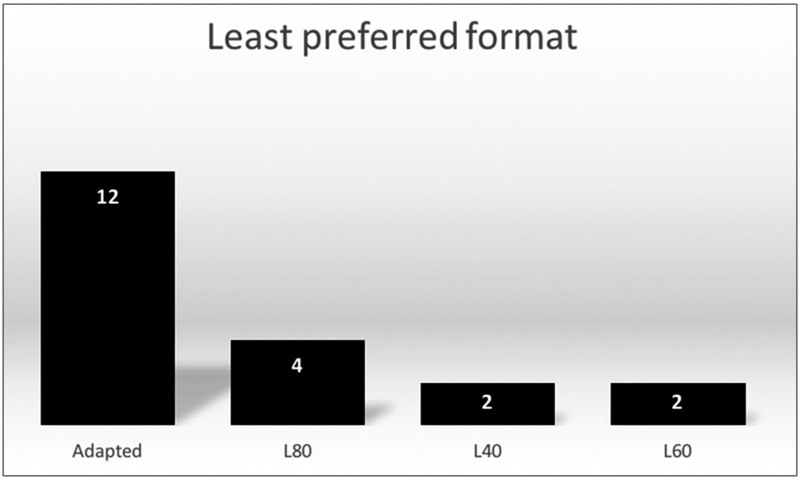


Fig. 4. Least favoured format, in number of participants.

We also computed the average ranking of each format (Fig. 5). The two wider formats, L80 and L60, scored 1.8 (std = 1.10) and 1.95 (std = 1.00), respectively, indicating that these two were the most favourable formats overall. The original adapted format scored the lowest overall, with an average rank of 3.3 (std = 0.60).

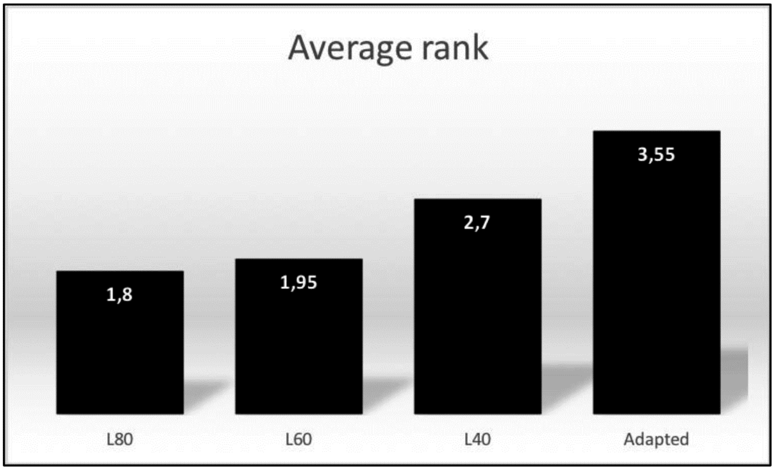


Fig. 5. Average ranking of each format.

5 Discussion

Our results indicate that the adapted format, which often means that longer sentences are split over several lines, are less favoured by dyslexics, compared to formats with longer lines. There seems to be a strong linkage between preferred format and the reading strategy applied.

Several participants disliking the adapted format, explained how the short lines disrupted their reading. A few participants mentioned how they used a reading strategy searching for the end of the sentence, and then going back to analyse the content of the sentence further, before reading the remaining text: "*I have to see the whole sentence before I understand it. Then I can go back to the rest of the text. It is much more work to go back in a split-up text, because then I lose the overview.*" Consequently, if a sentence was broken over several lines, the reading process was perceived as more difficult.

Many participants mentioned how skipping between lines made them confused and ruined the flow of their reading because they lost track of where they were in the text. For instance, one participant said that "*it would be easier to read if the lines were longer. Now I have to navigate to the next line often, losing where I was, and have to read the whole paragraph again.*" A wider format will be less likely to split a sentence over more than one or two lines and will hence make the reading process less confusing. One participant mentioned using a paper sheet to aid her in the reading, where she took a pause at the end of each sentence. This strategy was also difficult with many split sentences, implying that adapted books may interfere with the coping strategies of readers with dyslexia. This is an issue that needs to be investigated further.

The two most popular formats (L60 and L80), also received mixed feedback. Some participants disliked the widest format (L80), stating that the lines were too long and felt a bit overwhelming. Others liked how it made the overall block of text shorter, and in that respect, less overwhelming, which is in accordance with Rello and Baeza-Yates [20]. In addition, one participant specifically expressed how she felt more challenged by the wider format, giving her a sense of mastering reading. In her case, more challenging texts seem to be positive for her reading experience.

While the original and L40 were quite similar with respect to score in the 'most liked' category, the difference is surprisingly large with respect to be the least liked. Only 2 out of 20 listed L40 as their least favourable (Fig. 4). It seems like the adapted format is more provoking, and several participants commented on how the adapted format gave them associations to children's books or primary school books and gave them less interest in reading them. The comments about "feeling stupid" and "associations of children's books" support this claim.

Many of the dyslexic users expressed a general scepticism towards adaptation in general and stated that they "felt stupid" when presented with material targeted at dyslexics. For instance, one participant stated the following "*It should not be obvious that this book is made for someone with dyslexia. You shouldn't have to feel different*". Many participants also commented that it made them feel like they had to use the "elementary school version". One participant said, "*If they try to stupefy the text too much, it doesn't encourage to reading*".

A number of participants also stated that they would actually rather struggle with a difficult text than utilise the adapted versions, which is in accordance with Thiessen and Dyson [27]. "*It is nice with space between the lines, but not so adapted that one feels stupid*". The main conclusion regarding the attitudes towards adaptations was that it does not seem purposeful to use such an approach for this user group, who already are very conscious about "being different" and seem to perceive adapted products as stigmatising.

Participants enjoying the adapted format, explained how they liked what the format did to the text, making it more lyrical ("*like a poem*"). It was also commented on how the adapted format sometimes made text-dialogue more apparent, making the text easier to read. However, the short line lengths also felt confusing regarding genre: "*In one of the conditions I first thought I was reading a poem*". Another stated that "*I get a very poem-like feeling of the adapted text. I do not like reading that (…) it affects my reading flow*". Consequently, there may be other more purposeful strategies than reducing line lengths when developing books that are easily read by dyslexic users. However, this issue needs to be investigated further.

6 Conclusion

The findings reported in this paper are mainly focused on the reading experience as perceived by the dyslexics, and not on their actual reading performance. Two main issues are apparent in this study; namely confusion in the reading process and low self-esteem as results of the adapted reading format.

Several participants reported a reading strategy involving looking for the end of a sentence. The shorter formats meant often having to skip vertically across several lines, probably imposing more cognitive strain on the participants. Consequently, it is highly likely that shorter line lengths do not necessarily facilitate reading for users with dyslexia. Further, it seems very important to not develop products that feel adapted, but rather make books that may be more easily read by a variety of users, thus removing the feeling of "being stupid" and being stigmatised.

The findings of this study have implications for the development of different products and services directed at dyslexic users and may be transferred to areas such as user interfaces and various other ICT-based solutions targeted at dyslexic users. We argue against especially adapted solutions, and rather recommend the universal design paradigm. Further, the lack of consistency in this user group makes it difficult to develop guidelines and suggest that investing resources on developing a higher level of user control is probably more beneficial for this user group.

Acknowledgements. The authors are grateful to *Leser søker bok* (Books for Everyone) and Dyslexia Norway for helping with recruiting participants, and *to Leser søker bok* for contributing with gift cards to the participants.

References

1. Ahissar, M.: Dyslexia and the anchoring-deficit hypothesis. Trends Cogn. Sci. **11**(11), 458–465 (2007)
2. Mortimore, T., Crozier, W.R.: Dyslexia and difficulties with study skills in higher education. Stud. High. Educ. **31**(2), 235–251 (2006)

3. Majerus, S., Cowan, N.: The nature of verbal short-term impairment in dyslexia: the importance of serial order. Frontiers Psychol. **7**, article 1522 (2016)
4. Jones, M.W., Snowling, M.J., Moll, K.: What automaticity deficit? Activation of lexical information by readers with dyslexia in a rapid automatized naming Stroop-Switch Task. J. Exp. Psychol. Learn. Mem. Cogn. **42**(3), 465–474 (2016)
5. Lauterbach, A.A., Park, Y., Lombardino, L.J.: The roles of cognitive and language abilities in predicting decoding and reading comprehension: comparisons of dyslexia and specific language impairment. Ann. Dyslexia **2016**, 1–18 (2016)
6. Al-Wabil, A.: The Effect of Dyslexia on Web Navigation. City University London, London (2009)
7. Berget, G., Sandnes, F.E.: Searching databases without query-building-aids: implications for dyslexic users. Inf. Res. **20**(4), Article 689 (2015)
8. Carawan, L.W., Nalavany, B.A., Jenkins, C.: Emotional experience with dyslexia and self-esteem: The protective role of perceived family support in late adulthood. Aging Ment. Health **20**(3), 284–294 (2015)
9. Caroll, J.M., Iles, J.E.: An assessment of anxiety levels in dyslexic students in higher education. Br. J. Educ. Psychol. **76**(3), 651–662 (2006)
10. MacFarlane, A., Buchanan, G., Al-Wabil, A., Andrienko, G., Andrienko, N.: Visual analysis of dyslexia on search. In: Nordlie, R., Pharo, N. (eds.) Proceedings of the 2017 Conference on Human Information Interaction and Retrieval, pp. 285–288. ACM, New York (2017)
11. de Santana, V.F., de Oliveira, R., de Almeida, L.D.A., Baranauskas, M.C.C.: Web accessibility and people with dyslexia: a survey on techniques and guidelines. In: Vigo, M., Abascal, J., Lopes, R., Salomoni, P. (eds.) Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A 2012) Lyon, France, 16–17 April, Article no. 35. ACM, New York (2012)
12. Schneps, M.H., Thomson, J.M., Chen, C., Sonnert, G., Pomplun, M.: E-readers are more effective than paper for some with dyslexia. PLIS ONE **8**(9), e75634 (2013)
13. Rello, L., Baeza-Yates, R.: Good fonts for dyslexia. In: ASSESTS 2013 Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, Article no. 14. ACM, New York (2013)
14. O'Brien, B.A., Mansfield, J.S., Legge, G.E.: The effect of print size on reading speed in dyslexia. J. Res. Reading **28**(3), 332–349 (2005)
15. Rello, L., Pielot, M., Marcos, M.-C., Carlini, R.: Size matters (spacing not): 18 points for a dyslexic-friendly Wikipedia. In Brajnik, G., Salomoni, P. (eds.) Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, 13–15 May 2013, Rio De Janeiro, Brazil (W4A 2013), Article no. 17. ACM, New York (2013)
16. British Dyslexia Association: Dyslexia Style Guide. http://www.bdadyslexia.org.uk/common/ckeditor/filemanager/userfiles/About_Us/policies/Dyslexia_Style_Guide.pdf. Accessed 31 Jan 2018
17. Zorzi, M., Barbiero, C., Facoetti, A., Lonciari, I., Carrozzi, M., Montico, M., Bravar, L., George, F., Pech-Georgel, C., Ziegler, J.C.: Extra-large letter spacing improves reading in dyslexia. Proc. Nat. Acad. Sci. U.S.A. **109**(28), 11455–11459 (2012)
18. Marinius, E., Mostard, M., Segers, E., Schubert, T.M., Madelaine, A., Wheldall, K.: A special font for people with dyslexia: does it work and if so, why? Dyslexia **22**(3), 233–244 (2016)
19. Schneps, M.H., Thomson, J.M., Sonnert, G., Pomplun, M., Chen, C., Heffner-Wong, A.: Shorter lines facilitate reading in those who struggle. PLOS ONE **8**(8), e71161 (2013)
20. Rello, L., Baeza-Yates, R.: How to present more readable text for people with dyslexia. Univ. Access Inf. Soc. **16**(1), 29–49 (2017)

21. Jackson, J.E.: Towards universally accessible typography: a review of research on dyslexia. J. Technol. Persons Disabil. **2**, 155–165 (2014)
22. Petrie, H., Kheir, O.: The relationship between accessibility and usability of websites. In: CHI 2007, CHI Conference on Human Factors in Computing Systems, pp. 307–406. ACM, New York (2007)
23. Story, M.F.: Maximizing usability: the principles of universal design. Assistive Technol. **10**(1), 4–12 (1998)
24. Humphrey, N.: Teacher and pupil ratings of self-esteem in developmental dyslexia. Br. J. Spec. Educ. **29**(1), 29–36 (2002)
25. Riddick, B., Sterling, C., Farmer, M., Morgan, S.: Self-esteem and anxiety in the educational histories of adult dyslexic students. Dyslexia **5**(4), 227–248 (1999)
26. Alexander-Passe, N.: How dyslexic teenagers cope: an investigation of self-esteem, coping and depression. Dyslexia **12**(4), 256–275 (2006)
27. Thiessen, M., Dyson, M.C.: Typography for children with reading difficulties: preferences for type in reading books. Int. J. Book **6**(2), 115–122 (2009)
28. Høien, T., Tønnesen, G.: Ordkjedetesten [The Word Chain Test]. Logometrica, Bryne (2008)
29. Solheim, O.J., Uppstad, P.H.: Eye-tracking as a tool in process-oriented reading test validation. Int. Electron. J. Elementary Educ. **4**(1), 153–168 (2011)
30. Dahle, A.E., Knivsberg, A.-M.: Internalizing, externalizing and attention problems in dyslexia. Scand. J. Disabil. Res. **16**(2), 179–193 (2014)
31. Berget, G.: Search and find? An accessibility study of dyslexia and information retrieval. University of Oslo, Oslo (2016)
32. ISO: ISO 8596 Ophthalmic Optics: Visual Acuity Testing – Standard Optotype and its Presentation. ISO, Geneva (2009)
33. Zhang, P., Bobier, W., Thompson, B., Hess, R.F.: Binocular balance in normal vision and its modulation by mean luminance. Optom. Vis. Sci. **88**(9), 1072–1079 (2011)

Check for updates

Research on Sticker Cognition for Older Adult Using Instant Messaging

Cheih-Ying Chen(✉)

Department of Product Innovation and Entrepreneurship,
National Taipei University of Business, Taipei, Taiwan, R.O.C.
c.y.chen@ntub.edu.tw

Abstract. The growing ubiquity of smartphones has changed how older people live and the advent of instant messaging apps has reshaped their communication methods. Stickers are arguably the most popular feature of LINE, an instant messaging app. In this study, in-depth interviews were conducted with elderly users of instant messaging apps to explore the effects of the emotional meanings behind stickers on their social connections and level of understanding of the other party's meaning in an online conversation. The findings suggested that using stickers improved the participants' social connections, made them happier, and enhanced their involvement in LINE chats. Moreover, stickers with perceived positive and negative connotations enhanced the participants' social connections; in this regard, no noticeable differences were observed between stickers with positive and negative connotations. However, the participants had difficultly accurately interpreting stickers featuring highly expressive facial expressions or diverse physical movements; some even found such stickers ambiguous and misunderstood the implied meanings.

Keywords: Sticker · Instant messaging · Older adult · User experiences

1 Older Adult and Instant Messaging

1.1 Motivation

Health and social wellbeing are common concerns among older adults. It is estimated that one in four people in Taiwan will spend the remainder of their life living alone. Across Taiwan, older adults with limited contact with their relatives have small social networks and are frequently neglected. They tend to idle away time watching TV at home or sitting in front of their homes watching passers-by and passing vehicles. Older adults who seldom participate in social activities or events of interest exhibit less positive emotions. However, the Internet that enables diverse communication methods can be a solution for the feeling of loneliness in older adults [1]. The Internet can be harnessed to prepare people mentally for a solitary life and expand their social circle in later life after their children have become too busy to come home and their friends have limited time for gatherings. Thus, "solitude" may not be a synonym of "loneliness". Communication technologies including telephones, mobile phones, and the Internet enable various methods of interpersonal communication and allow users to maintain

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 248–256, 2018.
https://doi.org/10.1007/978-3-319-91238-7_21

relationships with others without time or spatial constraints. Older adults engage in interpersonal communication in a variety of manners; for example, holding social gatherings, using the telephone, and writing letters. They can also socialize through Internet technology–enabled features such as mobile phones, email services, and instant messaging apps. Thus, older adults can reduce the feeling of loneliness and expand their social network by using smartphones to maintain contact with their family members and relatives and connect with their friends.

1.2 Older Adults' Social Connections

A sense of connection refers to the state in which a person is connected or acquainted with others; such a state may involve family bonding, friendship, or acquaintance with neighbors. Such social connections provide joy and support, enable a person to feel a sense of belonging, and assist a person in creating social capital, thereby enabling society as a whole to operate effectively. Moreover, people with more active social lives tend to be more physically and mentally robust and more capable of handling major life changes [2].

The sense of connection is a fundamental need for human beings, and it is the sense of belonging that a person feels in his or her community or social network [3]. Paying a personal visit, holding a social gathering, making a phone call, or initiating a conversation via the Internet can induces feelings of comfort and belonging in a person. Social connection can be measured based on six criteria: whether phone and Internet connections are available at home, interaction with family members and friends, interaction between young people and their parents, level of trust in others, loneliness, and participation in voluntary work. Fincham and Beach [4] used the construct "relationship flourishing" to describe a relationship characterized by happiness, intimacy, growth, and resilience, which further heightens feelings of happiness and wellbeing.

1.3 Role of Instant Messaging in Social Connections

Instant messages are pieces of textual information transmitted between electronic devices and typically sent in the form of short passages of text between users [5]. Advances in smartphones and wireless communication technologies have given rise to many instant messaging and social media apps that enable smartphone and computer users to send textual, graphical, and audio information, photos, videos, and electronic files, thereby enabling users to communicate in a convenient, rapid, and varied manner [6].

A 2015 survey conducted by Foreseeing Innovative New Digiservices—a research branch of the Institute for Information Industry (III)—revealed that older adults living in Taiwan (a) increasingly considered retirement as a transitional period of life, (b) enriched their retirement lives by using high-tech devices to pursue the goals they had set for themselves when they were young, and (c) expected to remain up to date with technological developments to meet their recreational, social, and learning needs. The survey results also indicated that elderly people (a) depended increasingly on technology, particularly smartphones, (b) expressed a strong preference for LINE, and (c) used technology to access health-related information, contact their relatives, follow new episodes of soap operas, and listen to music. Overall, technology enables Taiwanese older adults to fulfill their expectations of daily living and live well-rounded and vibrant lives [7].

Statistics from SimilarWeb (2016) revealed that WhatsApp is the world's most popular instant messaging app, Facebook Messenger is the second most popular, and LINE has the highest degree of penetration of any messaging service in Taiwan. Both Facebook Messenger and LINE enable users to send text messages, pictures, audio, photos, and stickers, make phone calls, and deliver short messages to multiple users instantaneously. LINE accounts for as much as 91.9% of instant messaging app use in Taiwan because it helps to cement one's interpersonal relationships, facilitates communication, and is easy to use. In December 2016, the III [8] released a Report on 2017 Trends in New Media, which suggested that 4G Internet connection services had over 17 million subscribers in Taiwan, 90% of whom reported having a LINE account. Furthermore, the number of LINE users in Taiwan doubled within a 2-year period; the widespread use of smartphones for socializing and entertainment has contributed to the development of numerous commercial services and ample business opportunities.

1.4 Older Adults' Use of Instant Messaging Apps on Smartphones

Older adults differ somewhat from younger people in terms of lifestyle, technology use, and attitudes toward virtual commerce. The III's 2016 survey of older adults' needs for using 4G Internet found that 60.2% of respondents aged 55 years or older owned a smartphone; their smartphone use focused largely on making phone calls and they showed high preferences for LINE and Facebook. Additionally, approximately 60% of over-55-year-olds improved their interpersonal interactions through LINE; over 50% used the app to engage in video chats with their children or grandchildren, thereby strengthening their relationships [8].

2 Stickers

2.1 Stickers on Instant Messaging

Stickers are used extensively on instant messaging apps. The precursor of stickers is the *emoji*, which is designed by Shigetaka Kurita (a Japanese engineer) and etymologically consists of two Japanese language characters—*e* for "picture" and *moji* for "character." Just as languages evolve with social change, stickers have gradually influenced users' communication with each other and enriched instant messages.

A 2016 report released by Yahoo's Flurry Analytics described stickers on instant messaging apps as a contributing factor in the development of personalized mobile apps. Many existing instant messaging apps offer stickers as a means of nontextual communication. Sending stickers can facilitate communication via messaging apps by fostering understanding between users; in the workplace, using stickers judiciously can improve emotions involved in communication. Social factors are the primary factors in the use of instant messaging apps within a business organization; the more messaging app users in an organization and the more positively the management perceives use of the apps, the more frequently are the apps used. Thus, social factors affect the intention to use instant messaging apps. Stickers are related to the message input method used, and some are intended to reflect ongoing social events and mainstream culture and subcultures. However, for old-age users, the diversity of stickers can hamper communication via instant messaging apps.

Therefore, this study discussed older adults' difficulties in using stickers on messaging apps; this paper proposes solutions to these difficulties to increase older people's sticker usage. Stickers appear not only as pictures but also as accentuated images with sound effect, and a diverse range of genres of stickers are currently available on the market. This study analyzed older adults' perceptions of stickers.

2.2 Line with Stickers

A 2016 Nielsen survey found that LINE had a usage rate of over 90% in Taiwan, with over 17 million users; more than 90% of people aged 20 to 65 years used the app, and nine out of ten LINE users had added official business accounts to their contact lists to obtain complimentary stickers. Evidently, stickers are popular among messaging app users in Taiwan. However, the survey did not include statistics of LINE usage among people aged 65 or above. LINE stickers used widely in Taiwan are classified under such genres as male and female characters, cute, sweet, comforting, funny/hilarious, dialect/buzzwords, animals, and food. All these sticker genres were adopted as cartoonistic stickers in the present study.

As an enhanced variant of their conventional equivalents, stickers are virtual images created in a personified manner designed to vividly convey the mood or tone of a sent message. Emoticons are akin to facial expressions and serve to express certain nonverbal behaviors [9].

The Shannon–Weaver model of communication suggests that the delivery of a message is subject to external noise, which influences how the recipient interprets the message [10]. Accordingly, stickers are characteristically open to interpretation and can create misunderstanding and confusion among older LINE users.

3 User Experiences

To analyze older adults' difficulties in using stickers on instant messaging apps, this study explored older adults' experiences of sticker use to address their problems with stickers. Older adults tend to have difficulty typing on mobile devices because of their slow movement and the insufficiently large keypads on such devices. Moreover, older adults tend to have an inadequate understanding of stickers for instant messaging apps and younger people's use of language, and rarely use such stickers to express their feelings. However, using stickers can enable older adults to interact more readily and conveniently with their younger relatives and friends.

This study used various methods to investigate older adults' experiences of using instant messaging apps with a view to elucidating their difficulties in sending textual and graphical messages and stickers. These methods are detailed as follows.

3.1 Focus Groups

As a qualitative research method, a focus group enables a researcher to develop an effective, timely, and thorough understanding of a topic and gather data that can actually reflect the said topic. Despite some implementation-related limitations, this method is more effective than are other qualitative techniques. A focus group interview should be planned by specifying the participant group, the moderator, the length of the questionnaire to be administered, and other questionnaire-related factors in accordance with the interview topic; formulated in such a manner, the interview can be flexibly implemented, and thus high-value data can be collected [11].

In this study, focus group interviews were conducted in a structured manner, each with six to eight representative older adults. The moderator asked the participants to share their experiences of using messaging apps. Guided by open-ended questions, they stated their opinions and needs regarding stickers and discussed what types of stickers met their preferences and needs and facilitated their communication and mutual understanding. Through such discussion, the participants were inspired to have new thoughts concerning the topic, and thus the collected data can reflect opinions from various perspectives. By conducting the in-depth interview, the researchers can comprehensively understand older adults' perceptions and evaluation of stickers and identify the difference between the older adults' previous experience and current perceptions.

3.2 Contextual Inquiry

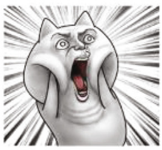
Analyzing user experiences enables product designers or engineers to better understand users' needs for a specific product or difficulty using it. Contextual inquiry is a qualitative research method where a researcher interviews a participant to identify his or her problems, explore the causes of said problems, and address said problems. Through this method, a product designer can ask a user questions about his or her behaviors from using the product in question for the first time to experiencing problems with it. In so doing, the product designer can face-to-face communicate with users and comprehensively understand users' difficulties and product usage experience [12]. Therefore, contextual inquiry involves collecting primary data, which encompass a user's reported difficulties in using the product; the designer can then collate these data to improve the design and functionality of the product.

Regardless of whether they live alone, older adults tend to have fixed habits of use. This study observed how older adults used stickers to respond to positive and negative critical events in various representative scenarios, as well as how they interpreted and felt about the stickers they used. The findings are expected to improve older adults' understanding of stickers used in messaging apps.

3.3 Experimental Procedure

The experiment comprised two stages. In the first stage, a focus group of five participants aged 55–60 years held a discussion to select 16 stickers representing typical human emotions and feelings (as shown in the Table 1).

Table 1. The participants' interpretations of all 16 stickers

Sticker				
The participants' interpretations	1.My apologies 2.Sorry 3.Can't recall 4.Acting cute	1.I am cool 2.It's you 3.You get me 4.I am confident on myself	1.I love you 2.I like it 3.Blowing a kiss 4.Over the moon;	1.Cool, isn't it? 2.I'm awesome! 3.I'm smart 4.Things are going well 5.Let me think
Sticker				
The participants' interpretations	1.Thank you 2.I need a hand 3.Help me	1.Acting cute 2.I like it 3.Blowing a kiss 4.Annoying 5.Joking 6.Making a face	1.Delighted 2.Blissful 3.Satisfied	1.Mocking 2.Ha ha ha! This is hilarious! 3.Making fun of a person's stupidity
Sticker				
The participants' interpretations	1.Sigh 2.Anyway 3.It doesn't matter 4.Resigned 5My bad	1.Having a yearning 2.Fantasizing 3.In a good mood 4.I like it	1.Speechless 2.Dumbass 3.At a loss for words 4.Resigned	1.Stupefied 2.Sad 3.Surprised 4.Weeping 5.Stunned by an unexpected incident
Sticker				
The participants' interpretations	1.I won 2.I'm a champion 3.Proud 4.Happy	1.Furious 2.Holding my breath 3.Tsundere 4.Roll my eye	1. Let me see 2. Want a date? 3. Seducing	1.Surprised 2.How could this be? 3.Taken aback 4.Terrified 5.This is impossible

In the second stage, an in-depth contextual inquiry was conducted to ask eight 55–78-year-old participants (who know how to use LINE) about their interpretations of these 16 stickers and how they used the stickers on LINE. The Table 1 presents the participants' interpretations of all 16 stickers.

4 Discussions and Conclusions

During the individual interviews of contextual inquiry, the participants shared a view that using instant messaging apps significantly increased the frequency they contacted with relatives and friends and the time they spent on making contact with them. This strengthened their social connections with their relatives and prompted them to frequently check up on their relatives, thus feeling less lonely. Moreover, stickers, which contain contextual cues including facial expressions and physical movements, were more intelligible to the participants in terms of the emotional overtones indicated by their chatting partners than were messages composed of pure text. Thus, using stickers in a LINE chat can improve a sender and recipient's perceptions of each other's emotions and further cement their relationship. Facial expressions and physical movements are easy to identify; facial expressions are particularly highly expressive and the most crucial form of human communication after textual and spoken information. Because faces are readily observable, any change in facial expression becomes the focus of attention. Particularly, when people make crucial judgements, they typically depend on the messages conveyed from facial expressions. Stickers can serve as symbols of facial expressions, which are invariably subject to interpretation; therefore, the perceived emotions and symbolic meanings conveyed by the same stickers varied among the participants.

The participants reported that stickers effectively expressed senders' positive and negative emotions, LINE stickers were appealing, and stickers representing cartoon characters and nonhuman characters (e.g., dogs, cats, and monsters) were easier to understand than were other stickers. In addition, the participants stated that they often received replies in the form of stickers from their younger relatives and were often unclear about the meanings behind these stickers. For example, when interpreting a popular LINE sticker that features a cute character with a wide mouth producing a red facial expression and shedding tears, some participants remarked that the sender was laughing to tears, whereas others perceived the meaning to be weeping bitterly. The emotional ambiguity of such a sticker can lead to misunderstanding and poor communication. The participants' opinions on the influence of stickers on communication via instant messaging apps are summarized as follows.

(1) The more expressive a character sticker, the greater difficulty the participants had understanding its meaning and the more likely they were to attempt to guess it. Additionally, the participants reported that crucial matters must be discussed through text messages or over the telephone. Considering that older adults often misunderstood the meanings of stickers, young people are advised to use stickers sparingly when talking to their older relatives via instant messaging apps.

(2) Some participants opined that the purpose of sending stickers was to enable the other party to understand the meaning. They did not realize that some stickers may be used to brush others off by young people. The participants added that they worked hard to decipher the meanings of stickers that they received and that ambiguous stickers should not be used.
(3) Few participants ignored the meaning of stickers and perceived them simply as cute images.
(4) All participants stated that stickers were lifelike and interesting. Because their children and grandchildren often sent cartoon stickers, they often responded in kind and tried to understand more about the emotional meanings of stickers.

An effective image conveys its intended message in a straightforward and concrete manner, whereas an ineffective image leaves the receiver struggling for clarity. A user's attitudes and adhesion toward messaging app stickers strengthen if the stickers in question accurately convey the user's emotions and are understandable to the recipient. By contrast, ambiguous stickers tend to cause misunderstanding, which is why stickers should be designed to be self-evident in meaning to prevent confusion among elderly people. Moreover, the findings indicated that stickers had direct and significant effects on the extent to which emotions are expressed through stickers; however, highly expressive or exaggerated stickers can be confusing to older adults. Stickers on instant messaging apps in earlier times were mostly graphical; as the time went by, an increasing number of stickers were accompanied by simple words (e.g., "OK" and "Yes") to complete express the intended meaning and prevent misunderstanding. Stickers without textual information can be understood at first glance only if they convey explicit emotions. Therefore, stickers should be designed in an animated style together with features such as perceptible physical motion, words, sounds, or scenes. The increase of nonverbal cues can effectively enhance the extent to which emotions are expressed through stickers. Future research is suggested to analyze whether stickers with textual and audio elements increase older people's understanding of the intended meanings.

References

1. Mausbach, B.T., Roepke, S.K., Depp, C.A., Patterson, T.L., Grant, I.: Specificity of cognitive and behavioral variables to positive and negative affect. Behav. Res. Ther. **47**(7), 608–615 (2009)
2. Wildevuur, S., van Dijk, D., Hammer-Jakobsen, T., Bjerre, M., Äyväri, A., Lund, J.: Connect: Design for an Empathic Society. BIS Publishers, Amsterdam (2014)
3. Visser, T., Vastenburg, M.H., Keyson, D.V.: Designing to support social connectedness: the case of snowglobe. Int. J. Des. **5**(3), 129–142 (2011)
4. Fincham, F.D., Beach, S.R.H.: Of memes and marriage: toward a positive relationship science. J. Fam. Theory Rev. **2**(1), 4–24 (2010)
5. Campbell, B., Rosenberg, J., Schulzrinne, H., Huitema, C., Gurle, D.: Session initiation protocol extension for instant messaging. RFC 3428 (2002)
6. Boyd, D., Ellison, N.B.: Social network sites: definition, history, and scholarship. J. Comput. Med. Commun. **13**(1), 210–230 (2007)

7. Institute for Information Industry (III) (2015). http://www.iii.org.tw/Press/NewsDtl.aspx?nsp_sqno=1603&fm_sqno=14
8. Institute for Information Industry (III) (2016). http://www.iii.org.tw/Press/NewsDtl.aspx?nsp_sqno=1886&fm_sqno=14
9. Derks, D., Bos, A.E.R., Grumbkow, J.V.: Emoticons and social interaction on the Internet: the importance of social context. Comput. Hum. Behav. **23**(1), 842–849 (2007)
10. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press (1963)
11. Masadeh, M.A.: Focus group: reviews and practices. Int. J. Appl. Sci. Technol. **2**(10), 63–68 (2012)
12. Beyer, H.R., Holtzblatt, K.: Apprenticing with the customer. Commun. ACM **38**(5), 45–52 (1995)

Check for updates

Cognitive Offloading and the Extended Digital Self

Kristy A. Hamilton(✉) and Mike Z. Yao

University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
{kristyh2,mzyao}@illinois.edu

Abstract. Memory is more than remembering. Beyond cognitive processes that reside within the confines of internal memory storage exists several extramnemonic processes that produce behavior that we typically understand as within the realm of memory. The proliferation of immersive computing, pervasive computing, ubiquitous computing, and ambient intelligence has brought attention to a myriad of new questions related to the dynamics of memory under the influence of technology, particularly for the shifting interplay between internal and external memory. Control of memory is particularly important in consideration of the current and future potential for offloading aspects of memory onto systems in contexts of novel technology use. If an individual can accurately monitor memory performance, she can make key decisions that will service her intellectual and behavior goals, such as how to recall information and when to terminate practice. In this presentation, I will discuss cognitive offloading in the context of various media experiences. I will also discuss how the ability to rely on external search may increase performance, but may also potentially distort how we understand our own memory and knowledge.

Keywords: Cognitive offloading · Technology-mediated behavior
Memory and cognition

1 Introduction

The proliferation of immersive computing, pervasive computing, ubiquitous computing, and ambient intelligence has brought attention to a myriad of new questions related to the dynamics of memory under the influence of technology—particularly for the shifting interplay between internal and external memory [1–3]. Although humans have tried to offload memory tasks well before the advent of modern digital technology [4], the unique affordances introduced into our affairs through new technological environments has forced consideration of the ways in which technology-mediated cognitive operations differ from previous forms of cognitive operations, and how these differences impact past and future scholarship. The purpose of this paper is to explore the implications of cognitive offloading in our current technological environment. To accomplish this task, we will first characterize memory and memory control to accommodate the intricate role of human memory in the context of the global information infrastructure. Then, we will explore the ways in which strategically offloading information in particular contexts of technology use can potentially expand, constrain, and alter the functioning of memory

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 257–268, 2018.
https://doi.org/10.1007/978-3-319-91238-7_22

and cognition. We end with an example from our own empirical research that illustrates the influence of technology features on our perceptions of knowledge.

2 A Characterization of Memory and Memory Control

Beyond cognitive processes that reside within the confines of internal memory storage exists several extramnemonic processes that produce behavior that we typically understand as within the realm of memory. We store the phone numbers of our friends, classmates, and roommates on our smartphones, although we decide that it may be useful to store our partner's number internally. Yet, if someone asks us if we know our roommates number, we are naturally inclined to say that we do if the information is accessible to us. In such a perspective, memory is not only related to storage capacity, but also extramnemonic skills that exists in the larger cognitive context of servicing intellectual and behavioral goals.

Memory capacity reflects the dynamic ways in which individuals strategically select information for encoding to enhance memory performance, while reducing cognitive demands on memory [5]. Accordingly, the accurate assessment of one's own memory performance is a crucial ability to memory and memory control. If an individual can accurately monitor learning, she can make key decisions that will service her intellectual and behavior goals, such as how to recall information and when to terminate practice. If judgments of learning are inaccurate, the allocation of subsequent study time will suffer. Still, how people think about and monitor their own learning is highly imperfect [6]. Take as an example—while traveling to London by train, Alan Baddeley [7] noticed a familiar face, and decided to attempt to resolve how he knew the man. First, he recalled two associations related to the man—the name Sebastian and something with children. This led to several other relevant (and irrelevant) associations until he successfully recalled that the two were in the same baby-sitting group. Now, imagine the same scenario, but this time after seeing the familiar man on the train, Baddeley noticed that he was pushing a child in a stroller, which allowed him to immediately remember the baby-sitting group. Based on these two scenarios, Baddeley would likely feel more confident that he will be able to remember the man in the second scenario because he was able to identify him immediately. However, because this inference relies on contextual cues, we cannot be sure that it is diagnostic of future learning. In fact, the scenario in which Baddeley spent more time searching for the correct answer might lead to better long-term retention, despite having taken longer to retrieve. In the same way, because learners use monitoring to efficiently obtain their learning goals, better monitoring accuracy is expected to be related to more effective learning and higher levels of retention.

To illustrate this perspective, Benjamin (2007) considers the memory behaviors of lifeloggers—individuals who engage in comprehensive external encoding using technologies such as video recorders, computers, mobile devices, and so forth. There are several advantages to lifelogging. First, because storage capacity on a hard drive is practically limitless, decisions do not need to be made about what and how to encode. Second, because external digital sources are increasingly accessible, information is more easily "recalled" from memory stores. Third, because outsourcing information reduces cognitive demands on memory, time and resources are freed up for other activities [5]. In Alan Baddeley's case, outsourcing the question of how he knew the man on the train to a digital device may have been optimal given that the information is, and will continue to be, one Facebook search away. But, what about the information that must be committed to memory, such as information relevant to creativity, problem-solving, expertise, and other complex domains of knowledge? Without the skills in interacting with memory, how can the lifelogger accommodate these cognitive tasks? Here we must consider one critical advantage of *strategic* mental encoding—higher-order cognition guides memory behavior, but also memory guides higher-order cognition [5]. Therefore, the ability to strategically control memory access and flexibly use the outputs of memory processes to serve specific tasks may be as important as the ability to store information in a place that is accessible.

Control of memory is particularly important in consideration of the current and future potential for offloading aspects of memory in new technological environments. With the arrival of ubiquitous computing and ambient intelligence, our personal devices have become an essential component to our own memory and knowledge [1, 2]. In fact, many memory processes are now accomplished with the help of digital technology (e.g., remembering birthdays, findings directions). Offloading responsibility for information is optimal in many cases, such as when accuracy is paramount, or when offloading unneeded information may reduce interference of new information [5]. In spite of these positive impacts, such as expanding the capacity of human cognition and improving the efficiency of information searching, we should be cautious in assuming that all features of technology that reduce the cognitive effort of interaction and improve performance will necessarily benefit long-term retention and transfer of information [8]. This said, our ability to adaptively integrate internal with external processes, and our ability to monitor the decision to do so, represents a defining feature of what it means to be a successful cognitive agent in a complex environment.

3 Cognitive Offloading

When initially thinking about cognitive offloading, one might find it easier to think of familiar experiences connected to the term, such as storing important contacts on a phone, using a navigation app to finding directions, or archiving e-mails for later use. In its most basic sense, it is the idea that people can offload some of their cognitive functions onto technology, thereby extended the performance capacity of their human faculty [9]. Generally, cognitive offloading is understood as associated with common cognitive technology, such as computers and smartphones. At first glance, instances of offloading cognition onto cognitive technology may seem clear. The issue, however, becomes more evident once we consider how cognitive offloading may manifest in new media environments, and also media environments of the future. The arrival of wearable computers, the Internet of Things (IoT), and virtual and augmented realities are dramatically changing human-computer and computer-mediated interactions. Through improved modality, agency, interactivity, and navigability [10], these technologies offer users greater levels of "presence" and an illusion of non-mediation [11]. The need for having a physical and designated technological device for information retrieval, display and exchange is diminishing. Some tech analysts have gone so far as to predict the "death" of smartphones in the next ten years [12]. These new environments force us to reexamine fundamental constructs such as what technological affordances constitute as within the realm of cognitive offloading, and what are the subsequent consequences on human memory and perception. To address this issue, recent work has expanded the purview of cognitive offloading to include, in a very general sense, actions that offload cognitive demands onto-the-body and into-the world:

> We tilt our heads while trying to perceive ambiguous images, we gesture while imagining spatial transformations, and we rely on smartphones and search engines to store and retrieve information. In other words, we often think using our bodies and the external world [13].

In each of these examples, an action is performed in a way that accommodates an ongoing cognitive act so as to reduce cognitive demands on memory and cognition. In this sense, cognitive offloading encompasses actions that offload cognition onto-the-body (e.g., gestures, physical movement) and into-the-wild (e.g., writing things down, setting reminders). Therefore, we settle on a definition of cognitive offloading as, "the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand" [13]. Nonetheless, new media environments have expanded the range of actions people can take to reduce the burden on memory and understanding this expansion of possibilities requires attention to what it is new media spaces afford to people. The complex and systematic ways that media technology influence different possibilities for action by the user raises important questions related to the ways that cognitive offloading manifest in new media spaces, and their potential consequences on human perception and behavior.

4 Cognitive Offloading in a Complex Media Space

The proliferation of immersive computing and ambient intelligence has brought attention to the variety of experiences made possible by features and affordances of emerging technology, and potential outcomes on human perception and behavior. Although the same cognitive tendency may lead people to offload information in a variety of technology-driven contexts, the outcomes of offloading this information—and accessing it in the future—may differ according to features of the environment and context of use. These emergent technologies expand human activities beyond the realm of physical reality or even create entirely new human experiences. Technology-mediated interactions have gone from serial and codified message exchanges to fully immersive experiences enriched with social cues and machine intelligence; and, as the boundaries between the virtual and physical spaces blur, a technology-mediated environment has emerged. In the next section, we will discuss particular human-technology dynamics that may have benefits for cognitive offloading, and also consequences for accurate monitoring of learning. The approach of this discussion will be to characterize cognitive offloading in the context of various media experiences, and to discuss potential consequences of this behavior. This will be explored through consideration of three prominent technological environments— immersive virtual environments, ambient intelligent environments, and ubiquitous and pervasive computing.

4.1 Immersive (Virtual) Environments

Immersive virtual environments (IVEs) alter our perceptions of ourselves and our surroundings. It does this by replacing sensory information with technologically synthetic content to manufacture experiences that feel real, even though they are mediated [14]. Wickens (1992) mentions several notable features of IVE technology which may create a greater sense of presence, such as three-dimensional viewing, dynamic displays, and enhanced sensory information (among other features). These structural features of IVE technology (e.g., virtual reality), have potential to reduce the cognitive effort required to navigate through and interpret information in the system. For instance, data overload in visual and auditory domains pose challenges to operators in a wide range of workplaces, such as aviation, medicine, or process control [15]. This said, multimodal interfaces may facilitate strategic offloading by allowing information to be distributed across channels in a more task-appropriate manner [16, 17]. Although this could be an effective way of reducing the cognitive load on a taxed memory system, we should be careful to assume that all features of IVEs will improve performance. This is especially true for features of IVE technology that eliminate desirable difficulties in the environment that may be necessary to promote flexible outputs and transfer of knowledge [8]. For instance, evidence suggests that guiding trainees through the correct landing path using flight simulations can produce error-free performance in immersive environments, yet produces poorer transfer to landing skills once augmentation is removed [18]. These findings suggest that realism itself will not invariably improve memory.

4.2 Ambient Intelligent Environments

Ambient Intelligent (AmI) environments refer to technology-mediated spaces that are sensitive and responsive to our requirements and desires. AmI environments incorporate aspects of context-aware computing, disappearing computers, and pervasive/ubiquitous computing to proactively support people in their daily lives [19]. It is an inconspicuous technological environment that is perceptive to the particular characteristics of human behavior and is capable of reciprocating with an intelligent response. There is a clear advantage of offloading cognition in AmI environments— not only do AmI environment retain the wealth of knowledge that is typical of common cognitive technology (e.g., the internet), but also, the ability for these systems to adapt to users needs and desires certainly reduces the burden on memory, and in some sense, removes the decision to offload all together. Nonetheless, the question of how this may affect memory behavior requires consideration. In a speculative vein, this type of "embedded" offloading may have consequences that are conceptually similar to those described by the "Google Effect" [1]. The Google effect is a phenomenon first described by Sparrow, Liu, and Wegner (2011) as the tendency to forget information that is perceived as easily accessible through Internet search engine such as Google. In the original study on the Google effect, the authors demonstrated that participants who typed to-be-remembered information into a computer that they expected would save the information remembered less than individuals who typed to-be-remembered information into a computer and did not think it would be saved (memory was assessed in both cases without the memory aid) [1]. In line with these findings, Storm, Stone, & Benjamin (2017) noted that using the internet to access information makes people more likely to use the internet to access new information, and less likely to rely on their own memory [20]. This said, while the vision of interacting with smart objects every day offers a great range of fascinating extensions to human performance, seamless reliance on external processes to satisfy cognitive demands may have negative consequences for intellectual and behavior goals that depend on mechanisms of information aggregation that are (presently) singular to human memory systems.

4.3 Ubiquitous/Pervasive Computing

Ubiquitous and pervasive computing is a concept where computing is made to appear anytime and everywhere. In these environments, information is always present or reachable and almost always delivers the information that is desired by the user. In its ideal form, technology of this type move with us through the world to build pervasive, yet inconspicuous systems for offloading tasks. People cannot possibly know everything. Therefore, the benefits of having information at our fingertips are obvious. In fact, searching the web may be even faster than searching internal memory; whereas efforts to recall information internally can be time-consuming—and often fruitless—search engines return search results instantly, often even faster than these questions can be asked. This being said, an informed user should be able to take advantage of this boundless access to information by utilizing effective encoding strategies that maximizes performance on intellectual and behavior goals, while minimizing cognitive demand. However, the evidence on users' abilities to effectively control encoding strategies is mixed, and therefore the benefits of these systems are questionable. For example, Henkel (2014) examined whether the act of photographing objects influences what is remembered about them. On a tour of an art museum, participants viewed 30 objects—15 of which were photographed and 15 observed. Their findings reveal that participants remembered fewer objects and fewer details about the objects remembered if they had been photographed [21]. Although these findings highlight potential consequences associated with having technology available anytime and everywhere, the counterargument is clear—if technology is made available *anytime* and *everywhere*, what does it matter if information is not stored internally? Although externally-stored memory has the advantage of retaining information with reliable precision, it does not hold the capacity for self-organization. On the other hand, internal (human) memory systems have an exceptional capacity for self-organization and reorganization, which explains why creativity and expertise derive from well-organized internal memory systems and not digital memory [5].

5 An Example: Influence of Technological Ownership and Modality on Perceptions of Knowledge

As mentioned previously, the ability to adaptively integrate internal with external processes, and the ability to monitor the decision to do so, represents a defining feature of what it means to be a successful cognitive agent in our complex media environment. Still, how people think about and monitor their own learning is highly imperfect [6], and this lack of cognitive control is exacerbating by the unique actions afforded in particular human-technology dynamics. We now proceed to empirically explore our theoretical proposition through a case of one human-technology interaction: using technology to find answers to common declarative knowledge questions. These are the sorts of questions that we might come across in daily life. For example, you might be having a beer with friends after work when someone asks, "What do you think is the best-selling beer in the United States?" When this situation arises, you have one of two options—you can use your own internal knowledge to give your friend your best guess, or you can Google it. The results presented here come from a series of studies where we seek to determine how the act of searching for the answer internally versus outsourcing the query to Google influences your perceptions of your own knowledge, and also how particular features of the device you use to google the answer moderate the effect. For this discussion, we focus only on the moderators. For the full report, please contact the authors.

5.1 Research Questions

The purpose of this experiment was to test whether offloading cognition onto cognitive technology influences self-perceptions of knowledge differently when the cognitive technology is personally owned (versus not owned) and when information is accessed on a mobile device (versus a stationary device). Comparing different levels of familiarity with an external source and their relative influence on metacognition has been explored in various contexts. Similar manipulations have been studied in the context of close human dyads [22] and online access points (Google vs. Lycos) [23]. Thus, we directly examined the influence of ownership (*own device*: offloading cognition onto a personally owned device; *control device*: offloading cognition onto an unfamiliar lab device) on inflated cognitive evaluations. Accordingly, we predicted that retrieving answer to trivia questions from a personal device would result in inflated cognitive evaluations compared to retrieving answers from a control device. The modality of the device we use to access information may also carry cues relevant to knowledge judgments. Because mobile external digital sources are increasingly accessible, information is more easily "recalled" from memory stores. Thus, we also directly examined the influence of modality (*smartphone*: offloading cognition onto a smartphone; *laptop*: offloading cognition onto a laptop computer) on inflated cognitive evaluations. We predicted that retrieving answer to trivia questions from a smartphone would result in inflated cognitive evaluations compared to retrieving answers from a laptop.

5.2 Procedure

We aimed to recruit about 30 participants per condition (total 120 participants) based on the minimum suggested power (80%) used to detect differences between groups [24, 25]. The final sample contained 115 undergraduate students (94 female, $M_{age} = 19.63$) at University of Illinois at Urbana-Champaign participated in a between-subjects design. Participants were randomly assigned to one of four independent groups divided by two independent factors: ownership (own device vs. control device) and modality (smartphone vs. laptop). This said, participants were instructed to use either their own smartphone (n = 27), their own laptop (n = 29), a control (lab) smartphone (n = 33), or a control (lab) laptop (n = 26) to find all their answers to a ten-item trivia quiz. Trivia quiz items were selected based on pre-tested fairness ratings [23]. Responses from the pre-test indicate that participants found the items to be fair, but not particularly obvious (e.g., "What is the densest planet in our solar system?"). After completing the trivia quiz, participants completed the dependent measures and then were debriefed before leaving.

5.3 Dependent Measures

Response Accuracy. Responses to the ten-item trivia quiz were scored such that participants received one point for each correct response. Responses were counted as "correct" if they very closely or exactly match the correct answers (slight misspellings or conceptual matches will count as "correct"). The judgments were made by a research assistant blind to condition and experimental hypotheses.

Cognitive Evaluations. Immediately after completing the trivia quiz, participants completed the Cognitive Self-Esteem Scale (CSE) [23]. This 14-item scale measures participants' beliefs about their cognitive abilities. The CSE scale contains three sub-components that assess confidence in the ability to think (e.g., "I am smart"), remember (e.g., "I am proud of my memory"), and locate information (e.g., "I have a knack for tracking down information"). Responses were coded on a 7-point scale (1 = strongly disagree to 7 = strongly agree), such that higher ratings would indicate higher levels of CSE. The CSE scale demonstrated good reliability ($\alpha = .93$).

5.4 Results

Response Accuracy. We conducted a one-way ANOVA to evaluate the relationship between response accuracy and cognitive evaluations. Results revealed a significant effect of condition on response accuracy, $F(1, 114) = 5.03, p < .05$. Follow-up tests were conducted to evaluate pairwise differences among the means. Participants who used their own mobile device scored significantly lower on the trivia quiz ($M = 8.52$, $SD = 1.05$) compared to participants who used their own laptop ($M = 9.34$, $SD = .90$) and participants who used a control laptop ($M = 9.23$, $SD = .71$) (ps $< .05$). All other comparisons were not significant (ps $> .05$). Because the interpretation of this finding is ambiguous and extraneous to our investigation, we hesitate to draw conclusions on this result. We will, however, take note of this finding with respect to our main prediction regarding the effect of ownership and modality on cognitive evaluations.

Cognitive Evaluations. We conducted a two-way ANOVA to evaluate the effect of ownership (owned versus control) and modality (smartphone versus laptop) on CSE ratings. As predicted, results indicate a significant main effect of ownership, such that participants had higher overall CSE scores when they used their own device ($M = 5.21$ $SD = .74$) compared to participants who used a control device ($M = 4.93$, $SD = .90$) to complete the experiment, $F(1, 115) = 4.54$, $p < .05$, $\eta_p^2 = .039$. Also, results indicate a significant main effect of modality, such that participants had higher overall CSE scores when they used a smartphone ($M = 5.30$, $SD = .86$) compared to participants who use a laptop ($M = 4.82$, $SD = .74$) to complete the experiment, $F(1, 115) = 11.54$, $p < .05$, $\eta_p^2 = .094$ (see Fig. 1). The interaction effect of ownership and modality on CSE was not significant ($p > .05$). This implies that the particular features of a device used to offload cognition, such as ownership and modality, may influence cognitive evaluations. Furthermore, cognitive evaluations are not related to actual performance.

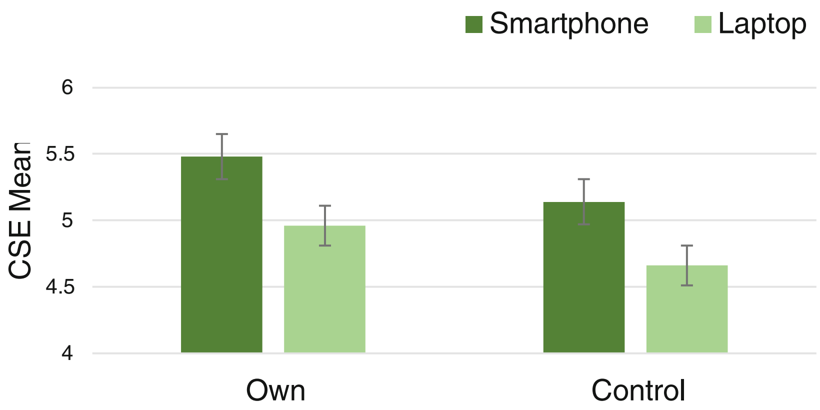


Fig. 1. Effect of ownership and modality manipulations on cognitive self-esteem (CSE) mean scores (N = 115). Error bars represent standard error.

5.5 Discussion

Media technology provide unique opportunities for offloading cognition in order to extend the capacity of our cognitive capabilities. Although the availability of these actions have wide-ranging implications, many of which are beneficial and valuable, they bring with it novel consequences. This example illustrates a particular instance of this dynamic human-technology interaction. Although participants correctly retrieved 9 out of 10 trivia questions on average, cognitive evaluations following the task reveal that features of a technological device used to offload cognition play an important role in influencing attributions of knowledge. More specifically, participants who used their own device to find answers reported inflated cognitive evaluations compared to participants who used a control device, and participants who used a smartphone reported inflated cognitive evaluations compared to participants who used a laptop. Our findings are consistent with the notion that individuals misattribute outcomes and characteristics of technology to the self while judging their own knowledge, which have potential consequences on strategic control of memory decisions, such as when to strategically encode information. For instance, a student who uses Google to study for an upcoming exam by "confirming" definitions he thinks he "mostly" understands may be surprised when he is not able to recall the information from memory during the exam. Likewise, a student who uses a navigation app to drive home to visit family may be caught off guard when they are unable to articulate the directions to a friend.

6 Closing Remarks

Taken together, the claims forwarded through this discussion are not intended to promote a technologically deterministic stance toward the positive or negative consequences of technology use. Instead, our discussion is meant to illuminate unique contexts of cognitive offloading made possible by a new, complex media space; and hopefully, lead our readers toward asking new questions—not about *whether* memory should be extended, but rather—about how to offer new answers to old questions *given* that memory has been extended. From the preceding discussion, we can see how and why the use of cognitive technology to facilitate information retrieval has become a pervasive habit of human behavior. Our growing digital memory repositories bring with them several potential advantages, such as when accuracy is crucial, or when the Internet is available and its use is contextually appropriate. Given that we will likely become increasingly connected to our digital memory, one may even argue that technological advancements will eventually obsolesce the need for strategic memory encoding. However, we should be cautious to assume that all features of technology that reduce cognitive demands, and even improve performance, will necessarily service our intellectual and behavior goals. After all, memory is more than just remembering [5]. So, a person immersed in a technologically rich environment may be more likely to remember a class essay deadline than a person who chooses to rely on memory alone. But, they may both have trouble deciding what to write about—both have "memories" inundated with trivial facts and details. Yet, the person who strategically encoded information is able to seamlessly navigate a self-organizing internal memory system to piece together the contents of the essay, the person who offloaded their knowledge in word documents, lecture images, and recordings is still sifting through their external "memory" for content relevant to the task at hand.

The broad assertion forwarded by this research is that new technological environments do not necessarily supplant human activity but rather changes it, often in unintended and unanticipated ways, and as a result poses new coordination demands on the user [26]. The unique affordances introduced into our affairs through interacting with new, complex media environments has forced consideration of the ways in which technology-mediated memory behavior deviate from previous memory practice, and how these changes beg reconsideration of antecedents and consequences of memory and memory control. This said, research is needed to explore the ways in which technology expands, constrains, and alters the functioning of memory and cognition in order to offer new answers to old questions, such as how to determine what information is most important to encode for some intellectual or behavioral goal, how to optimize retrieval practice to enhance that selective learning, and how to train people in this process.

References

1. Sparrow, B., Liu, J., Wegner, D.M.: Google effects on memory: cognitive consequences of having information at our fingertips. Science **333**, 776–778 (2011)
2. Nestojko, J.F., Finley, J.R., Roediger III, H.L.: Extending cognition to external agents. Psychol. Inquiry **24**(4), 321–325 (2013)
3. Storm, B.C., Stone, S.M.: Saving-enhanced memory: the benefits of saving on the learning and remembering of new information. Psychol. Sci. **26**(2), 182–188 (2015)
4. Yates, A.J.: Psychological deficit. Ann. Rev. Psychol. **17**(1), 111–144 (1966)
5. Benjamin, A.S.: Memory is more than just remembering: strategic control of encoding, accessing memory, and making decisions. Psychol. Learn. Motiv. **48**, 175–223 (2007)
6. Soderstrom, N.C., Bjork, R.A.: Learning versus performance: an integrative review. Perspect. Psychol. Sci. **10**(2), 176–199 (2015)
7. Baddeley, A.D.: Domains of recollection. Psychol. Rev. **89**(6), 708 (1982)
8. Wickens, C.D.: Virtual reality and education. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 842–847. IEEE, Chicago (1992)
9. Dror, I., Harnad, S.: Offloading Cognition onto Cognitive Technology. John Benjamins Publishing, Amsterdam (2008)
10. Sundar, S.S.: The MAIN model: a heuristic approach to understanding technology effects on credibility. In: Metzger, M., Flanagin, A. (eds.) Digital Media, Youth, and Credibility, pp. 73–100. MIT Press, Cambridge (2008)
11. Lombard, M., Ditton, T.: At the heart of it all: the concept of presence. J. Comput. Med. Commun. **3**(2), 107–133 (1997)
12. Weinberger, M.: Will the rise of AR mean the end for smartphones and TVs? World Economic Forum. https://www.weforum.org/. Accessed 21 Jan 2018
13. Risko, E.F., Gilbert, S.J.: Cognitive offloading. Trends Cognit. Sci. **20**(9), 676–688 (2016)
14. Bailey, J., Bailenson, J. N., Won, A. S., Flora, J., Armel, K. C.: Presence and memory: immersive virtual reality effects on cued recall. In: Proceedings of the International Society for Presence Research Annual Conference, Philadelphia, PA, pp. 24–26 (2012)
15. Sarter, N.: Multimodal displays: conceptual basis, design guidance, and research needs. In: Lee, J.D., Kirlik, A., Dainoff, M.J. (eds.) The Oxford Handbook of Cognitive Engineering. Oxford University Press (2013)
16. Sarter, N.B.: Multimodal information presentation in support of human-automation communication and coordination. In: Advances in Human Performance and Cognitive Engineering Research, pp. 13–36. JAI Press, New York (2002)
17. Sarter, N.B.: Multimodal information presentation: design guidance and research challenges. Int. J. Ind. Ergonomics **36**, 439–445 (2006)
18. Lintern, G., Roscoe, S.N.: Adaptive perceptual motor training. In: Roscoe, S.N. (ed.) Aviation Psychology, pp. 239–250. Iowa State University Press, Ames (1980)
19. Cook, D.J., Augusto, J.C., Jakkula, V.R.: Ambient intelligence: technologies, applications, and opportunities. Pervasive Mob. Comput. **5**(4), 277–298 (2009)

20. Storm, B.C., Stone, S.M., Benjamin, A.S.: Using the Internet to access information inflates future use of the Internet to access other information. Memory **25**(6), 717–723 (2017)
21. Henkel, L.A.: Point-and-shoot memories: the influence of taking photos on memory for a museum tour. Psychol. Sci. **25**(2), 396–402 (2014)
22. Wegner, D.M., Giuliano, T., Hertel, P.T.: Cognitive interdependence in close relationships. In: Ickes, W. (ed.) Compatible and Incompatible Relationships. Springer Series in Social Psychology, pp. 253–276. Springer, New York (1985). https://doi.org/10.1007/978-1-4612-5044-9_12
23. Ward, A.F.: One with the Cloud: Why people mistake the Internet's knowledge for their own. Doctoral Dissertation. Harvard University, MA (2013)
24. Cohen, J.: A power primer. Psychol. Bull. **112**(1), 155 (1992)
25. VanVoorhis, C.W., Morgan, B.L.: Understanding power and rules of thumb for determining sample sizes. Tutor. Quant. Methods Psychol. **3**(2), 43–50 (2007)
26. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **30**(3), 286–297 (2000)

Deconaby: Animations for Improving Understandability of Web Images

Kazuyuki Hikawa(✉), Kohei Matsuda(✉), and Satoshi Nakamura(✉)

Meiji University, Nakano, Tokyo, Japan
kazuyuki.hikawa@gmail.com, matsuda0201@gmail.com, satoshi@snakamura.org

Abstract. There are many "how-to" pages on the Web. However, on such pages, descriptions and images are often not clearly linked. As a result, it is difficult for viewers to understand which parts of the text description refer to which pars of the images. This study aims to solve this problem. With the developed Deconaby system, users can add animations just by clicking on an image on a Web page and moving their mouse. Our system then generates a URL that consists of the how-to page's original URL and a hash ID for sharing the animation-added version with others. In this manner, our system can improve the understandability of Web pages with little time and effort. We conducted an experiment to test the effectiveness of our system. The results showed that our system succeeded in increasing the understandability of difficult-to-understand pages. We also showed that a page written in an unfamiliar language can be understood with the help of appropriate Deconaby animations.

Keywords: Web · Image · Animation · Mouse operation · Support · Share

1 Introduction

Many "how-to" pages are posted on the Web. These pages describe, for example, how to install PC applications (see Fig. 1), access restaurants from their nearest station, fill out and submit application forms, and draw illustrations. When people want to know something, they usually use a search engine such as Google and Bing to obtain information from these how-to pages.

To make their how-to Web pages easier to understand, authors often use images such as screenshots, maps, photographs, and illustrations. However, the descriptions and the images in how-to pages are often not clearly linked. As a result, readers cannot easily understand what part of a text description refers to what part of an image (see Fig. 2). In addition, these how-to pages sometimes require expertise in fields such as programming, software development, and foreign languages. Japanese students sometimes give up trying to understand a page if it is written in English. Likewise, elderly people cannot easily understand a page written with many technical terms (see Fig. 3).

On a different note, people cannot help but gaze at something that is moving. For example, presenters often use a laser pointer to direct the audience's attention to a point of note on the screen when giving a presentation using PowerPoint, Keynote, and other

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 269–282, 2018.
https://doi.org/10.1007/978-3-319-91238-7_23

software applications. In addition, creators of instructional videos about application installation, programming, and similar fields aid their viewers' understanding by using the mouse cursor to indicate specific parts of content. Thus, if we add movement to how-to pages, we might improve those pages' understandability.

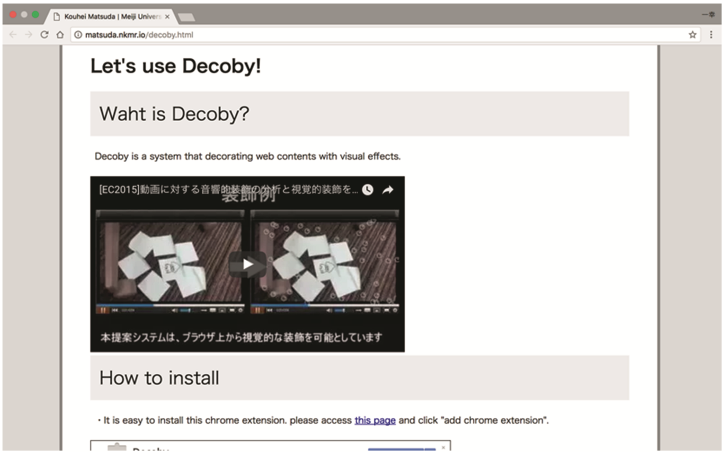


Fig. 1. Example of a "how-to" page.

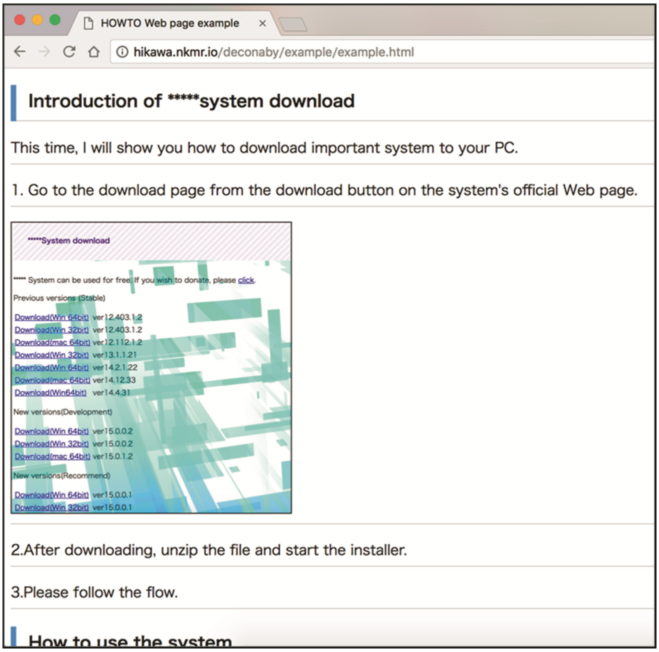


Fig. 2. Example of an incomprehensible image on a web page due to several overlapping descriptions such as repeated use of the term "download."

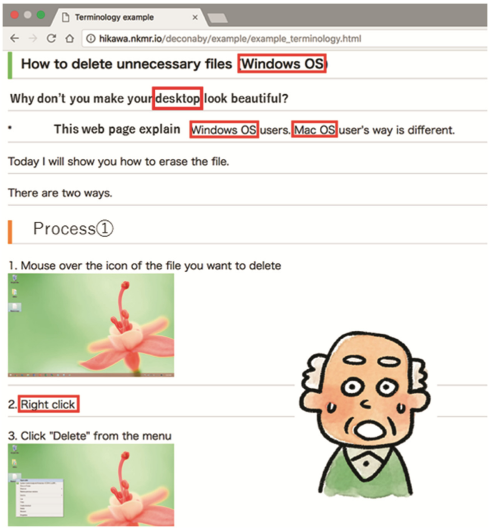


Fig. 3. Example of how technical terms (highlighted by red rectangles) confuse elderly persons. (Color figure online)

Aiming to solve the problems described above, we propose a system called "Deconaby" that enables a user to add animations directly to images on Web pages and to easily share the pages with the added animations with others. A mouse-cursor, pointing finger, or walking person animation can be added to images. The user can improve others' understanding of Web pages by using the proposed system with little time and effort.

By using Deconaby, a user can use original content without infringing on its copyright as well as create and browse originally animated pages tailored to their needs. We expect that Deconaby will enhance the understanding of users unfamiliar with Website operations such as clicks and swipes, and prevent users from suffering unnecessary stress and confusion.

First, we describe related works and propose the Deconaby system. Then, we describe its implementation and show its effectiveness by conducting an experimental test. Finally, we describe the application of our system and discuss future work.

2 Related Work

Methods for modifying the content of pages on the Web have been developed extensively. For example, the Web browser extension Decoby enables users to add effects to images and movies posted on the Web and share those effects with others [1]. By using Decoby, users can enjoy content in their preferred form. In addition, Nakamura et al. proposed a system that can aurally decorate movies on the Web [2]. This system not only enables users to make decorations of their choice for videos but also creates a new movie-viewing experience. These systems have the advantage of enhancing the user experience by modifying the content on the Web. However, they do not increase the understandability of Web content.

WATSS (Web annotation tool for surveillance scenarios) is a Web-based annotation tool that allows users to annotate video content taken by a surveillance camera [3]. Web Annotator enables users to annotate e-learning materials [4]. InkAnnotation enables users to use styluses to add handwritten annotations to teaching materials [5]. Another Web-based annotation tool enables users to emphasize parts of teaching materials by highlighting them [6]. These tools enable annotations to be made on actual notes used in e-learning. Virtual Notes can embed annotations in the text; that is, annotations are inserted as “sticky notes” with a title and make it easier for users to understand the text [7]. These studies propose appropriate methods for the purpose of annotation; our research differs from them because its purpose is supporting user understanding.

A publication system that enables communication such as chatting between users browsing the same Web page has been proposed [8]. It allows users to exchange opinions and information. Another proposed system shows annotations made by support staff on users’ screens in real time as a help system for Web applications [9]. However, these systems differ from Deconaby, which enables asynchronous sharing irrespective of timing between users. In addition, a graphic annotation platform that provides annotated lines and figures on screenshots of Web pages and stores them in a database for sharing has also been proposed [10]. These annotation systems allow text and graphic annotations. In contrast, Deconaby does not insert text or graphics; instead, it supports the user by adding animations of icons such as a mouse cursor and a pointing finger.

3 Deconaby

3.1 Proposed Method

One way to increase the understandability of an image on a Web page is to highlight a part of the image by drawing a red rectangle around that part or by pointing to it with a red arrow. However, preparing such image content requires much time and effort. In addition, if a Web creator wants to highlight several parts of an image in order to indicate the order in which a user should click several buttons or indicate the route from a station to a destination, the image content might become incomprehensible because of too many highlighted parts. In addition, average users cannot fix these incomprehensible images themselves.

To solve this problem, we propose a method that enables average users to add animation to image content on the Web as a decoration. This method emulates actions done by users with superimposing animations, such as a mouse cursor, pointing a finger and moving person, on the image content (see Fig. 4). These animations enhance the user’s understanding without modifying the original content of the how-to pages.

For example, when Web creators want to make a page describing how to install an application on a PC, they usually take a screenshot of the page distributing the application and embed the screenshot in their web page. Then, they explain in the text the parts of the image that can be clicked on. When a reader visits the Web page to find out how to install the application, they sometimes cannot easily understand which parts they should click. If guidance with an animation of, for example, a cursor clicking on the relevant part of the screenshot could be given directly, the user would not have to check the text to determine where to click.

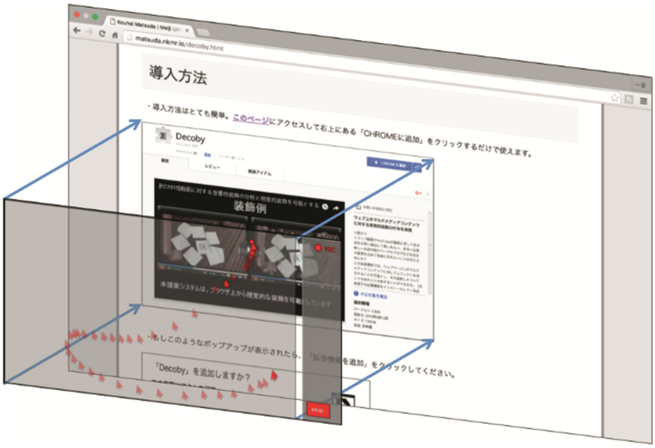


Fig. 4. Overview of Deconaby. (Color figure online)

As another example, when Web creators want to create a page explaining how to access a facility from a train station, they usually upload an image of an appropriate map and describe the location of the facility, the nearest station, routes of access, and so forth in the text. However, if the map is too complicated, readers who visit the page to determine how to access the facility might get lost when they actually go. If it were possible to present a suitable route to the user on such a map using an animation, users would be able to find their way easily without viewing the page for a long time.

An image showing how an animation is assigned on a Web page by using Deconaby is shown in Fig. 4. An additional canvas on the target image is created automatically, and an animation is drawn on the canvas. The animation can be added by moving the cursor and clicking on this canvas.

Deconaby allows users to easily share the animations assigned to images on Web pages with others. Specifically, it saves the animations that a user adds to a page and issues a new URL. The new URL consists of the original URL and a hash ID generated automatically for each decorated version. A user can share this decorated version only by sending this new URL to other users, who can access the animation-decorated Web page by using Deconaby and clicking this new URL.

3.2 Implementation

We implemented the Deconaby system as a Google Chrome browser extension using JavaScript and its libraries jQuery and P5.js[1], and we implemented its server system

[1] https://p5js.org/.

using PHP and MySQL for storing animation information and sharing animations assigned by users.

When a user uses Deconaby to visit a Web page, all images on the page are listed, except for small images such as buttons, icons, and advertisements, which are excluded because they are unsuitable for adding animations. Next, transparent additional layers (canvases) are generated by P5.js (as many as the number of elements of the acquired IMG tag of HTML) according to the size of each image. These layers are overlapped according to each image's position on the page.

When a user wants to add a mouse-cursor animation to an image on the page, they simply click the part of the image that they want to animate. At that time, if the image is designated as a link by the A tag of HTML, the Web browser judges that the user wants to access the linked image. In the animation mode wherein users animations, our system disables its link, as described below.

When the user assigns an animation to the page, the animation information is sent to the recording server (in JSON format) so that it can be shared with other users. In the recording server, the received data is directly collected in a database. However, it is possible that one page may be decorated with animations from multiple users. Therefore, the page to which the animation is assigned is linked to the animation information by attaching an ID to the original URL. The animation ID is given at the end of the URL as a hash value (i.e. the value after the "#" of a URL). IDs are automatically generated so that data do not overlap. For example, when a user decorates a page by adding an animation to it, a URL ("http://example.com/example.html") and an ID (12345) are generated, and a new URL ("http://example.com/example.html#decoID=12345") where other users can see the animation-decorated page is provided.

When this URL is opened with Deconaby in a Web browser, the "decoID" is detected from the URL, and animation information is requested by Ajax with ID "12345" from the recording server. The recording server retrieves the animation information matching the ID from the database and sends the information to Deconaby, which reproduces the animation on the relevant image of the page. If the user does not have Deconaby installed, the Web browser ignores the values after the "#" of the URL and shows the original page.

3.3 Animation Types

When adding animation to various how-to-type pages, users might want to use several kinds of animations. For example, how-to pages about PCs and smartphones have different UIs for each application, so there is likely to be significant demand for explanations of the different interfaces. To meet that demand, appropriate mouse-cursor and finger animations were prepared for those pages. In addition, a footprint animation (which can be used to express a trajectory or order of movement that is difficult to explain verbally) was also prepared.

The mouse-cursor animation (see Fig. 5) supports explanations of PC operations and can signify a mouse click with a pulsing circle. Users can indicate the order of clicking by recording the points that they click by clicking multiple times. The finger animation (see Fig. 6) supports explanations of smartphone operations. In this mode, a finger image moves along the trajectory recorded by mouse clicks, and an animation illustrating a tap on the smartphone's touchscreen can be assigned to the image.

Fig. 5. Mouse-cursor animation.

Fig. 6. Finger animation.

The footprint animation (see Fig. 7) explains routes on a map. In this mode, a person-like icon animates the trajectory recorded by mouse clicks while leaving a trail of footprints. When the animation is played, the person-like icon stops for a while at the clicked location, so the user can understand that they are being guided to that location.

Fig. 7. Footprint animation.

By appropriately using these three types of animation, users can add their intended commentary to the image.

3.4 Operation

When a user installs Deconaby, a mode-change button is displayed at the lower right of their browser. Clicking this button switches between animation mode and decoration mode. Animation mode only shows the assigned animations. Decoration mode enables the users to add animations to certain images on a Web page. Deconaby's default setting is animation mode. When users want to add an animation to an image on a web page, they turn on the decoration function by clicking the mode-change button.

In decoration mode, a decoration menu is displayed when a user moves their cursor over an image on the Web page (see Fig. 8). The user can change the animation type by clicking the appropriate button on the decoration menu. If the user clicks on the image, an animation is created and recorded by monitoring their mouse movement. During the recording, an afterimage of the locus of the mouse is displayed. When the image is clicked on, the animation starts recording at the clicked point. The user can indicate the order of the clicks by clicking two or more times without stopping recording of the animation. When the user clicks the stop button in the decoration menu, the recording stops. The trajectory on the image from the last clicked position to the stop-button click point is not recorded because the mouse movement after the last click is unnecessary information. If the user does not click on the image, it is judged that the user does not intend to decorate the image, so the locus is not recorded.

Fig. 8. Decoration menu for a mouse-cursor (top) and animation recording (bottom). (Color figure online)

After assigning animations to all selected images on the page, the animations can be saved in the server by clicking the “Issue” button in the menu, and a page with an ID in the URL is opened in another tab of the browser. The user can use this URL to share the animation-added Web page with others.

3.5 Additional Functions

Deconaby is implemented on and distributed through our Web site (http://deconaby.club/). On the basis of feedback from users, the following two additional functions were added.

- **ID list display:** Usually, users cannot view decorated pages without knowing the animation ID. However, they sometimes want to view animations when they cannot understand the page. The ID list display allows users to browse animations added by other users by selecting a decoration ID from the list. Also, if another user adds an animation to the page that the user is viewing, the color of the menu changes to green. As a result, the user can know whether an animation has been added to that page.
- **Request:** If a user cannot understand a page, and the page has not been improved with Deconaby, the user can send a request asking others to improve the page. In this manner, it is possible to get many users to add animations to the image to increase its understandability.

4 Experiment

4.1 Pre-experimental Analysis

Deconaby was evaluated with an experiment using 14 university students (aged 20 to 23) as subjects. The experiment used 22 how-to Web pages, namely pages explaining the operations of PCs and smartphones or giving directions to a destination (10 pages for PC operation, 5 for smartphone operation, and 7 for route directions). Each subject was asked to use Deconaby to explain a page or improve its understandability for their families.

Examining the animated pages made by the subjects revealed that they tended to click on the same position on the page concerning how to operate applications or settings. In addition, most of the things described in the text on the page were reproduced with animation.

The subjects sometimes used animation to emphasize multiple positions in an image in the order of operation. Sometimes, when text explanations were omitted because the operation was basic, there were still decorations to supplement the instructions. In other words, the animations can be classified into two types: one type reproduces the content written in the text as animations; the other supplements insufficient explanations on the page.

The assigned animations sometimes reflected the users' situations. For example, several operating systems and several versions of Windows OS are in use around the world. Accordingly, some applications can run on several different operating systems and versions. Pages that distribute these applications list several links to download for each operating system and each version. In our experiment, the subjects chose the indicated download link depending on the operating system and versions running on their families' computers. As for when the subjects animated a page that gives directions from a station to a building, half of them used animation to indicate the shortest path, and the other half indicated the least crowded path.

Some animations added by the subjects encircled parts of an image, and these animations traced the designated part with the finger animation. These animations were added to emphasize certain parts of the image. In addition, some animations pointed to a station or destination with the finger animation, rather than the footprint animation, on the map. These explanatory animations were unexpected.

According to the feedback from the subjects, many subjects felt that using Deconaby made it easy to understand pages visually by animating them and that it was handy and easy to use. These impressions demonstrate that Deconaby has utility and usability. However, some subjects expressed the opinion that they would like to signify certain operations, such as double-clicking and dragging-and-dropping, as animations; animations for these operations are not implemented in Deconaby. It would indeed be useful if users could use such animations. However, other subjects expressed the opinion that Deconaby is handy because it is simply operated by mouse-clicking only. Therefore, when increasing the functionality of Deconaby, we must take care not to overcomplicate its operation.

As for other opinions on the functionality of Deconaby, some subjects felt that it would be helpful to be able to add animations to explain where to click on an entire Web page. Currently, Deconaby only targets images on Web pages for animation because many how-to pages contain difficult-to-understand images. If the entire page is targeted, Deconaby would need to consider the window size of the browser and the position of content depending on its size. However, we aim to develop Deconaby further so that it can be used to decorate an entire Web page and be more usable for giving general commentary with animations.

4.2 Analysis of Results of Evaluation Experiment

The extent to which Deconaby improved the understandability of how-to pages by assigning animations was experimentally tested as follows. One hundred pages were prepared. Among them, fifty were considered difficult to understand, so Deconaby was used to add animations to each of those 50 difficult-to-understand pages. The understandability was evaluated on five levels (-2 to $+2$) before and after the animations were added. A total of 50 evaluations were made by each subject, evaluating pages that were seen for the first time; specifically, each subject evaluated 25 unanimated pages and 25 animated pages.

The numbers of pages sorted into each how-to category used in the experiment are listed in Table 1. In the 50 pages examined, the average understandability score indicated increased easiness-to-understand for 32 pages and decreased easiness for 11 pages. The statistical significance of the difference was verified using a sign test ($p < .01$). In addition, the evaluation score of 14 of the pages was negative before decoration. This result means that the original pages were highly difficult to understand. Of these pages, the average score for 13 of the 14 increased (Table 2).

Table 1. Average understandability score of how-to pages classified by category.

Category	Decrease	±0	Increase	14 highly difficult pages
PC	2	5	13	8
Smartphone	2	3	7	1
Map	2	0	3	2
Photo	3	0	1	0
Other	1	1	7	3

Table 2. Average understandability score for each highly difficult-to-understand page.

Page number	Animation	
	Without	With
4	−1.2	−0.4
9	−0.25	0.6
10	−0.4	0
21	−0.8	−0.2
23	−0.8	−0.2
25	−0.6	0.8
28	−0.6	−0.6
33	−0.6	0.6
38	−0.4	0
41	−0.8	−0.4
44	−1	0.2
46	−0.8	−0.6
48	−1.4	−0.4
49	−1	1.6

As shown in Table 2, although there are seven pages in the smartphone category for which the understandability score increased, only one page was originally particularly difficult to understand. All of these pages were easy to understand by indicating the order of operations, and because there were many understandability values that increased, it was a case that demonstrated the advantage of using Deconaby to show order. Considering these categories, the mouse-cursor and finger animations are necessary, but the footprint animation needs to be further studied. The footprint animation is a mode that can indicate a locus that cannot be indicated with other modes. From now on, it will be necessary to further verify the usefulness of having an animation that shows tracks and to implement other optimal animation.

5 Application

Even if a page is written in a language that is unfamiliar to the viewer, it is possible to understand the page by assigning appropriate animations with Deconaby. Even if animation cannot facilitate perfect understanding, it is at least helpful in improving understanding. In addition, we believe that Deconaby is useful for enhancing e-learning content. Deconaby can potentially be used to support learning by animating, for example, the order in which characters are written and of flow charts. For example, many pages explain algorithms for problems such as searching a binary search tree, solving a shortest-path problem, and giving the best route to solve the Seven Bridges of Konigsberg problem. Deconaby could also be used to indicate the stroke order of Chinese characters and to show the best route to escape a labyrinth.

The current prototype of Deconaby only handles simple patterns with limited animations, namely, a mouse cursor, footprints, and fingers. It is thought that it is effective to show images of food ingredients on a page explaining how to cook a meal and to use an animation of a kitchen knife to show how to cut certain ingredients properly. Also, for pages explaining how to assemble things such as tents, plastic models, and origami, animation suitable for the content can be added. The addition of such animation to Deconaby will also be studied in the future.

Animation can be added to an image with an attached URL. We can use Gyazo, which shares screen captures with URLs, and users can add animation to the screen captures they take [11]. Gyazo has a GIF-shooting function. However, Gyazo's GIF-shooting function has a time limit, so it cannot capture a long operation. It is thought that Deconaby has advantages not found in the GIF-shooting function.

With Deconaby, animation can be added only to still images, but the animation for moving images can be added with continuous animations such as those created by Decoby [1]. For example, it is possible to add commentary animation too, for instance, a person in a sports movie. One could also use animation to a dance movie to instruct viewers on how to move their foot or hand. One could also use an animation of a finger with a movie of someone playing music to instruct viewers on how to play.

6 Summary and Future Work

We proposed and implemented a system called Deconaby for assigning explanatory animations to the images of how-to pages to make them easier to understand visually and for sharing those animations with others. The effectiveness of Deconaby was tested by conducting an experiment.

Three animation modes—namely, a mouse cursor, fingers, and footprints—were implemented in Deconaby. However, they are not enough for users to improve all of the vast types of how-to pages on the Web. In future work, we plan to consider other types of animation. In the prototype system, the user has to install our extension to Google Chrome. Users who do not install Deconaby cannot see the animations. Accordingly, we are planning to create a Web service that allows users to see the animation without installing Deconaby.

Acknowledgements. This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

References

1. Matsuda, K., Nakamura, S.: A method for decorating web contents with visual effects. In: HCG Symposium, Toyama, pp. 1–7 (2015)
2. Nakamura, S., Ishikawa, N., Watanabe, K.: A decorating and sharing method to change the enjoyment of video clip on the web. In: WISS 2013, Kochi, pp. 1–6 (2013)
3. Federico, B., Lorenzo, S., Giuseppe, L., Svebor, K., Alberto, D.: WATSS: a web annotation tool for surveillance scenarios. In: Proceeding of the 23rd ACM International Conference on Multimedia, MM 2015, Brisbane, pp. 701–704 (2015)
4. Dale, R., Sam, J.: Web annotator. In: Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education, SIGCSE 2003, Reno, pp. 386–390 (2003)
5. Silva, A.C.: Ink annotation: an annotation tool for e-learning environments. In: International Conference on e-Learning, e-Business, Enterprise Information System, and e-Governance, pp. 73–74 (2015)
6. Pereira Nunes, B., Kawase, R., Dietze, S., Bernardino de Campos, G.H., Nejdl, W.: Annotation tool for enhancing e-learning courses. In: Popescu, E., Li, Q., Klamma, R., Leung, H., Specht, M. (eds.) ICWL 2012. LNCS, vol. 7558, pp. 51–60. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33642-3_6
7. Geyer-Schulz, A., Koch, S., Schneider, G.: Virtual notes: annotation on the WWW for learning environments. In: Americas Conference on Information Systems, pp. 136–138. Milwaukee (1999)
8. Samuel, S., Ivanoff, A.: Enabling communication between users surfing the same Web page. Patent Application Publication US 2006/0026233 A1, 1–6 (2006)
9. Karam, M., Ibrahim, M.: Synchronous online help support with visual instruction aids for workflow-based MVC web applications. In: Proceeding of the 27th ACM International Conference on Design of Communication, SIGDOC 2009, Bloomington, pp. 105–114 (2009)
10. Giordano, D., Mineo, S.: A graphical annotation platform web-based e-learning. In: 3rd International Conference on Multimedia and Information and Communication Technologies in Education, pp. 1–5 (2005)
11. Gyazo. https://gyazo.com/. Accessed 8 Feb 2018

Check for updates

Study on Children's Toy Design Based on Perceptual Evaluation

Bin Jiang[1], Haoran Ni[1(✉)], and Yingying Miao[2]

[1] School of Design Arts and Media, Nanjing University of Science and Technology, 200, Xiaolingwei Street, Nanjing 210094, Jiangsu, China
631603555@qq.com, 304708122@qq.com

[2] Engineering Training Center, Nanjing University of Science and Technology, 200, Xiaolingwei Street, Nanjing 210094, Jiangsu, China
398991222@qq.com

Abstract. With the development of society, the demand of consumers shifts from material demand to emotional demand. The consumer - centered designer is asking for more. Perceptual technology, is a kind of "consumers' perceptual translated into product design elements of technology", which aims to explore some of the emotional reaction, for the product to convert these reactions to design products and consideration should be paid attention to by the conditions. Perceptual evaluation can help designers to grasp the emotional needs of consumers scientifically and accurately extract design elements. In view of the importance of children in family, it is particularly important to study the emotional needs of children's toy consumers. Methods by using the method of psychological measurement, semantic research for consumers, by interview or questionnaire to the respondents information such as language, actions, expressions, gestures were analyzed, and the master consumers emotional appeal for children's toys. Through the material image experiment, the perceptual evaluation of the material is studied, and the design language is acquired. Conclusion the emotional demand of children's toy consumers is analyzed, and the emotional tendency is quantified, and the excellent design language is extracted accurately. It provides scientific theoretical support for the product design of children's toys, and provides reference for the selection of color, material and processing technology of the products.

Keywords: Emotional evaluation · Children's toys · Emotional design
The semantic method

Preface: with the development of the society, the progress of science and technology, make people's material needs are met, people no longer just for survival, and gradually began to the pursuit of life, high level, high quality, enjoy life. This puts forward higher requirements for the designers who focus on users and consumers. The design of products no longer stops at the requirements of function and safety, but tends to satisfy people's emotional needs and enhance the added value of products. As the perceptual technology, is put forward by the Japanese city of Hiroshima university a "consumer's perceptual translated into product design elements of technology", which aims to explore some of the emotional reaction, for the product to convert these reactions to design products and consideration should be paid attention to by the conditions. Perceptual

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 283–294, 2018.
https://doi.org/10.1007/978-3-319-91238-7_24

design provide us with a product quality and functionality are going to promote the design method of the competitive advantage, at the same time also can make designer constantly in the design, pay more attention to the emotional needs of the people. Children, as the new force of human society and the hope of the future, are widely concerned, and the industry related to children is worth exploring and redesigning. Children's toys are of great significance to children's intelligence exploration and communication, which can arouse children's enthusiasm, enhance perceptual knowledge, and help children to think and imagine actively. For this reason, this article will research related to children's toys perceptual engineering theory and analysis in the design of children toy should be considered when the problem such as colour, material qualitative, craft, so as to provide new study idea for the design of children toy and practical way.

1 Current Research on Children's Toys

From the perspective of toy development, science and technology level, it can reflect the each stage of history can reflect different countries, different nationalities, different education concepts, customs, different religious beliefs, aesthetic temperament and interest. Like other civilizations, toys have evolved over time. Toys for children is more important, is not only indispensable to accompany them grow "mentor", at the same time also cultivate their intelligence and promote the healthy growth of children learning and intelligence auxiliary tool. As the children grew up, lu xun once said, "games are children's work, toys are the angels of children." "To pay attention to the growth of young children, we must pay attention to children's games"; Attention to children's game, it is necessary to focus on use of toys, children can see toys in developing plays an important role because of toy contain the social or the culture of The Times, customs and values, any toys are all in the broadest sense of the term "education"; And from a narrow point of view, the education function of some toys in the toy design, can promote children's learning a new skill or size, shape, color, position, type, number, classification, comparison and calculation, such as concept and training, etc. Can be seen that the toy in preschool children, especially the role of education should not be underestimated, need to be taken seriously, at present, although there are many different kinds of toys, toy market in our country the number of dazzling, but the author in writing this thesis, through the access network and literature, etc., found at home and abroad about scarcity, toy toy theoretical research is relatively weak. It can be seen that the role of toys in children, especially preschoolers, has not yet attracted enough attention. With the rapid development of China's economic construction, the family living standard is also improving. As a result, parents in today's society are more than happy to buy toys for their children, and the annual expenditure of toys also shows a clear trend of growth. But how do parents buy toys? What kind of toys? What are the effects of toys on children? And so on. [1, 2] For suppliers, sometimes too much emphasis on toys commercial, considering the manufacturer's profits and earnings, and neglect the function of the toy, therefore, the design and production of toys too much emphasis on entertainment functions.

At present, there are some scholars involved in the research of toy design in China. An early childhood educator, once put forward such a view on the design of toys. He believes that "manufacturing materials should be firm and not damaged"; "interesting shapes can stimulate children's interest"; "can wash without fading, modeling is not ugly, to meet the child's aesthetic needs"; "the structure of the toy needs to be changed, can be moved The children can play on their own and can open up; adapting to the child's ability. It can be seen that the toy designed by Mr. Chen is from the material, shape, color, structure and aesthetic requirements of the toy. The articles of the design principle of preschool children's toys, she puts forward five aspects of the design of toy design principles: Security, namely the toy design should consider when playing with toys by children, will not be any harm; The toy design should be concise and easy to use, which conforms to the operation mode and usage habits of children. Toy design should be interesting, because interesting toys can arouse children's attention and arouse their interest, and cultivate their exploration spirit, imagination, creativity and practical ability; Fourthly, it should be helpful to stimulate the brain development of children and promote the development of their thinking and cognitive abilities. 5. Interactivity, that is, when children play with toys, they can also play different roles to make their social and other aspects develop. Horse town Lin the articles of the four concepts into toy design, toy design puts forward four Suggestions: one, to highlight the theme, namely toy design should comply with the corresponding age children's physical and mental characteristics, toy design fits their needs; Ii. Functional integration, that is, toy design should consider having more functions, such as disassembling and assembling, etc.; Third, the innovation, that is, toy design should also keep pace with The Times, constantly add new elements of the society, let it give the flavor of The Times; Focus on serialization.

2 The Main Function of Children's Toys

Toys, as the name implies, are used to play. Therefore, any kind of toy, entertainment and games should be the most basic characteristics of toys. Apart from the characteristics of entertainment and games, the education effect of toys should not be overlooked. The world health organization (who) led the organization of children's growth and development cooperation center under 6 years old children's psychological development level and family situation, according to a survey of the toys to children's physical activity, language, cognitive and social development has a role in promoting. The author believes that the role of toys in preschool children is mainly reflected in the following aspects.

2.1 Help Children Learn

Our country some scholars pointed out that: good toys to children bring happiness, not only can they provide a rich perceptual experience stimulation and social culture, helping young children to understand and master the usage of some articles for daily use. It can be seen that toys can promote children's learning [1]. One of the early adopters of the learning process was the British thinker John Locke. He transformed the blocks into building blocks that could help children learn the alphabet, and tried to use them at home to learn. Some toys are designed to promote children's learning of certain skills or concepts. If thimble plastic toys, not only in a variety of colors: red, yellow, blue, green, violet wait for color, common but also vary the shape of a ring: round, square, triangle, heart shape, rectangle, etc., children can combine at will assembling. In the process of playing, I not only learn to distinguish shapes and colors, but also learn to combine orders according to certain rules, from easy to difficult, and train children's logical thinking ability. Such as all kinds of makeup, also on every board with different Numbers, letters or animals, not only let the children of Numbers, letters, and learn the cognition to the animal, can also through different combination of simple addition and subtraction and learning of English words.

2.2 Help Children Get More Experience

Former Soviet educator sukhomlinski once said that the development of children's intelligence is reflected in the fingertips. Toys with there are bright colors, beautiful shapes, pleasant sounds, different materials, etc. [3] Toys can give children a variety of feelings. The excitement of an official children gain many different experiences by looking, listening, touching and sniffing toys. The development of ability has good promoting effect.

2.3 Help Children Learn to Explore

For preschoolers, to play with toys, they must first observe, study and play with toys. Some toys have some difficulty in playing, which requires children to use observation, imagination and memory at the same time. Psychology studies suggest that preschool children observation has no purpose, also is not very stable, easily affected by other things, to the training on preschool children in observation, will find a first can cause they observe the stimulation of interest. [1] Toys because of its beautiful shape, bright color is easy to attract the attention of young children, plus the artists also like hidden in the toy "mystery", children need to explore to find by my own observation, a variety of fun so that they, in their observation of persistent training there is a huge role in promoting. Such as various types of jigsaw puzzle toys, children by looking for patterns and characteristics of each first, and then put all of the puzzle together, together a complete picture of the children in the process of observation obtained the very good training [6]. For example, children play cars on the plane of different materials (wood, glass, cloth, etc.). By observing the different speeds of cars, they feel the existence of friction in the process of playing. In conclusion, the observation ability has been well developed in the process of playing with toys.

3 Explore Children's Toys in an Emotional Way

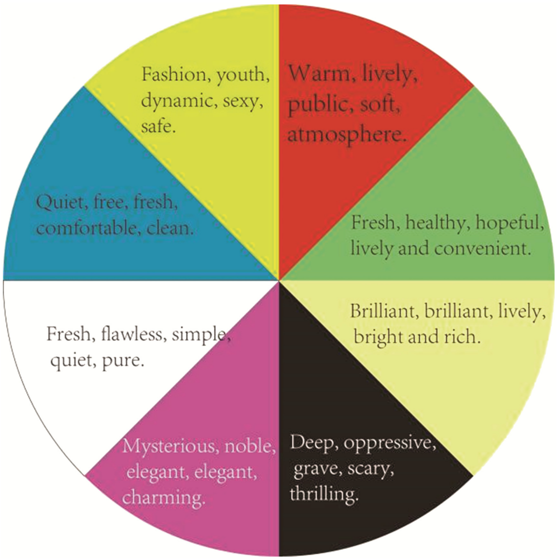
3.1 Color Image Experiment

Select Research Cases
Network through on-the-spot investigation, collection, magazine refer to a variety of ways such as sample collection, preliminary get 100 samples of children's toys, considering visual characteristics, to all sorts of color classification analysis, selection of representative and able to related comprehensively covers product place to choose color, finally select 15 representative sample.

The Selection of Perceptual Vocabulary
Choose the seven colors commonly used in toys: red, orange, blue, green, purple and black. The steps are as follows:

(1) Collection of vocabulary: from product books, magazines, toy product introduction manuals, and Internet, 150 representative emotional imagery words related to children's toy materials were collected;
(2) Meaning analysis: the respondents were randomly selected 50 mass consumers, through which questionnaires were issued and according to the conclusion.
After selecting the representative sample images of children's toys, we selected 54 perceptual images that were in accordance with the criteria.

Table 1. The sense of the color corresponds to the vocabulary



(3) The semantic object: random invited 30 mass consumer's perceptual image semantic clustering experiment, statistics based on the number of the same listed after the similarity matrix, after multiple scale method classification and cluster analysis by class, on the basis of the status of the material classification is divided into six groups, eventually take each group of center distance for the group of

representatives recently, the resulting in a group of representative perceptual image semantics, as shown in Table 1.

3.2 Data Sorting

The Participants

Before the investigation, the respondents were divided into designers with children's toy design experience and children's toys.

Use experienced consumer 2 groups.

Quantization of Semantics

The semantic differential component method is used to obtain the numerical value for statistical analysis, as shown in Fig. 1.

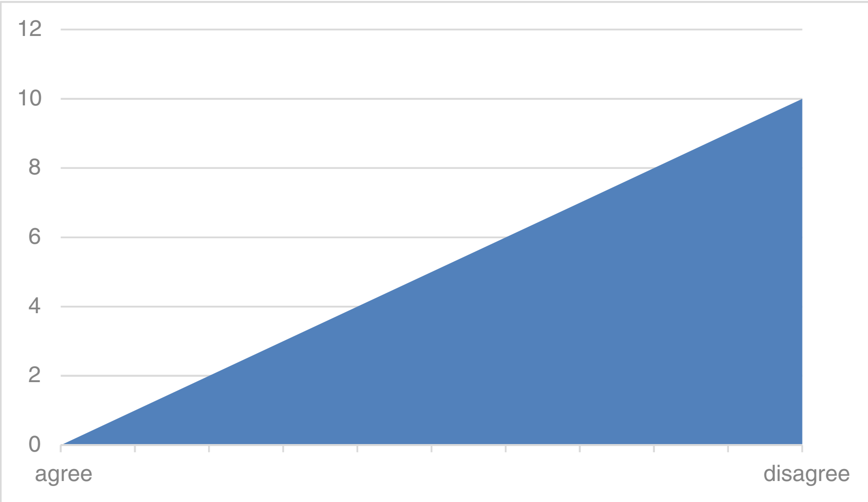
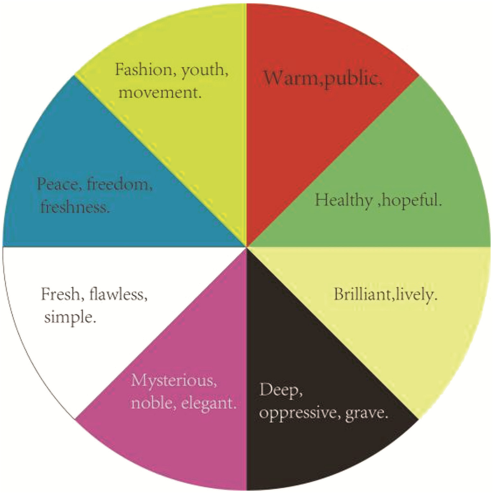


Fig. 1. Semantic quantization graph

Questionnaire Survey. After screening, finally determine the 6 groups of 18 adjectives, let consumers and designers get emotional information from children's toys on the material for each group of adjectives scores, thus build a perceptual image relation model between consumers and designers, as shown in Table 2. Selection of survey objects, one group of children toy for 15 are using consumer, and the other is a set of 15 have certain practical experience and a toy designer, to obtain the image evaluation of children's toys material, as shown in Figs. 2 and 3.

Table 2. Group diagram of perceptual vocabulary.



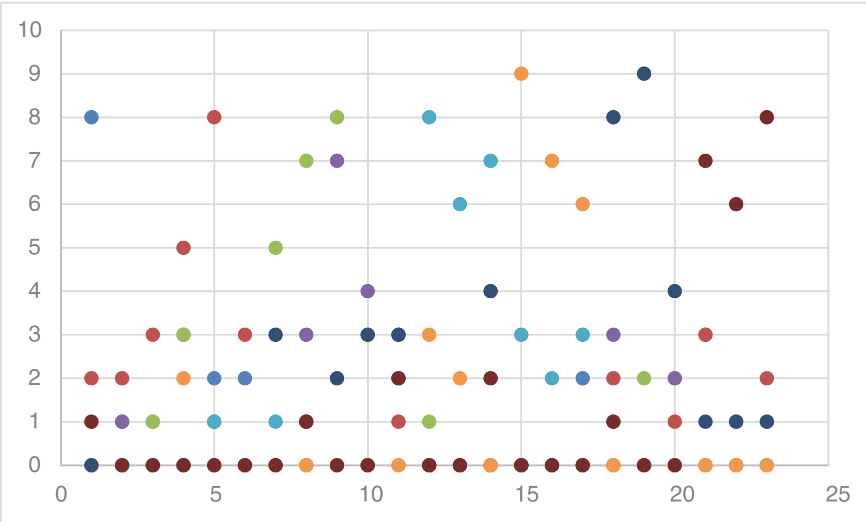


Fig. 2. Consumer information graph (Color figure online)

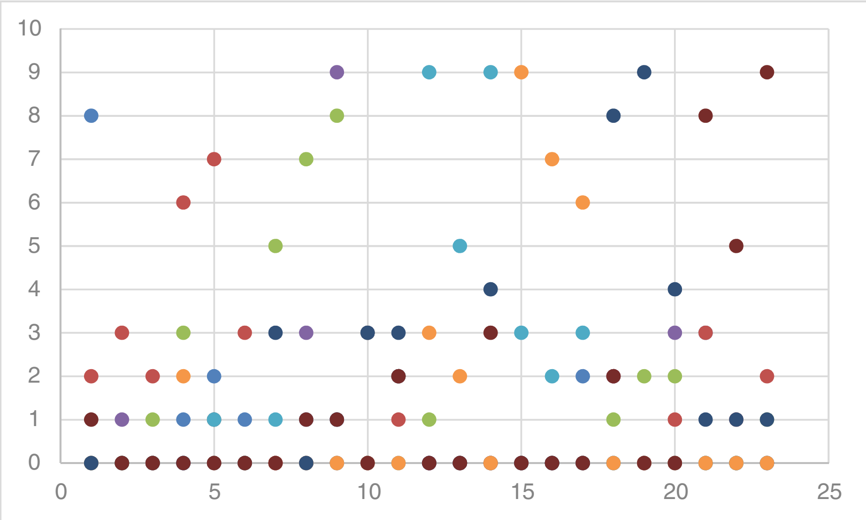


Fig. 3. Designer information diagram (Color figure online)

3.3 Material Related Results

Above from the perspective of children's toys material application, analyzes the selection of key points, children's toys and combined with market research and questionnaire survey and mathematical statistics analysis method, a detailed analysis, establish a toy design by means of the quantitative evaluation model of the conclusion. [7] In terms of material classification, it is suitable for children's toy materials to design plastic, wood and leather. Metal, rattan bamboo and cloth art are not suitable for making children's toys.

3.4 The Influence of Color on Children

Due to age restrictions, children have some limitations on the awareness of the surrounding things. They do not have the ability to think independently. This special physiological condition should be taken seriously by the designer. For children, the colors of children's toys attract attention more than shape, text, and sound [5]. Colors are used in children's toys to enrich the use of toys, stimulate children with color, enhance children's memory of pictures, and attract their attention. But with the ever-increasing interaction design, the rich use of color is no longer a simple picture, the picture will affect the story requires the use of color rendering, so that color and children's feelings interact [8]. Color is one of the important design elements of children's toys and is a necessary manifestation of children's toys.

The colors of physiological and psychological feelings of different ages are different, so as a designer, in the design of children's toys, in addition to the application of color color design features, the physical and mental states of the application are divided into different age groups. The festival will focus on analyzing the cognitive experience of children of different ages. Child psychologists divide children into three phases according to their physiological and psychological characteristics: Different ages and colors have different effects on people's psychology and emotions, especially children. Due to the particularity of children, we should also design them in the design of children's toys.

The development of newborn vision is the lowest of all sensory abilities. Newborns have a range of 1/10 to 1/30 of the average adult's vision, and by six months they have almost as much vision as adults [4]. Some researchers use conventional methods to study babies' visual abilities and discover the world's colors when they are born. The rapid development of the nerve center and sensory channel makes the children's color perception ability improve rapidly. 3 years old, due to its physiological development, the level of vision, hearing and language is limited. The response to color is only a conditioned reflex of the brain, a physiological response to color, not psychological. The famous professor of psychology at valentine to analyze this stage children of color perception and color preference, in kindergarten conducted an interesting experiment: the one hundred red bricks and one hundred blue bricks together, and let one hundred two to three years old children to choose their want to block, the results more than 90% of children chose the red color blocks, after obtaining blocks and was very happy. The results of this test confirm that children in this stage have a clear perception of color and can distinguish between red and blue colors. Children three years old have a special preference for red. They like to wear red clothes and red shoes. They also like red pens when they doodle. Young children are attracted by red color, red color is strong, can produce certain stimulation to the vision, help the growth and development of the baby brain. The deep color such as blue, black can make the child feel fear, adverse to their physical and mental development.

Therefore, as the designer, when designing children's toys for children's children, we can try to use bright colors to avoid the gray color.

Preschool age children in early childhood have certain thinking ability, they began to have their own ideas, also has made great progress in the language, you can say what they want, favorite things, and so on. This stage belongs to the color of a sensitive stage, by thought alone, they are not thinking of the feelings of color, when you see a certain color, will have a clear psychological reaction, and very direct through facial expressions, language, or physical behavior. "Child psychology research shows that four years of age had 98% of the children can correct to say "red", 94% of the children can correct a "black", 92% of children correctly say "green", 78% of children are able to correctly say "yellow". By the age of five, more than 70% of children can correctly name eight colors. By the age of six or so, more than 55% of children can correctly name 12 colors. This study proves that children in this stage have good visual foundation and color discrimination ability. They like the color of high purity such as red, yellow, blue and green. This color will give them a warm and cheerful color feeling. Preschool age children in addition to correctly identify part color, also to have certain color changes in temperature, light and heavy feeling, hard and soft feeling, and as the growth of the age, this kind of colour feeling more intense and accurate.

Children and preschool children's sense of color is more dependent on perceptual thinking, and has a certain preference for color purity and higher brightness. And 6 children as the growth of the age, also some changes have taken place in physical structure, visual and language expression ability has reached the level of adult, have own judgement ability, can the independent thinking ability. This age children are in elementary school, so different from infancy, preschool age children, environment changed from home to school such a large group, and parents communication, less interaction with peers became more frequent. Them for their favorite things have the initiative, want to wear what color clothes, what color back schoolbag have their own ideas, now the young parents will think of some way to try to meet to the requirement of the child. This stage children's understanding of color began to mature, no longer favor only in the colour of bright, excitant strong, gradually to the lightness and purity, low dark and cool color tones have interest. The color psychological effect also gradually reflected in them: they began to recognize the color symbolism, the color is rich in association, can feel the color of the cold, the sense of light. For example, see red in infancy and preschool age children, intuitive thought is a flower and the colour of the sun, and childhood is different, they can feel this is a happy color, also can produce sweet taste sense; Seeing black, they can feel the dark atmosphere that this color brings, still can have the psychological feeling such as hard feeling, weight, melancholy. Therefore, we should clarify the color psychological characteristics of children in this stage and start from their thinking logic.

In addition, "pre-school children have a certain gender difference in color perception". As is known to all, little girls' favorite color is pink, while boys are different, most of them like blue. Pink light tone, easy to remind them of the beautiful skirt, lenovo to the beautiful flowers in the garden; The boys, however, prefer a sedate tone that suits them better. Over time, in the people's inertia is considered strong warm girl's gender attributes, cool color to emphasize the boy's gender attributes, with color for girls and boys on the gender cues.

The Influence of Different Age or Gender on Toy Color. A questionnaire survey of toys for children of different ages and genders was conducted. Children with different age and age have different color preferences for toys. The color preference of children between 1 and 3 years old is not related to gender, the preference is blue and red, the boy of 7–10 years old is fond of blue, the girl likes red, yellow and orange are not very different. Blue and red are the most popular colors in different colors, followed by yellow and orange. Purple and green were the least favored by children (23.4%). As shown in Fig. 4.

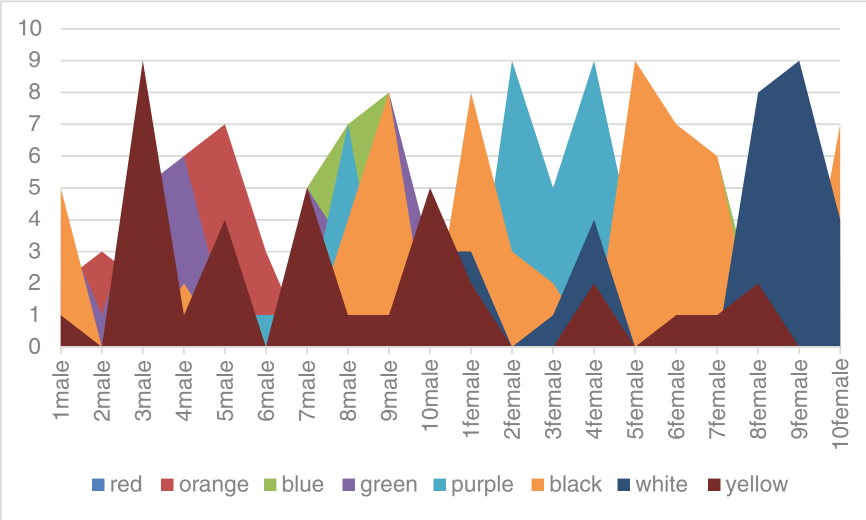


Fig. 4. Color of toys of different ages and sexes (Color figure online)

Through the above investigation, it can be found that children's toy color cognition is of age difference and gender difference. First of all, before the age of 3, children had no difference in color. Generally speaking, boys over four years old prefer cool colors, such as blue and green. Girls are more likely to enjoy warm colors, such as high purity yellow, red, pink, etc. Therefore, as the age increases, the children in this stage will be bored with the usual colors of red, yellow, blue and green, and do not like simple color combinations. Also is prominent, the color is lenovo at this stage in the children's consciousness, they will feel warm nervous, see red see yellow will feel warm, soothing, see will feel dark black, fear, and so on. Therefore, in children's toy color design, the first step is to strengthen the color complexity, from simple color to gradual change, with the brightness change, from concrete to abstract.

4 Conclusion

Children are the future of the country. Children's toys play a crucial role in the development and development of children. Currently, the design of children's toys is more entertaining and ignores the important education function. In the design of the toy, the consideration of perceptual engineering is deficient, and the related research is very weak. However, children's toys which are conducive to intellectual development and conform to the scientific basis of perceptual engineering are the urgent needs of children and parents. This study on the current children's toys design and research present situation has carried on the induction summary, emphasized the important role in children's toys, apply the scientific method of perceptual engineering and experiment of children's toys products material and colour is analyzed. Using the semantic difference component method to study and summarize the materials suitable for children's toys. Comparative study (0–3 years old) early childhood; (3–6 years old); (6 to 12 years old) children childhood for physiological and psychological characteristics of color, children's toys of different age and gender love colour has carried on the questionnaire survey, found that children's toys color cognition has the age difference and gender difference. After clarifying the physical and psychological characteristics of children, the designer can obtain the use demand more accurately, so as to make more rational use of the design language for children's toy design. In a follow-up study, the author will from a broader investigation scope, increase the number of samples, for voice, process more dimensions of research, children's toys to realize more scientific, complete and system design.

Acknlowledgement. The authors are grateful for the financial support provided by "The general project of Humanities and social sciences of Ministry of Education" (17YJA760022).

References

1. Zhang, Y.: Research on the status of children's toys. Shanghai Normal University (2012)
2. Zheng, S.: "Childhood study" needs to introduce children's voices. Chinese Social Science News, 2010-06-01(005)
3. Jiang, M.: A brief analysis on the color design of children's toys. Sci. Technol. Inf. **33**: 340 + 338 (2009)
4. Wang, D.: Color study of children's toy design in China. Jilin Art Institute (2012)
5. Chun, L., Tian, F.: Development of children's toy design based on sensory experience. J. Packag. **4**(03), 83–87 (2012)
6. Chen, Q., Lei, Y., Hong, L.: Influence of color, shape and size similarity and variability on children's inductive reasoning. Psychol. Dev. Educ. **27**(01), 17–24 (2011)
7. Guo, J., Yuan, Z.: Analysis and research on Kansei engineering of children's furniture. Furnit. Inter. Decor. **11**, 100–103 (2015)
8. Wang, J.: Application of color in children interactive ebooks. Hubei University of Technology (2012)

User Performance for Vehicle Recognition with Visual and Infrared Sensors from an Unmanned Aerial Vehicle

Patrik Lif(✉), Fredrik Näsström, Fredrik Bissmarck, and Jonas Allvar

Swedish Defence Research Agency, Linköping, Sweden
{patrik.lif,fredrik.nasstrom,fredrik.bissmarck,
jonasallvar}@foi.se

Abstract. In many situations it is important to detect and recognize people and vehicles. In this study the purpose was to examine human performance to detect and recognize vehicles on the ground from synthetic video sequences captured from a simulated unmanned aerial vehicle. A visual and an infrared sensor was used on an unmanned aerial vehicle with camera scan rate of the field of view on the ground relative to the ground of either 8 m/s or 12 m/s. The results from this study demonstrated that performance was affected by type of sensor, camera scan rate and type of vehicle. Subjects performed worse with infrared than with visual sensor and increased camera scan rate caused more errors. Also, the results show that recognition performance varied between 67 and 100% depending on type of vehicle. Recognition of specific vehicles was also affected negatively by interference from vehicles of similar appearance. Consequently, a vehicle with unique appearance within the set was easier to recognize.

Keywords: Vehicle recognition · Visual sensor · IR sensor · UAV
Human factors

1 Introduction

Gathering information with new and better sensors is positive since users can access more information, but it is necessary to have an understanding of what is the most vital information in a given situation. To accomplish this, users' need a good understanding of the whole system. Data overload may be a serious problem, and how to help human cognition using e.g. computers is fundamental to ensure good situation awareness and good user performance. Regardless of type of system it is also necessary to have a good understanding of the user and the context. The ecological approach [1] and representation design [2] describes a cognitive triad between *environment*, *interface* and *users*. There is a reciprocal coupling between the user and the environment, which often is mediated by a user interface. The interface effectiveness is determined by the mapping between the environment and interface (correspondence) and the mapping between the user and interface (coherence). To develop an effective and user-friendly system all these three parts must be taken into account. Information that reaches the user has often been

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 295–306, 2018.
https://doi.org/10.1007/978-3-319-91238-7_25

acquired with some type of sensor system that involves signal processing, acting as a filter between the environment and the interface. In order to be able to understand the complete picture of study sensor-related aspects the model has to be extended to also include environment, interface and human aspects. Since the sensor is a central part in our research, we add the sensor to the representation visualization (Fig. 1).

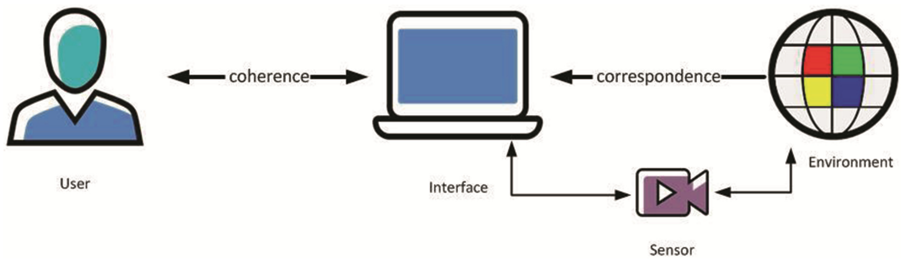


Fig. 1. The relation between user, interface, environment and sensor. Icons were adopted from Iconshock [3].

Even though the whole system always have to be taken into account, the main focus here is on the ability and limitations of the users and their performance to extract correct information from sensor data.

Seeing an object could mean different things, but one way to analyze observers' ability to perform visual tasks is to use the Johnson criteria [4, 5] that distinguish between *detection* (i.e. whether there is something of potential interest), *recognition* (e.g. the difference between a human and car) and *identification* (e.g. whether it is a friend or foe). According to Johnson criteria, possible detection distance is calculated based on how many pixels an object must contain. In order to detect static objects it requires 2×2 pixels, orientation 8×2.8 pixels, 8×8 pixels for recognition, and identification required 12.8×12.8 pixels [6]. However, this should be interpreted as values under best possible conditions. There is also a variety of factors that must be considered, including the contrast between objects and background, atmospheric disturbances, the number of objects in the picture, light, contextual clues, color and type of optics. Moreover, performance is affected by the type of task, the experience of the participants and their level of training for the specific task, motivation, and the relative importance between quick decisions and correct results [4]. Also the methods Triangle Orientation Discrimination (TOD), Targeting Task Performance (TTP) and Thermal Range Model (TRM) could be considered. For further description of these methods see Näsström et al. [7], Wittenstein [8], and Vollmerhausen and Jacobs [9].

Even though theoretically calculated values (e.g. Johnson criteria) could be of some value to get an indication of what objects that can be detected, recognized or identified, experiments with users should be conducted to get a better understanding of a real situation. There is an obvious risk of confusion regarding the interpretation of concepts, since the concepts are used by researchers in different context without a standardized definition. It is absolutely necessary to clarify and define the concepts used.

Identification of friend or foe is different from actual identification of a face from memory or a database. In many situations one must be absolutely certain about the identity of a person or vehicle to make a decision whether to use military force. Also, it is necessary to have a good understanding of roles of engagement (ROE), when military force can and cannot be used. Friendly fire, where a soldier accidentally opens fire on his own troops, is a well-known phenomenon that must be avoided. In other cases, such as intelligence, is it important to describe what is seen according to a predetermined classification scheme and not just describe what users think they see.

In military contexts, it is sometimes important to find a particular type of vehicle among other similar military vehicles, and it is also important to distinguish between military and civilian vehicles. To increase knowledge about this, our work involves assessing actual sensor performance but also investigating how operators use and interpret sensor information. Even though the interest from a human factors perspective is mainly on user performance to detect and recognize people and vehicles, we also conduct technology driven sensor studies [10] and thorough investigation of the real setting [11]. There are many interesting studies focusing on detection, recognition and identification. Colomina and Molina [12] discuss the evolution and use of unmanned aerial systems in photogrammetry and remote sensing that can be used in both military and civilian operations, e.g. search and rescue missions. Other research with unmanned aerial vehicle (UAV) and target detection focus has a more technical approach, e.g. develop algorithms for autonomous target detection [13] or autonomous UAVs for search and rescue [14]. There are also interesting studies using multiple cooperative vehicles [15] or a swarm of unmanned vehicles [16] which shows that multiple vehicles can improve performance. Other research has a clearer connection to human factors issues and user performance. Hixson et al. [17] used soldiers to investigate the relation between performance in the laboratory and in the field for tasks including detection, recognition and identification. The results shows that perception laboratory performance using real or simulated imagery relates well to imagery performance in the field.

The research question in the first experiment was to investigate how fast and to what degree of correctness can users detect and recognize one selected military vehicle among other similar vehicles and how is performance affected by type of sensor, camera scan rate of the field of view on the ground (hereafter referred to as scan rate) and distance? The research question in the second experiment was to investigate to what degree of correctness can users recognize eight military vehicles with an infrared sensor, at camera scan rate of 8 m/s at a distance of 400 meters?

It is important to investigate and understand the sensors' pros and cons in different situations. Only the infrared sensor can be used at night while both the visual and infrared sensor can be used during daytime. However, it is not obvious which sensor is preferred during daytime in different situations and it is therefore important to investigate this. In some situations it is certainly better to use the visual sensor, but sometimes the vehicle can be partly hidden under e.g. branches or trees and then it is advantageously to use the infrared sensor also during daytime. From a tactical perspective it may be advantageous to fly the unmanned aerial vehicle at night, but then only the infrared sensor can be used. Also, at night there are significantly fewer civilian vehicles in motion and less vehicles that gives heat signatures which facilitates detection and recognition of military vehicles. If performance decreases with one of the sensors quantification would be important. It is preferable to use high camera scan rate since larger geographic areas can be covered, but if it results in decreased performance it may be necessary to use a lower speed. Even though user performance is expected to decline at increased camera scan rate it is important to objectively quantify performance decrease. If the unmanned aerial vehicle fly at high altitude there are tactical advantages such as lower risk for the UAV being detected, but if it results in decreased performance it is not recommended.

Here, two experiments were conducted that is part of a larger study where the overall goal is to investigate how different sensors should be used in unmanned aerial vehicles to gather information. The purpose with these two experiments were to investigate subjects' performance of vehicle detection and recognition from a simulated unmanned aerial vehicle. In the first experiment, detection and recognition of one selected vehicle among a total of eight vehicles was investigated at two different camera scan rate (seen from the UAV) with visual- and IR-sensor. In the second experiment recognition of all eight vehicles was investigated at a camera scan rate of 8 m/s with an IR-sensor. Although the results here are only presented and analyzed strictly linked to these experiments, later it can be analyzed and compared to other experiments. Also, this information can be used to better understand how information from different sensors can be aggregated to increase performance. However, this is not the focus here and is therefore not presented in this paper.

2 Experiment 1 – Detection and Recognition of Selected Vehicle

In the first experiment, detection and recognition of one selected vehicle among a total of eight vehicles was investigated.

2.1 Method

Participants watched synthetic video sequences captured from an UAV. All video sequences were generated by a sensor simulation system [10]. The task was to detect and recognize a selected vehicle among other vehicles. A within-group design with *two visualizations (*visual and IR) × *two distances (400 and 520 meters)* × *two camera* scan rate (8 and 12 m/s) was used.

Subjects

Twelve subjects (5 women and 7 men) between 25 and 48 years participated in the experiment. Half of the participants' had military background and the other participants' were well acquainted with military activities through their civilian jobs. However, none of the participants were experts on the vehicles presented in these experiments and therefore trained prior to the experiments started. All had adequate vision with or without correction.

Apparatus

The video sequences were presented on a Dell Latitude 7240 with 12.5 inch display with a resolution of 1366 × 768 pixels. The computer had 4th generation Intel® Core i5 and i7. A self-developed software was used to present stimuli and record participant's response time.

Stimuli

A total of eight videos (640 × 480 pixels) were generated during a clear sunny day with shadows from targets on the ground to depict sensor information from a visual- and infrared sensor (Fig. 2). The overall mission was similar to a real UAV flying along a predefined path with vehicles stationary on the ground.

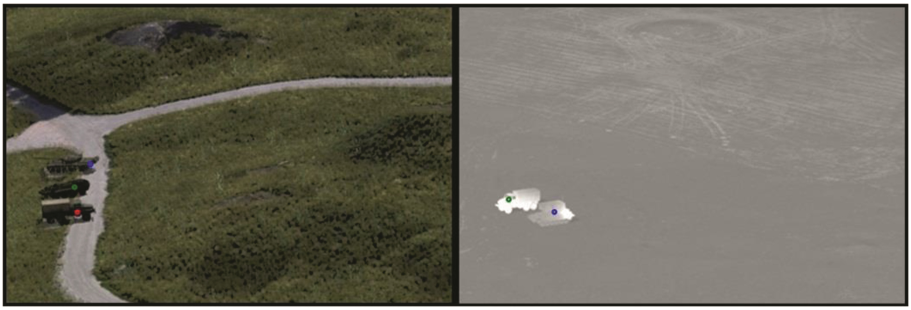


Fig. 2. Still images from the visual sensor (left) and the IR sensor (right).

The task was to detect and recognize one selected vehicle among a total of eight vehicles. The eight vehicles were BMP-3, BTR-80, MT-LB, SA-19, T-72, TOS-1, Ural 4320 Ammunition truck, and Ural 4320 fuel truck (Fig. 3).

Fig. 3. The eight vehicles used.

Four scenarios were generated with the visual- and infrared sensor respectively. Each scenario had 18 areas with different target positions. The same areas and positions were used for the visual- and infrared scenarios. A total of eight videos were generated according to the aforementioned design. The visual- and infrared scenarios were presented in a balanced order between subjects', and within each sensor the four scenarios were presented in a randomized order.

Procedure

After welcoming the participants individually and briefing them about the experiment purpose and procedure they received written information and had the opportunity to ask questions to the experiment leader. Then an introduction was given to make sure that the participants were familiar with the situation and test material. They were introduced with both visual- and infrared image visualizations and received training, which consisted of two three minutes scenarios, one for visual- and one for infrared stimuli. The participants watched the videos and answered by first pressing the space bar whereby the response time (RT) was recorded and then used the left mouse button to annotate in the image to indicate the selected vehicle position. The annotation was later used to calculate number of correct answers. The participants were instructed to always focus on the screen with the stimuli. Because the task was mentally demanding it was divided into eight separate videos with the possibility to rest before continuing with the next one.

2.2 Results

The results include statistical analysis of time to detect targets and recognition of the selected vehicle. The data were analyzed with a three-way ANOVA [18] with type of visualization (visual and infrared), camera scan rate (8 and 12 m/s), and distance (400 and 520 meters). Tukey HSD was used for post hoc testing [19].

Detection

The ability to detect targets was measured by response time (RT) and analysis was performed by ANOVA repeated measures. The results showed no significant main effects of response time ($p > .05$).

Recognition of one selected vehicle

The ability to recognize one selected vehicle was analyzed by ANOVA repeated measurement, where mean values for each condition was used for each participant. The results showed a main effect for type of sensor $F(1, 11) = 9.02$, $p < .05$, where participant's performance were lower with the infrared sensor than with the visual sensor (Fig. 4).

There was also a significant main effect of camera scan rate $F(1, 11) = 8.75$, $p < .05$, where higher camera scan rate caused more errors (Fig. 5). There was no significant main effect of distance, and no significant interaction effects $p > .05$.

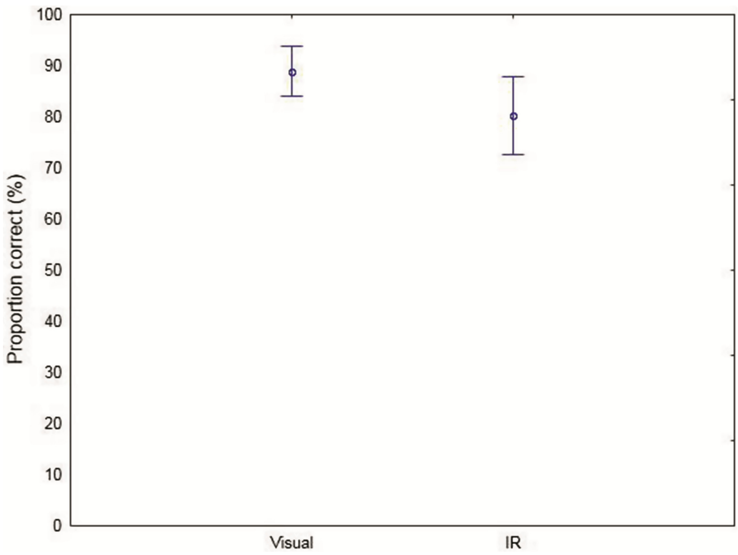


Fig. 4. Mean and standard error of mean for proportion correct answers for visual- and infrared (IR) sensor.

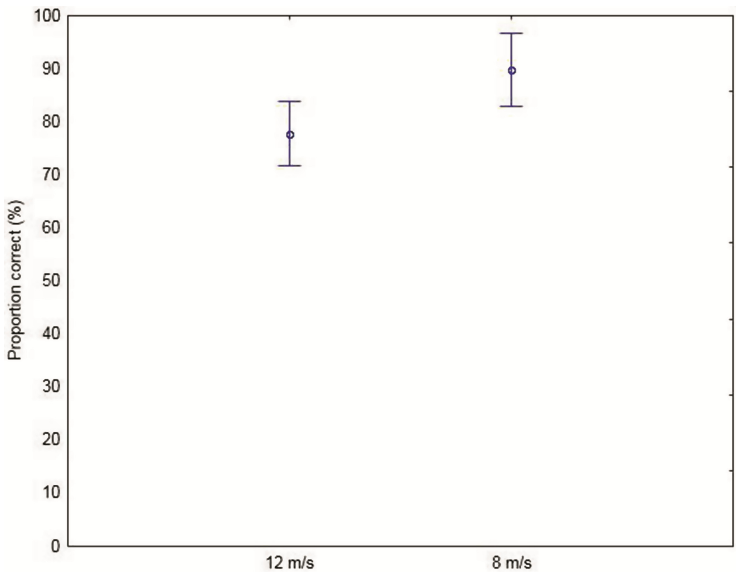


Fig. 5. Mean and standard error of mean for proportion correct answers for 12 m/s and 8 m/s.

3 Experiment 2 – Recognition of Eight Vehicles

In the second experiment, detection and recognition of a total of eight vehicles were investigated.

3.1 Method

From Experiment 1, the scenario with an infrared sensor, distance of 400 meters, and camera scan rate 8 m/s was selected. For this setting recognition of eight different vehicles was investigated. In this experiment the focus was on proportion correct recognized vehicles only. The subjects' watched the video sequences for five seconds and then reported their answers, no response time was measured.

Subjects
Twelve subjects (4 women and 8 men) participated in the experiment. Five of the participants' had military background and the other participants' were well acquainted with military activities through their civilian jobs. However, none of the participants were experts on the vehicles presented in these experiments and therefore trained prior to the experiments started. All had adequate vision with or without correction.

Apparatus
See experiment one for technical description. Superlab [20] was used to present the video sequences and to record proportion correct answers.

Stimuli
Two video sequences (640 × 480 pixels) with a total of nine stops, where the subject's task was to recognize vehicles from a total of eight vehicles. The same eight vehicles were used as in experiment 1, but in this experiment the task was to identify all eight vehicles, not only one selected vehicle.

Procedure
Overall the procedure was as in experiment 1. However, there were some differences due to another design. In experiment 2, the video sequences was paused at predefined occasions, and one vehicle was indicated by a circle. The subject's answered by pressing the number 1–8 on the keyboard, and then the next stimuli was indicated by a circle. The procedure was repeated 1–4 times at each scenario stop depending on the number of vehicles in that particular stop.

3.2 Results

The results include statistical analysis of recognition of the eight vehicles, which are analyzed with one-way ANOVA [10]. Tukey HSD was used for post hoc testing [11]. The results showed a significant effect of vehicle type $F(7, 77) = 5.54$, $p < .001$ (Fig. 6).

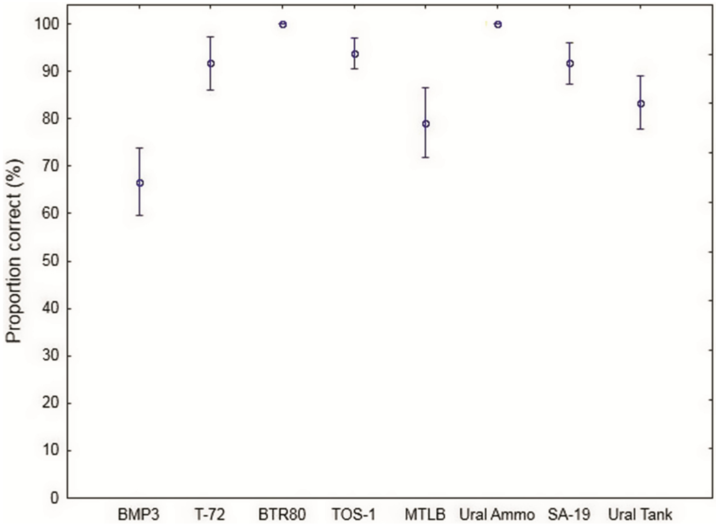


Fig. 6. Mean and standard error of mean for proportion correct answers of the eight vehicles.

Tukey HSD showed that BMP-3 was significantly harder to recognize than most of the other vehicles (except the MT-LB). Vehicles T-72, BTR-80, TOS-1, Ural 4320 ammunition truck, and SA-19 were recognized in 90% of cases or more, while BMP-3, MT-LB and Ural 4320 fuel truck were more difficult. BMP-3 was mainly confused with BTR-80 and MT-LB. MT-LB was mainly confused with BTR-80, and Ural 4320 fuel truck was mainly confused with Ural 4320 ammunition truck. For a more detailed description see the confusion matrix (Table 1). The first column show the vehicle name and the second column show percentage correct recognition. Column three to eight show which other vehicles the vehicle (in first column) was confused with.

Table 1. Confusion matrix that shows which other vehicles the eight vehicles were confused with. All numbers are presented in percent, and each row add up to 100%

Vehicle	% correct	BMP-3	T-72	BTR-80	TOS-1	MT-LB	Ural 4320 ammunition	SA-19	Ural 4320 fuel truck
BMP-3	66,6		4,2	12,5		12,5		4,2	
T-72	91,7		8,3						
BTR-80	100								
TOS-1	93,8						4,2	2,0	
MT-LB	79,1	4,2		12,5					4,2
Ural 4320 Ammunition	100								
SA-19	91,7	2,8			5,5				
Ural 4320 fuel	83,3						16,7		

4 Discussion

The experiments presented here is part of a larger study where user performance to detect and recognize people and vehicles is investigated, technology driven sensor studies are performed [10], and thorough investigation of the real setting [11] conducted. The combination of improved technical knowledge, understanding of the real environment, and user performance is seen as a good interdisciplinary combination to better understand how a final system can make a difference in real settings.

The purpose with these two experiments were to investigate subjects' performance of vehicle detection and recognition from a simulated unmanned aerial vehicle. The results shows that the ability to recognize vehicles is affected by type of sensor, camera scan rate, and type of vehicle that is to be recognized. User performance to recognize the selected vehicle among a total of eight vehicles was significant lower with the infrared- than with the visual sensor, and significant lower at camera scan rate 12 m/s than at 8 m/s. Also, the results show that recognition performance varied between 67% and 100% depending on type of vehicle. The results from the second experiment clearly shows that vehicle recognition with the infrared sensor is problematic, even though short distance (400 meters) and slow camera scan rate (8 m/s). The results also show that certain types of vehicles are particularly difficult to recognize which is an important operational information in military contexts.

In situations where the vehicles are placed in open terrain as in these experiments, it is advantageous to use the visual sensor. However, the infrared sensor allows detection of vehicles in situations where there is no clear view, in situations with low visibility and at night. In these situations the heat signature from the infrared sensor can be used to detect and recognize vehicles. Another possibility is to combine information, either by switching between the two sensor images or by fusing the sensor images into one image. However, this was not investigated in these experiments and therefore not reported here.

From a scientific perspective it is important to understand perceptual and cognitive possibilities and limitations. As a part of this we investigated how the type of information presented (visual and infrared), camera scan rate and distance affected user performance. Although there are a number of other factors that affect performance, this contributes to knowledge about vehicle detection and recognition in this military context. For practical and economic reasons it is not always possible to conduct field studies and therefore laboratory studies can be used as an important compliment. The results presented here can also be correlated with results from similar field studies (not yet performed). Results from laboratory experiments are especially valuable if they can predict performance in real environments, which for us is a future challenge.

It is also important to use systematic methods for data collection and result analysis, which gives the possibility to compare and analyze the results relative to other scientific results. The results from our experiments can later on be analyzed and correlated with calculated values from e.g. the Johnson criteria or relative to sensors' technical performance to get an understanding of correlation between human performance and technical performance. In this study, no specific sensors have been presented, but the results from this work can be used for evaluating the existing sensors that the simulation here were based on. This work remains to be done and is therefore not presented here.

Even though these experiments and prior experiments [21] gives a good understanding of user performance to detect and recognize people and vehicles from an unmanned aerial vehicles there are some limitations. In these studies we used a predefined flying path, as often is the case in real settings, but it would be interesting to let the users manually control the sensor direction and give them the possibility to zoom-in to targets. Also, in these experiments it was daytime and strong sunshine, which gave clear shadows of vehicles. It would be interesting to compare the results achieved in this study with results from a daytime scenario with cloudy weather without clear and sharp shadows visible. Also, night time scenarios would be interesting to investigate. In this study, vehicles were placed on open surfaces in the terrain, but it would be interesting to see how different camouflage (such as nets or trees) would affect the ability to recognize vehicles especially with visual camera.

Detection and recognition was investigated in this study, but it would be interesting to also investigate identification of people and vehicles in a similar setting as the experiments presented here. One limitation of this study is that although the visual and infrared sensor data are realistic, no scientific verification has been made to confirm the similarity between the used stimuli material and real data from sensors. However, one researcher compared the simulated videos with real sensor information and confirmed that the material looked similar [10]. In the future, this procedure need to be improved with standardized objective measures.

The information presented from this study is important since user performance and technical knowledge can be aggregated and used to understand operational performance and limitations. Issues such as camera scan rate, type of sensor, flight altitude, weather conditions, and time of day, are important and can be put in a broader context to understand how a task best can be solved.

References

1. Flach, J.M., Hancock, P.A.: An ecological approach to human-machine systems. In: Proceedings of the Human Factors Society Annual Meeting, vol. 36, no. 14, pp. 1056–1058 (1992)
2. Woods, D.: Toward a theoretical base for representation design in the computer medium: ecological perception and aiding human cognition. In: Flach, J., Hancock, P., Caird, K., Vicente, K. (Eds.) An Ecologixal Approach to Human Machine System, Erlbaum, Hillsdale, New Jersey, pp. 157–188 (1995)
3. Iconshock: “Icon man,” Smashingmagazine (2017). SmashingMagazine.com
4. Donohue, J.: Introductionary review of target discrimination criteria. Wilmington, MA (1991)
5. Sjaardema, T.A., Smith, C.S., Birch, G.C.: History and Evolution of the Johnson Criteria. Albuquerque, New Mexico (2015)
6. Kopeika, N.S.: A System Engineering Approach To Imaging. SPIE Optical Engineering Press (1998)
7. Näsström, F., Bergström, D., Bissmarck, F., Grahn, P., Gustafsson, D., Karlholm, J.: Prestationsmått för sensorsystem, Linköping, FOI-R–4139–SE (2015). (In swedish)

8. Wittenstein, W.: Thermal range model TRM3. In: Proceedings Volume 3436, Infrared Technology and Applications XXIV, vol. 3436, p. 413 (1998)
9. Vollmerhausen, R.H., Jacobs, E.: The Targeting Task Performance (TTP) Metric A New Model for Predicting Target Acquisition Performance. Fort Belvoir, VA, Technical Report AMSEL-NV-TR-230, 2004
10. Näsström, F., Allvar, J., Deleskog, V.: Simulation framework for research of 'intelligent' reconnaisance systems (2017)
11. Näsström, F., et al.: Värdering av sensorsystem, Linköping, FOI-R–4474–SE (2017). (In swedish)
12. Colomina, I., Molina, P.: Unmanned aerial systems for photogrammetry and remote sensing: a review. ISPRS J. Photogramm. Remote Sens. **92**, 79–97 (2014)
13. Hinas, A., Roberts, J.M., Gonzalez, F.: Vision-based target finding and inspection of a ground target using a multirotor UAV system. Sensors **17**(12), 2929 (2017)
14. Sun, J., Li, B., Jiang, Y., Wen, C.: A camera-based target detection and positioning UAV system for search and rescue (SAR) purposes. Sensors **16**(11), 1778 (2016)
15. York, G., Pack, D.J., York, G., Pack, D.J.: Ground target detection using cooperative unmanned aerial systems. J. Intell. Robot. Syst. **65**, 473–478 (2012)
16. Altshuler, Y., Pentland, A., Bruckstein, A.M.: The cooperative hunters – efficient and scalable drones swarm for multiple targets detection. Swarms and Network Intelligence in Search. SCI, vol. 729, pp. 187–205. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-63604-7_7
17. Hixson, J.G., Teaney, B.P., May, C., Maurer, T., Nelson, M.B., Pham, J.R.: Virtual DRI dataset development. In: Proceedings of Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXVIII, vol. 10178 (2017)
18. Hays, W.: Statistics, 5th edn. Harcourt Brace College Publishers, New York (1994)
19. Greene, J., D'Oliveira, M.: Learning To Use Statistical Tests In Psychology. Open University Press, Philadelphia (1982)
20. Cedrus: "SuperLab 5," Cedrus (2017). https://www.cedrus.com/superlab/. Accessed 20 Nov 2017
21. Lif, P., Näsström, F., Tolt, G., Hedström, J., Allvar, J.: Visual and IR-based target detection from unmanned aerial vehicle. In: Yamamoto, S. (ed.) HIMI 2017. LNCS, vol. 10273, pp. 136–144. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58521-5_10

Reading Machine-Written News: Effect of Machine Heuristic and Novelty on Hostile Media Perception

Bingjie Liu(✉) and Lewen Wei

Donald P. Bellisario College of Communications, The Pennsylvania State University, University Park, PA 16802, USA
{bxl5252,lpw5086}@psu.edu

Abstract. The use of artificial intelligence (AI) in news production has been increasing these days. Its inanimate nature might bring opportunities to suppress hostile media effect by reducing readers' emotional involvement. This study conducted an online experiment ($N = 175$) to examine how the identity of news writers, human vs. robot, might impact readers' perceptions of news and source credibility. Findings suggest that readers holding machine heuristic, i.e., those who saw machine as free of intention, experienced less emotional involvement when the news was purported to be written by a news writing algorithm, as compared with when reading human-written news. Lower level of emotional involvement further led to less perception of bias in the news and the extremity of news slant. However, perceived novelty associated with robot news writer, although enhanced positive perceptions of the news, intensified readers' emotional involvement, which further heightened hostile media perceptions. Findings in this study identified the mechanisms underlying effects of AI writer in the context of controversial news topics. Implications are discussed under theoretical frameworks of general information processing, hostile media effect and MAIN model.

Keywords: Robot journalism · Machine heuristic · Emotional involvement
Hostile media effect

1 Introduction

On February 1st, 2013, eight minutes after an earthquake of 3.2 magnitude shaking the coast of San Simeon in California, *The Los Angeles Times* published a news story written by their news algorithm about what just happened (Marshall 2013). More impressively, the news was not distinguishable from that written by journalists (Latar 2015). The algorithm was programmed to ask questions that an experienced journalist would ask. In the case of earthquake, for example, it was set with filtering criteria such as location (i.e., California) and earthquake magnitude to judge if the happening was newsworthy (Marshall 2013).

In order to manage cost of news production and boost profit margin, news organizations have been motivated to engage artificial intelligence (AI) in newsroom. As early as in 2006, Thomson Reuters decided to use algorithms to generate finance

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 307–324, 2018.
https://doi.org/10.1007/978-3-319-91238-7_26

news (Wired 2006). In 2010, algorithms were developed in university labs with the purpose to substitute human journalists (Latar 2015).

Technology-assisted news production is hardly any news as journalists used computers to help produce news stories since 1952 (Cox 2000; Örnebring 2010). However, the idea of replacing professional journalists with AI seems to be disturbing and has elicited concerns among both professional journalists and the public (Latar 2015; van Dalen 2012). In spite of all these concerns, the idea of machine writing news might not be all bad. As the public tends to think of machine as inanimate and thus free of intentions and biases (Minsky 2006; Sundar 2008; Sundar and Nass 2001; van Dalen 2012), recent advancement in AI and its application in journalism practice seems to bring some hope in suppressing hostile media effect (HME) by suppressing readers' emotional involvement (Matthes and Beyer 2015).

Although promising, research has not yet reached a unanimous conclusion on how well machine-written news is received by human judges and whether it can actually enhance fairer judgment on news content (e.g., Edwards et al. 2014; Gambino and Kim 2015; Van der Kaa and Krahmer 2014). Against the backdrop of the rapid development of AI and its application in news production (Latar 2015), this study aims to examine the impact of robot journalism on hostile media perceptions (HMPs) and explore the potential underlying mechanisms.

1.1 Hostile Media Effect

News coverage on controversial issues tends to induce HMP, especially for partisans with high issue involvement, i.e., the extent to which one perceives the issue as personally relevant (Cho and Boster 2005; Gunther and Liebhart 2006). News that "most nonpartisans find evenhanded and objective" is perceived as biased simply because partisans are "exerting pressure in the hope of more favorable media treatment" (Vallone et al. 1985, p. 578). Even in cases when the news is in favor of one's own side, they judge the magnitude of this favoritism as less compared with that judged by their opposing party (Gunther and Chia 2001; Gunther et al. 2001).

Besides issue involvement, research has found emotional involvement a crucial process above and beyond cognitive involvement through which HMP develops (Matthes 2013; Matthes and Beyer 2015; Vallone et al. 1985). Emotional involvement, characterized by both emotional arousal and experience of specific emotions (Matthes 2013), is defined as the emotional reactions associated with an attitude object (Wirth 2006).

1.2 Effects of Robot Journalism I: Machine Heuristic and Reduced Mind Attribution

The rise of robot journalism seems to bring some hope in suppressing HMP in that it is widely held that machine, given its inanimate nature, is free of intention and bias (Minsky 2006; Sundar 2008). Heuristics are mental short cuts or rules of thumb developed from experience and ready to be applied in judgements making (Bellur and Sundar 2014; Chen and Chaiken 1999; Kruglanski and Thompson 1999). Machine heuristic is the association between inanimate technology and the qualities of being objective, free of bias, morally neutral and intentionless (Sundar 2008). As the MAIN model suggests, interface cues of digital media might trigger certain associated heuristics and influence users' impressions of the content quality and credibility (Sundar 2008). When reading machine-written news, readers' heuristics associated with this identity cue might be activated, influencing readers' evaluations of the source and the content.

For example, news story purported to be selected by a computer terminal was evaluated as of higher quality compared with the very same news purported to be suggested by news editors because machines were perceived as "truly random and hence representative" (Sundar and Nass 2001, p. 68). In the context of personal assistant, people were more willing to reveal their credit card information to Siri than to a human service agent who helped them book flight ticket because they believed that Siri, as a machine, was free of malicious intentions such as stealing and abusing the information (Kim and Sundar 2016).

Intention attribution also underlies the development of HMP. Research suggests that HME is more prominent when the source is deemed to have more potential in swaying public opinion towards the opposite side. For example, a news story written by a journalist triggered more HMP than the very same news purported to be written by a student (Gunther and Schimitt 2004; Gunther and Liebhart 2006) since a journalist is supposed to have higher motivation and ability to exert influence on public opinion than a student. Another study found football fans perceived news purported to be published on the newspaper of their rival university's town as more biased compared to that from hometown and neutral-town newspaper (Arpan and Raney 2003) as a newspaper from the rival's town is supposed to be more motivated to report in favor of their own university team. In persuasion literature, intention attribution is also regarded as the cause of reactance and arousal such that source's influence intention is perceived as a threat or restriction to the target's freedom. Attribution to persuaders' influence intention further leads to psychological reactance characterized by anger and negative cognitions (Brehm and Brehm 1981; Dillard and Shen 2005; Rains 2013) and defensive attitudes and behaviors (Grandpre et al. 2003; Miller et al. 2007; Shen 2011), such as derogating the source as biased and unfair (e.g., Worchel 1974), which also characterizes HMP (Hwang et al. 2008).

As machine is believed to be free of bias and intention, using AI to write news might reduce emotional reactions and general arousal by suppressing intention attribution. Therefore, the lack of intention attributed to robot writer might have the potential to reduce HMP by reducing readers' emotional involvement. Moreover, effects of robot writer on emotional involvement should only apply to those who hold the machine heuristic that robot is free of intentions and bias but not for those who do not.

1.3 Effects of Robot Journalism II: Novelty and Eeriness

Although machine's inanimate nature might suppress intention attribution and thus emotional involvement, machine-written news as a new phenomenon to the general public (Latar 2015) might induce perception of novelty and eeriness as suggested by existing research. Specifically, perceived novelty associated with robot writing news might lead to positive psychological experience and positive evaluation (e.g., Gambino and Kim 2015; Sundar et al. 2014). Whereas perceived eeriness (Mori 1970; Stein and Ohler 2017) caused by inanimate agent presented with too much humanness, may further lead to negative perceptions in terms of the news' credibility and quality (e.g., Gambino and Kim 2015).

Besides the halo effect of novelty and eeriness on readers' overall evaluations of the news, novelty and eeriness associated with reading machine-written news might increase their emotional involvement while reading. Findings in neuroscience suggests that novelty stimulates affective reaction involving amygdala activation distinctively from stimuli's valence and arousing level (Moriguchi et al. 2011; Schomaker and Meeter 2015; Weierich et al. 2010). Eeriness, by definition is a repulsive response characterized by arousal and psychological discomfort experienced along with cognitive dissonance (Elliot and Devine 1994; Ferrey et al. 2015). Readers' emotional involvement might be enhanced by the sense of novelty and eeriness as a result of excitation transfer such that the arousal due to the novelty and eeriness associated with robot writing news might amplify one's emotional reaction to news content (Zillmann 1971), which further intensifies HMP.

Based on discussions above, we proposed the following hypotheses regarding the different mechanisms underlying effect of robot news writer on HMP.

H1. For individuals holding machine heuristic, robot writer leads to less emotional involvement as compared to human writer.
H2. Compared with human writer, robot writer leads to more perceived novelty, which further increases readers' emotional involvement.
H3. Compared with human writer, robot writer leads to more perceived eeriness, which further increases readers' emotional involvement.
H4. Compared with human writer, robot writer leads to less emotional involvement, which further reduces HMP.
H5. Compared with human writer, robot writer leads to more perceived novelty, which further reduces HMP.
H6. Compared with human writer, robot writer leads to more perceived novelty that further increases readers' emotional involvement, which intensifies HMP.
H7. Compared with human writer, robot writer leads to more eeriness, which further increases HMP.
H8. Compared with human writer, robot writer leads to more perceived eeriness that further increases readers' emotional involvement, which intensifies HMP.

1.4 Conceptualization and Operationalization of Hostile Media Perception

HMP is defined as the perception that news media are not in favor of one's own stand (Gunther 1992). However, in empirical studies, HMP has been operationalized in multiple ways. To better explore the effects of robot journalism on HMP, this study operationalized HMP in four different ways, namely (a) news slant extremity (Matthes and Beyer 2015), (b) bias perception, (c) news credibility (Appelman and Sundar 2016), and (d) source trustworthiness. As Matthes (2013) argues, although credibility seems not to be a direct measure of HMP, "bias is often considered a marker of the larger construct of news credibility" (Arpan and Peterson 2008, p. 325) and reduced credibility has been considered as a direct outcome of HMP and has been found to be highly correlated to bias measures (Arpan and Peterson 2008; Choi et al. 2009; Tsfati and Cohen 2005).

2 Method

2.1 Study Design

To examine the effect of robot news writer relative to its human counterpart, identity of news writer was manipulated. To avoid ending up with lack of variability on the dependent variables with a too simple news article, participants were randomly assigned to read either a short spot news article or a longer interpretive news article (DeMott 1973) about climate change purported to be written either by a human writer or by a news bot. Experiment was administered online with Qualtrics. All the news articles were embedded in the template of *The New York Times* and displayed as screenshots in the questionnaire.

2.2 Participants

Participants were 212 Mechanical Turkers. After manipulation check screening, only those who correctly identified the news writer's identity were remained. The final sample was composed of 175 participants, aging from 18 to 74 years old (M = 38.35, SD = 13.97), with 42.70% males, and 77.13% ethnically affiliated as Caucasians, 8.00% Latino/Hispanic, 7.43% Asian/Pacific Islander, and 7.43% African American.

2.3 Stimuli Creation

We chose the topic of climate change issue for the following reasons. Climate change has been a controversial issue both domestically and internationally that attracts considerable amount of public attention (IPCC 2014; McCright and Dunlap 2011; Nisbet and Myers 2007). Although much scientific evidence suggests the happening of climate change and global warming (IPCC 2014), polarized voices come from mass media and celebrities, especially politicians, endorsing different sides regarding the existence of climate change, the urgency of taking measures, and what measures to take, if necessary at all (Nisbet and Myers 2007). Past research has also found the occurrence of HME on climate change issue (Feldman et al. 2015).

Source manipulation appeared in three places on the screenshot of the news article (see Fig. 1). Specifically, for human-written news, under the article title, it says "By Andrew C. Revkin," a name made up by the researchers. Below the article where introduction and contact of the author often appear, it writes "Andrew C. Revkin is a correspondent for The New York Times" and "Email: ARevkin@nytimes.com; Twitter:

@ARevkinNYT." For news purported as robot-written, changes were made accordingly as shown in Fig. 1.

The New York Times

Science

Here's What We Can - and Can't - Say About Climate Change and Hurrican Matthew

By New York Times News Robot
NEW YORK, October 07, 2016

As Hurricane Matthew continues to churn through the Atlantic, leaving more than 260 dead in the Caribbean and threatening the Florida coast, there's an uncomfortable question we should wrestle with: are hurricanes getting worse because of human-induced climate change?

According to a US government report, global warming is expected to make hurricanes produce about 20% more rain near the eye of the storm. By 2100, tropical hurricanes are expected to be 2% to 11% more intense because of global warming, citing UN data.

Many environmental groups contend that industrial gases are magnifying the heating powers of the sun. Al Gore, in a speech in Washington, said "global surface temperatures could increase from an average of 2°F to 8°F over the next century, a rate unseen on this planet for at least the last 10,000 years."

On the other hand, however, over 31,000 scientists have signed on to a petition saying humans aren't causing global warming. More than 1000 scientists signed on to another report saying there is no global warming at all.

Documented data showed that the temperature of the planet has essentially been flat for 18 years since 1997. The planet was cooling from 1940-1975 and the upswing in temperature afterward only lasted 22 years till 1997. And the loss of Arctic ice seemed not to be true, as Arctic ice is actually up 50% since 2012 as the European Space Agency satellite CryoSat measurements show, and has been growing steadily these years according to data reported by National Snow & Ice Data Center.

New York Times News Robot is an advanced artificial intelligence writing algorithm

Email: AINewsBot@nytimes.com; Twitter: @NYTNewsBot

Fig. 1. Sample stimulus.

To enhance the external validity, news articles were composed of excerpts from existing news reports on climate change published on mainstream media outlets such as *CNN*, *Los Angeles Times*, *Townhall*, etc. In the news stimuli, factual information endorsing each side was equally presented to maintain a neutral standing. For example, the side in favor of climate change activism was backed up with research findings suggesting taking immediate measures to stop climate change. The side against climate change activism was endorsed by including findings denying the happening of climate change or suggesting the unnecessity of human intervention. A short spot news was first created with the above-mentioned criteria being met, based on which we created the longer interpretive news article by adding elaborations derived from the factual information. The spot news article ended up with 236 words whereas the interpretive news had 681 words. Two news articles were pretested and found to be different in perceived amount of interpretation but are identical in terms of their neutral standing.

2.4 Measurement

Manipulation Check. Participants were asked to answer "The news you just read was written by" by choosing among "human," "robot," and "I don't know." Only those answered this question correctly were remained for further data analyses.

Measure of Key Variables. Unless specified, all items were measured on a 7-point Likert-type scale.

News Slant Extremity. News slant extremity is the perceived *extremity* of the news' slant or the *magnitude* of bias and was measured with the same method as in Gunther and Schmitt (2004). Participants were first asked three questions regarding the standing of the news by choosing one of the three choices, (a) pro-action, (b) against-action, (c) neutral. For those who answered neutral, they were coded as "0" on this question; for those who chose (a) or (b), a follow-up question was asked regarding the degree to which the news story was endorsing the side they chose on a 7-point Likert scale. Values on this follow-up question was remained as the score one had on that question. Final news slant extremity score was calculated by averaging the scores of these three questions (Cronbach's $\alpha = .84$).

Bias Perception. Bias perception was measured with two items "How much do you think the news is (a) biased; and (b) objective on reporting climate issues" (Cronbach's $\alpha = .79$).

News Credibility. News credibility was measure by measures adapted from Appelman and Sundar's study (2016). Participants were asked to answer questions "How much do you think the news you just read is (a) accurate, (b) authentic, and (c) believable" (Cronbach's $\alpha = .94$).

Source Trustworthiness. Source trustworthiness was measured with 5 semantic differential items adapted from McCroskey and Teven (1999), Ohanian (1990) that are applicable to both human and robot writer such as "Undependable - Dependable" and "Dishonest - Honest" (Cronbach's $\alpha = .93$).

Belief in Machine Heuristic. One's belief in machine heuristic was measured with the following three items "Compared with human, robot (a) has no intentions, (b) is more objective, and (c) has no bias" (Cronbach's $\alpha = .74$).

Novelty. Novelty was measured with 10 items adapted from Wells et al. (2010), and the originality subscale from Sundar et al. (2014), such as "I found reading the news written by the news bot (writer) to be a novel experience" (Cronbach's $\alpha = .94$).

Eeriness. Eeriness was measured by the 8-item semantic differential scale developed by Ho and MacDorman (2010), with items such as "reassuring - eerie" and "predictable - thrilling" (Cronbach's $\alpha = .90$).

Emotional Involvement. Emotional involvement was measured with three items adapted from Matthes (2013). Participants were asked to indicate their agreement with the following statements, "Reading this news aroused my feelings," "I emotionally reacted to the issue of climate change while reading," and "I feel really emotionally involved while reading this news" (Cronbach's $\alpha = .92$).

Issue Involvement. Issue involvement was measured with seven items adapted from the value-relevant involvement scale developed by Cho and Boster (2005), such as "The values that are the most important to me determine my stand on climate change issues" (Cronbach's $\alpha = .89$).

Attitudes Towards Climate Change Activism. Participants' attitude towards climate change activism was assessed with three items, "In terms of whether human has caused climate change, which side do you take?", "In terms of the seriousness of climate change, which side do you take?" and "In terms of how urgent human should take any actions, which side do you take?" with higher score meaning more activist attitude (Cronbach's $\alpha = .96$).

Attitude Extremity. Attitude extremity score was calculated by folding the attitude measure, specifically by subtracting 4 (the middle point on a 7-point scale) from one's attitude score than taking the absolute value, which varies on a 0 to 3 scale.

3 Results

3.1 Manipulation Check

After reading the assigned news article, participants answered the question who the news writer was by choosing from (a) A human writer, (b) news robot, (c) I don't know. Among 212 participants, 175 correctly identified the purported identity of the writer. Therefore, only these 175 participants were included in further analyses.

3.2 Effects of Robot News Writer

In response to the overarching question of this study, i.e., whether news writer's identity would cause any perceptual differences among news readers, multivariate analysis of covariance (MANCOVA) test was conducted with independent variable being news writer identity (robot vs. human), dependent variables being perceived emotional involvement, novelty, eeriness, news slant extremity, bias perception, news credibility, and source trustworthiness. Individual differences in issue involvement, attitudes to climate change issue, attitude extremity, and news type, were controlled for throughout the analyses in this study if not specified.

Writer's identity had an effect on the perceptual variables, Wilks' $\Lambda = .68$, F (8, 157) = 9.29, $p < .001$, partial $\eta^2 = .32$. Following univariate analyses results showed that writer identity had an effect on perceived novelty and news credibility. Reading robot-written news was perceived as more novel ($M = 4.78$, $SE = .16$) than reading human-written news ($M = 3.44$, $SE = .16$), F (1, 164) = 36.17, $p < .001$, partial $\eta^2 = .18$. In terms of news credibility, robot-written news was perceived as less credible ($M = 4.23$, $SE = .18$) than human-written news ($M = 4.78$, $SE = .18$), F (1, 164) = 4.44, $p = .04$, partial $\eta^2 = .03$. Although we found no effect of writer identity on other measures of HMP, hypotheses on news writer's effects on HMP suggest the existence of counterbalancing mechanisms which might drive the total effect or news writer on other HMP measures non-significant.

We conducted the following mediation analyses to probe the mechanisms as we have hypothesized using PROCESS Macro (Hayes 2013).

3.3 Mechanisms of Robot Writer Effect

Effects on Emotional Involvement. To test H1-3 about how robot writer influences readers' emotional involvement via counterbalancing mechanisms, the model shown in Fig. 2 was tested using PROCESS Macro (Hayes 2013), Model 5, with 5,000 bootstrapping samples and 95% bias-corrected confidence intervals (CI) estimated, with news writer as independent variable, emotional involvement as the dependent variable, eeriness and novelty as parallel mediators. Meanwhile, in order to test H1, one's belief in machine heuristic was tested as the moderator of the direct path, examining the effect of robot's inanimate nature on emotional involvement.

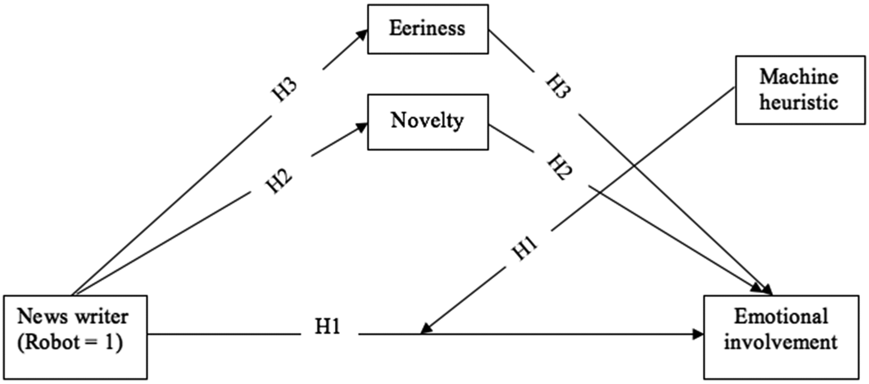


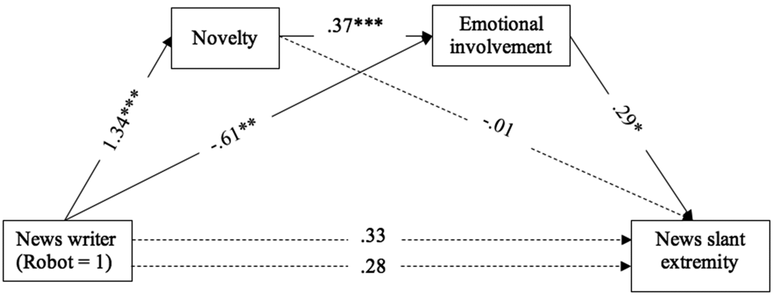
Fig. 2. Model of H1-H3.

In support of H1, results revealed a significant moderation effect of machine heuristic on robot's direct effect on emotional involvement, such that only for those who highly (belief > 6.17, one standard deviation above the mean) or moderately (belief $= 4.82$, mean) believed in machine heuristic, robot had a significant negative direct effect on emotional involvement (high believers: $B = -.67$, $SE = .29$, 95% CI [−1.25, −.10]; moderate believers: $B = -.43$, $SE = .21$, 95% CI [−.84, −.02]). For those who did not believe in machine heuristic (belief $= 3.46$, one standard deviation below the mean), effect of robot in reducing readers' emotional involvement was not statistically significant, $B = -.18$, SE = .27, 95% CI [−.71, .35].

In support of H2 that robot writer enhances readers' emotional involvement via perception of novelty, results showed a significant positive indirect effect of robot on emotional involvement via novelty, $B = .30$, $SE = .13$, 95% CI [.06, .57]. Disconfirming H3, indirect effect of robot writer via perceived eeriness was not statistically significant, $B = .04$, $SE = .08$, 95% CI [−.12, .21].

Indirect Effect of Robot Writer on HMP via Novelty. To test H4-6 regarding effects of robot writer via novelty and emotional involvement on readers' HMP, Model 6 in PROCESS Macro (Hayes 2013) was tested with news writer as independent variable, with novelty as first mediator, and emotional involvement as the second mediator in the mediation serial, with four different measures capturing HMP as dependent variables respectively.

News Slant Extremity. As shown in Fig. 3, in support of H4a that robot writer reduces emotional involvement which further reduces HMP, effect of robot writer via this specific path was significant, $B = -.18$, $SE = .11$, 95% CI [−.46, −.02]. Not supporting H5a, the indirect effect of news writer via novelty on news slant extremity was not significant, $B = -.01$, $SE = .21$, 95% CI [−.47, .39]. In support of H6a that novelty associated with robot writer enhanced emotional involvement, which further increased perception of news slant extremity, results showed that the indirect effect of robot writer via this specific path was significant, $B = .15$, $SE = .08$, 95% CI [.01, .35].

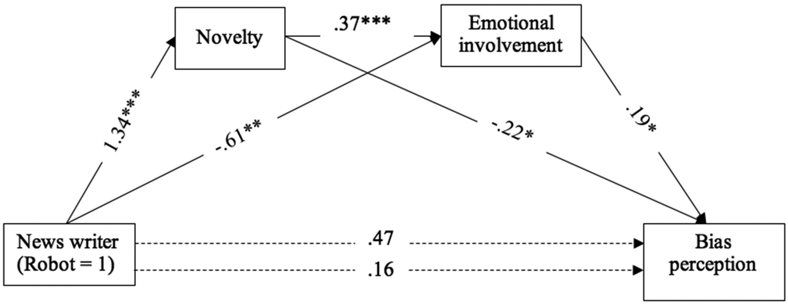


Note. *$p < .05$, **$p < .01$, ***$p < .001$. Numbers are unstandardized coefficients. The direct effect is .33, and the total effect is .28. *News type*, *issue involvement*, *attitude*, and *attitude extremity* were controlled as covariates.

Fig. 3. Mechanisms underlying effect of news writer on news slant extremity.

Bias Perception. A similar pattern was found for bias perception as well. As shown in Fig. 4, in support of H4b, the path that robot writer reduces emotional involvement which further reduces HMP was significant, $B = -.12$, $SE = .08$, 95% CI [−.33, −.00]. In support of H5b, readers perceived robot writers as more novel, which led to reduced bias perception, $B = -.29$, $SE = .14$, 95% CI [−.62, −.07]. In support of H6b, novelty associated with robot writer enhanced emotional involvement, which further increased perception of news slant extremity, $B = .10$, $SE = .06$, 95% CI [.00, .23].

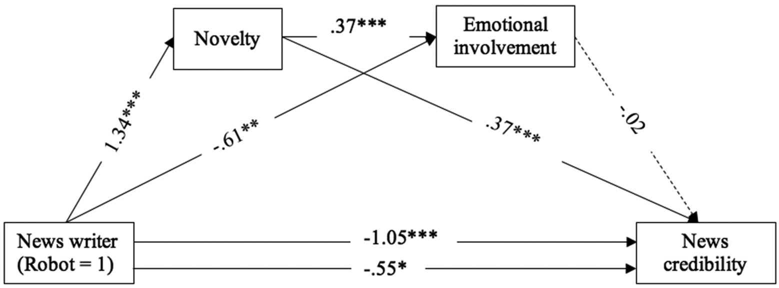
News Credibility. As shown in Fig. 5, only the path suggested by H5c that robot writer elicits novelty which further enhances one's evaluations of news credibility was found significant, $B = .49$, $SE = .17$, 95% CI [.21, .88]. The indirect path suggested by H4c that robot writer decreased emotional involvement and thus reduced perceived news credibility was not significant, $B = .01$, $SE = .08$, 95% CI [−.15, .19], neither was the path suggested by H6c that novelty associated with robot writer enhanced emotional



Note. *p < .05, **p < .01, ***p < .001. Numbers are unstandardized coefficients. The direct effect is .47, and the total effect is .16. *News type*, *issue involvement*, *attitude*, and *attitude extremity* were controlled as covariates.

Fig. 4. Mechanisms underlying effect of news writer on bias perceptions.

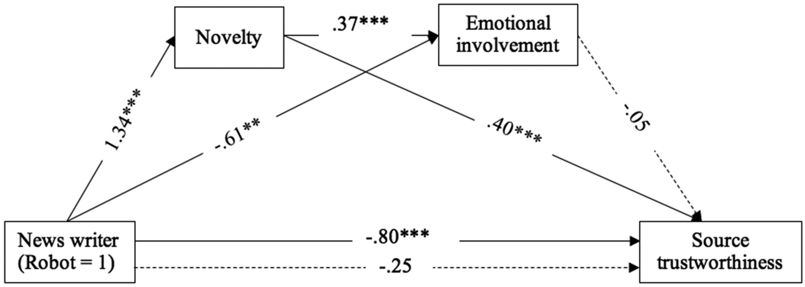
involvement, which further increased perceived news credibility, $B = -.01$, $SE = .06$, 95% CI [−.13, .13].



Note. *p < .05, **p < .01, ***p < .001. Numbers are unstandardized coefficients. The direct effect is -1.05 ($p < .001$), and the total effect is -.55 ($p < .05$). *News type*, *issue involvement*, *attitude*, and *attitude extremity* were controlled as covariates.

Fig. 5. Mechanisms underlying effect of news writer on news credibility.

Source Trustworthiness. Effect of robot writer on perceived source trustworthiness had the same pattern as that on perceived news credibility. As shown in Fig. 6, only the path suggested by H5d that robot writer elicits novelty which further enhances perceived source trustworthiness was found to be significant, $B = .54$, $SE = .14$, 95% CI [.30, .87]. However, the indirect path suggested by H4d that robot writer decreased emotional involvement and thus reduced perceived source trustworthiness was not significant, $B = .03$, $SE = .06$, 95% CI [−.08, .18], neither was the path suggested by H6d that novelty associated with robot writer enhanced emotional involvement, which further increased perceived source trustworthiness, $B = -.02$, $SE = .05$, 95% CI [−.12, .07].



Note. *$p < .05$, **$p < .01$, ***$p < .001$. Numbers are unstandardized coefficients. The direct effect is -.80 ($p < .001$), and the total effect is -.25. *News type*, *issue involvement*, *attitude*, and *attitude extremity* were controlled as covariates.

Fig. 6. Mechanisms underlying effect of news writer on source trustworthiness.

Indirect Effect of Robot Writer on HMP via Eeriness. To test H7 and H8 regarding effects of robot writer via perceived eeriness on readers' HMP, Model 6 in PROCESS Macro (Hayes 2013) was tested with news writer as independent variable, with perception of eeriness as first mediator, and emotional involvement as the second mediator in the mediation serial, with four different measures capturing HMP as dependent variables respectively. Disconfirming H7 and H8, neither of the hypothesized paths was statistically significant for all the dependent variables.

Table 1. Summary of hypotheses testing results

Hypotheses		
Supported	H1	Machine heuristic moderates Robot writer → emotional involvement
Supported	H2	Robot writer → novelty → emotional involvement
Unsupported	H3	Robot writer → eeriness → emotional involvement
	H4	Robot writer → emotional involvement → HMP
Supported	a	Robot writer → emotional involvement → news slant extremity
Supported	b	Robot writer → emotional involvement → bias perception
Unsupported	c	Robot writer → emotional involvement → news credibility
Unsupported	d	Robot writer → emotional involvement → source trustworthiness
	H5	Robot writer → novelty → HMP
Unsupported	a	Robot writer → novelty → news slant extremity
Supported	b	Robot writer → novelty → bias perception
Supported	c	Robot writer → novelty → news credibility
Supported	d	Robot writer → novelty → source trustworthiness
	H6	Robot writer → novelty → emotional involvement → HMP
Supported	a	Robot writer → novelty → emotional involvement → news slant extremity
Supported	b	Robot writer → novelty → emotional involvement → bias perception
Unsupported	c	Robot writer → novelty → emotional involvement → news credibility
Unsupported	d	Robot writer → novelty → emotional involvement → source trustworthiness
Unsupported	H7	Robot writer → eeriness → HMP (a, b, c, d)
Unsupported	H8	Robot writer → eeriness → emotional involvement → HMP (a, b, c, d)

3.4 Summary of Findings

As summarized in Table 1, robot writer induced less emotional involvement due to the machine heuristic held by readers that machines are free of intention and bias, which further reduced perceived news slant extremity and news bias, but not news credibility and source trustworthiness.

Results also showed that higher novelty elicited by robot writer, on one hand, led to evaluations that are more positive, i.e., lower bias perception, higher news credibility, and higher source trustworthiness, but made no difference on perceived news slant extremity. On the other hand, sense of novelty enhanced emotional involvement, which further increased perceived news slant extremity and bias perception, but did not change news credibility and source trustworthiness. Not as hypothesized, perception of eeriness was not found to play a role in news writer effect on all the measures of HMP.

4 Discussion

4.1 Discussion

Informed by research on the mechanisms of HME (e.g., Matthes 2013; Matthes and Beyer 2015) and MAIN model (Sundar 2008), this study examined the potential of applying AI in news production to reduce HMP and to evoke fairer evaluations of news given the pervasively held belief that machine is free of intentions and unbiased. Results showed that the label of robot writer had impact on readers' bias perception and credibility evaluation via multiple mechanisms.

One major finding of this study is that robot written news reduced readers' emotional involvement for those who believe in machine heuristic, suggesting machine heuristic was triggered by the identity cue, i.e., the ontological nature of the news source, which supports the MAIN model that for online news consumers, not only news content influences their judgments, but also other peripheral cues embedded in the interface by triggering relevant heuristics associated with those cues (Sundar 2008).

The lack of intention attribution for machine-written news led to less emotional involvement and therefore, less perceived news slant and bias perceptions. Such findings are in line with implications from extant studies that the influence intention attributed to the source might result in HMP by inducing affective reactance and arousal. The findings also corroborate that emotional involvement is a significant predictor of HMP independent from issue involvement (Matthes 2013; Matthes and Beyer 2015). As shown in the results, the more affectively involved the news readers were, the more they perceived the news as biased.

Another important finding of this study is the halo effect of novelty elicited by robot writer. As found in the current study, sense of novelty was associated with reduced bias perception, higher perceived news credibility and source trustworthiness. Consistent with existing findings on automated journalism (Gambino and Kim 2015) and findings in new technology adoption (Wells et al. 2010), sense of novelty associated with use of new technology can often enhance positive perceptions globally towards the experience with it.

Although the sense of novelty is associated with positive evaluations of the news, it seems to be a double-edged sword in the context of controversial news topics, which has not been quite tested in existing studies on automated journalism. On one hand, enhanced novelty promoted positive evaluations of the news and the source. On the other hand, sense of novelty also increased readers' emotional involvement, which further increased readers' news slant extremity and bias perception.

In contrast to findings in past studies, robot-written news was not perceived as eerier than that written by human writer as found in past studies (e.g., Gambino and Kim 2015). First, most studies that found eeriness associated with robotic agents were conducted with embodied robotic agents (Kätsyri et al. 2015; Seyama and Nagayama 2007). Second, perceived eeriness of machine-written news might only occur for news of certain topics (Gambino and Kim 2015). Given the current topic, i.e., climate change, is in the domain of hard science, machine writing on this topic may not be as uncanny as its writing on more personal topics (Gambino and Kim 2015).

This study used four measures to capture HMP. However, effects of news writer were found to be different on these four measures. As shown in the results, perceived news slant extremity and news bias (i.e., traditional HMP measures) were influenced by emotional involvement, whereas credibility measures (news credibility, and source credibility) were not found to be impacted by emotional involvement but more impacted by news writer and novelty directly. Findings suggest that evaluating news credibility and source trustworthiness is more than judging whether the news or the source is biased against or in favor of a side, but also involves other aspects (Arpan and Peterson 2008).

Besides theoretical contribution, this study, motivated by the goal of reducing HMP by applying AI in news production, also has practical implications on robot journalism. The overall findings seem to suggest that robot writer has two opposite effects that turn out to cancel each other out on HMP. Heightened sense of novelty leads to more favorable judgements overall but intensifies emotional involvement and therefore, HMP. To amplify the positive utility of robot writer in reducing HMP, news organizations might consider taking measures to trigger machine heuristic, enhance novelty, but tuning down the overall emotional involvement at the same time.

4.2 Limitations and Future Work

This study also has several shortcomings. First, we only tested the hypotheses with the climate change issue. However, existing research has suggested that perceptions of machine-written news could be contingent on news topics. For example, health news written by AI was less acceptable than finance news (Gambino and Kim 2015). In light of this, future research should expand the variety of news contexts.

Secondly, this study did not provide much explanation for why besides the positive effect brought by novelty, robot writer was perceived not as good as its human counterpart on credibility measures. As found in the study, no direct effect of robot was found to have on news slant extremity and bias perception but there was a negative direct effect of robot writer on the credibility measures. It suggests that being objective is not enough for news to be perceived as credible and source to be trustworthy. Future research should investigate the existence and the effect of other heuristics related to AI that matter in news production.

Thirdly, we did not address the potential impact of news organization. In the context of a controversial political issue, readers might question that the opposite party could manipulate the algorithm to intentionally seek or create information unfavorable to their side and in favor of the opponent's side (Carroll 2017), as partisans may naturally engage in defensive processing of incoming information (Gunther and Liebhart 2006). However, because we only had *the News York Times* in all the conditions, we could not assess the effect of news organization.

5 Conclusion

In conclusion, this study finds the use of AI in news production could be a double-edged sword in terms of its effect on HMP. On one hand, because of the widely held machine heuristic that machine is inanimate and therefore it must be free of intention, robot writer reduces readers' emotional involvement, which further reduces HMP. However, novelty associated with machine-written news, although enhances perception of credibility, also intensifies one's emotional involvement, which further increases bias perception. We suggest that by emphasizing the mechanical nature of robot, enhancing the novelty in the reading experience, and tuning down other factors that might intensify emotional arousal, news producers are able to reduce defensive reactions from readers, and foster positive evaluations of the news products.

References

Appelman, A., Sundar, S.S.: Measuring message credibility: construction and validation of an exclusive scale. Journal. Mass Commun. Q. **93**, 59–79 (2016)

Arpan, L.M., Peterson, E.M.: Influence of source liking and personality traits on perceptions of bias and future news source selection. Media Psychol. **11**, 310–329 (2008)

Arpan, L.M., Raney, A.A.: An experimental investigation of news source and the hostile media effect. Journal. Mass Commun. Q. **80**, 265–281 (2003)

Bellur, S., Sundar, S.S.: How can we tell when a heuristic has been used? Design and analysis strategies for capturing the operation of heuristics. Commun. Methods Meas. **8**, 116–137 (2014)

Brehm, S.S., Brehm, J.W.: Psychological Reactance: A Theory of Freedom and Control. Academic Press, San Diego (1981)

Carroll, E.C.: Making news: balancing newsworthiness and privacy in the age of algorithms. Georget. Law J., 1–42 (2017)

Chen, S., Chaiken, S.: The heuristic-systematic model in its broader context. In: Chaiken, S., Trope, Y. (eds.) Dual-Process Theories in Social Psychology, pp. 73–96. Guilford Press, New York (1999)

Cho, H., Boster, F.J.: Development and validation of value-, outcome-, and impression-relevant involvement scales. Commun. Res. **32**, 235–264 (2005)

Choi, J., Yang, M., Chang, J.J.: Elaboration of the hostile media phenomenon the roles of involvement, media skepticism, congruency of perceived media influence, and perceived opinion climate. Commun. Res. **36**, 54–75 (2009)

Cox, M.: The development of computer-assisted reporting. Paper presented to the Newspaper Division, Association for Education in Journalism and Mass Communication, Southeast Colloquium, March 1718, University of North Carolina, Chapel Hill (2000)

DeMott, J.: 'Interpretative' news stories compared with 'spot' news. Journal. Q. **50**, 102–108 (1973)

Dillard, J.P., Shen, L.: On the nature of reactance and its role in persuasive health communication. Commun. Monogr. **72**, 144–168 (2005)

Edwards, C., Edwards, A., Spence, P.R., Shelton, A.K.: Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. Comput. Hum. Behav. **33**, 372–376 (2014)

Elliot, A.J., Devine, P.G.: On the motivational nature of cognitive dissonance: dissonance as psychological discomfort. J. Pers. Soc. Psychol. **67**, 382–394 (1994)

Feldman, L., Hart, P.S., Leiserowitz, A., Maibach, E., Roser-Renouf, C.: Do hostile media perceptions lead to action? The role of hostile media perceptions, political efficacy, and ideology in predicting climate change activism. Commun. Res., 1–26 (2015)

Ferrey, A.E., Burleigh, T.J., Fenske, M.J.: Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. Front. Psychol. **6**, 249 (2015)

Gambino, A., Kim, J.: An algorithm wrote this? Psychological responses to AI news writers: eeriness, branding, and credibility. Poster presented at the 2015 Human-Computer Interaction Consortium (HCIC): Theory, Watsonville, CA, July 2015. http://comm.psu.edu/medialab/research-article/man-versus-machine

Gunther, A.C.: Biased press or biased public? Attitudes toward media coverage of social groups. Public Opin. Q. **56**, 147–167 (1992)

Grandpre, J., Alvaro, E.M., Burgoon, M., Miller, C.H., Hall, J.R.: Adolescent reactance and anti-smoking campaigns: a theoretical approach. Health Commun. **15**, 349–366 (2003)

Gunther, A.C., Chia, S.C.Y.: Predicting pluralistic ignorance: the hostile media perception and its consequences. Journal. Mass Commun. Q. **78**, 688–701 (2001)

Gunther, A.C., Liebhart, J.L.: Broad reach or biased source? Decomposing the hostile media effect. J. Commun. **56**, 449–466 (2006)

Gunther, A.C., Schmitt, K.: Mapping boundaries of the hostile media effect. J. Commun. **54**, 55–70 (2004)

Hayes, A.F.: Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. Guilford Press, New York (2013)

Ho, C.C., MacDorman, K.F.: Revisiting the uncanny valley theory: developing and validating an alternative to the Godspeed indices. Comput. Hum. Behav. **26**, 1508–1518 (2010)

Hwang, H., Pan, Z., Sun, Y.: Influence of hostile media perception on willingness to engage in discursive activities: an examination of mediating role of media indignation. Media Psychol. **11**, 76–97 (2008)

IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change/Geneva, Switzerland (2014)

Kätsyri, J., Förger, K., Mäkäräinen, M., Takala, T.: A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Front. Psychol. **6**, 390 (2015)

Kim, J., Sundar, S.S.: Human control or machine control-which do we trust? The role of control and machine heuristic in online information disclosure. Paper presented at Annual Conference of Association for Education in Journalism and Mass Communication, Minneapolis, MN (2016)

Kruglanski, A.W., Thompson, E.P.: Persuasion by a single route: a view from the unimodel. Psychol. Inq. **10**, 83–109 (1999)
Latar, N.L.: The robot journalist in the age of social physics: the end of human journalism? In: Einav, G. (ed.) The New World of Transitioned Media. TEICE, pp. 65–80. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09009-2_6
Marshall, S.: Robot reporters: A look at the computers writing the news. Journalism.co.uk, 12 March 2013. https://www.journalism.co.uk/news/robot-reporters-how-computers-are-writing-la-times-articles/s2/a552359/
Matthes, J.: The affective underpinnings of hostile media perceptions: exploring the distinct effects of affective and cognitive involvement. Commun. Res. **40**, 360–387 (2013)
Matthes, J., Beyer, A: Toward a cognitive-affective process model of hostile media perceptions: a multi-country structural equation modeling approach. Commun. Res., 1–24 (2015)
McCright, A.M., Dunlap, R.E.: Cool dudes: the denial of climate change among conservative white males in the United States. Glob. Environ. Change **21**, 1163–1172 (2011)
McCroskey, J.C., Teven, J.J.: Goodwill: a reexamination of the construct and its measurement. Commun. Monogr. **66**, 90–103 (1999)
Miller, C.H., Lane, L.T., Deatrick, L.M., Young, A.M., Potts, K.A.: Psychological reactance and promotional health messages: the effects of controlling language, lexical concreteness, and the restoration of freedom. Hum. Commun. Res. **33**, 219–240 (2007)
Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Pantheon, New York (2006)
Mori, M.: The uncanny valley. Energy **7**, 33–35 (1970)
Moriguchi, Y., Negreira, A., Weierich, M., Dautoff, R., Dickerson, B.C., Wright, C.I., Barrett, L.F.: Differential hemodynamic response in affective circuitry with aging: an FMRI study of novelty, valence, and arousal. J. Cogn. Neurosci. **23**, 1027–1041 (2011)
Nisbet, M.C., Myers, T.: The polls—trends twenty years of public opinion about global warming. Public Opin. Q. **71**, 444–470 (2007)
Ohanian, R.: Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. J. Advert. **19**, 39–52 (1990)
Örnebring, H.: Technology and journalism-as-labour: historical perspectives. Journalism **11**(1), 57–74 (2010)
Rains, S.A.: The nature of psychological reactance revisited: a meta-analytic review. Hum. Commun. Res. **39**, 47–73 (2013)
Schomaker, J., Meeter, M.: Short-and long-lasting consequences of novelty, deviance and surprise on brain and cognition. Neurosci. Biobehav. Rev. **55**, 268–279 (2015)
Seyama, J.I., Nagayama, R.S.: The uncanny valley: effect of realism on the impression of artificial human faces. Presence Teleoperators Virtual Environ. **16**(4), 337–351 (2007)
Stein, J.P., Ohler, P.: Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. Cognition **160**, 43–50 (2017)
Shen, L.: The effectiveness of empathy-versus fear-arousing antismoking PSAs. Health Commun. **26**, 404–415 (2011)
Sundar, S.S.: The MAIN model: a heuristic approach to understanding technology effects on credibility. In: Metzger, M.J., Flanagin, A.J. (eds.) Digital Media, Youth, and Credibility, pp. 73–100. The MIT Press, Cambridge (2008)
Sundar, S.S., Nass, C.: Conceptualizing sources in online news. J. Commun. **51**, 52–72 (2001)
Sundar, S.S., Tamul, D.J., Wu, M.: Capturing "cool": measures for assessing coolness of technological products. Int. J. Hum. Comput. Stud. **72**, 169–180 (2014)

Tsfati, Y., Cohen, J.: The influence of presumed media influence on democratic legitimacy: the case of Gaza settlers. Commun. Res. **32**, 794–821 (2005)

Vallone, R.P., Ross, L., Lepper, M.R.: The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. J. Pers. Soc. Psychol. **49**, 577–585 (1985)

van Dalen, A.: The algorithms behind the headlines: how machine-written news redefines the core skills of human journalists. Journal. Pract. **6**, 648–658 (2012)

Van der Kaa, H., Krahmer, E.: Journalist versus news consumer: the perceived credibility of machine written news. In: Proceedings of the Computation+Journalism Conference, pp. 24–25. Columbia University, New York (2014)

Weierich, M.R., Wright, C.I., Negreira, A., Dickerson, B.C., Barrett, L.F.: Novelty as a dimension in the affective brain. Neuroimage **49**, 2871–2878 (2010)

Wells, J.D., Campbell, D.E., Valacich, J.S., Featherman, M.: The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective. Decis. Sci. **41**, 813–843 (2010)

Wired: Robots wrote this, 29 August 2006. https://www.wired.com/2006/08/robots-wrote-this/. Accessed 2 Nov 2018

Wirth, W.: Involvement. In: Bryant, J., Vorderer, P. (eds.) Psychology of Entertainment, pp. 199–213. Erlbaum, Mahwah (2006)

Worchel, S.: The effect of three types of arbitrary thwarting on the instigation to aggression. J. Pers. **42**, 300–318 (1974)

Zillmann, D.: Excitation transfer in communication-mediated aggressive behavior. J. Exp. Soc. Psychol. **7**, 419–434 (1971)

Courting the Visual Image: The Ability of Digital Graphics and Interfaces to Alter the Memory and Behaviour of the Viewer

Damian Schofield(✉)

State University of New York, Oswego, NY 13126, USA
schofield@cs.oswego.edu

Abstract. An intrinsic connection exists between humans and the memories they create; they define who we are, where we came from and our accomplishments and failures. However, decades of research has shown how fragile human memory can be. Almost all human computer interfaces involve vision and most rely on vision as the primary means of passing information to the user [1]. It is worth considering that perhaps this specific form of media interaction requires special care and attention due to its inherently persuasive nature, and the undue reliance that the viewer may place on information presented through a (potentially photo-realistic) visualisation medium. Their influence on human memory and behaviour cannot be underestimated.

This paper will introduce research undertaken by the author over the past 25 years that has experimented with, and examined a range of visual based presentation technology into courtrooms all over the world. Courtrooms are environments where the decisions made (based on human memory and comprehension) can significantly affect the lives of others. This paper describes research undertaken to assess the effect of visual technology on users (in particular their memory and decision making abilities) and describes some of the issues raised by the experimental results. The work presented in this paper connects psychological research with human cognitive and perceptual processes and limitations, to allow the evaluation and optimisation of visual interfaces. The paper concludes with a discussion of the potential benefits and problems of designing interactive visual technology when considering the impact on human cognition.

Keywords: Visual images · Computer graphics · Psychology · Evidence · HCI

1 Introduction

In a modern courtroom, the presentation of forensic evidence by an expert witness can bring about the need for arduous descriptions by lawyers and experts to get across the specific details of complicated scientific, spatial and temporal data. Technological advances have also meant that experts have had to develop new ways to present such increasingly complex evidence in court. As courts begin see an increasingly use multi-media and cinematic displays, this has profound implications for the legal processes taking place that are intrinsically tied to the application of such technology. Courtrooms

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 325–344, 2018.
https://doi.org/10.1007/978-3-319-91238-7_27

are environments where the decisions made (based on human memory and comprehension) can significantly affect the lives of others.

Digital visual evidence presentation systems (including digital displays, computer-generated graphical presentations and three-dimension virtual simulations) have already been used to present evidence and illustrate hypotheses based on scientific data in many jurisdictions [2]. These visual tools can be used to present evidence and illustrate hypotheses based on scientific data, or they may be used to depict the perception of a witness, and to illustrate what may have occurred (seen from a specific viewpoint) during a particular incident. Digital reconstruction technology may also be applied in a court to explore and illustrate 'what if' scenarios and questions, testing competing hypotheses and possibly exposing any inconsistencies and discrepancies within the evidence [3] Examples are shown in Fig. 1.

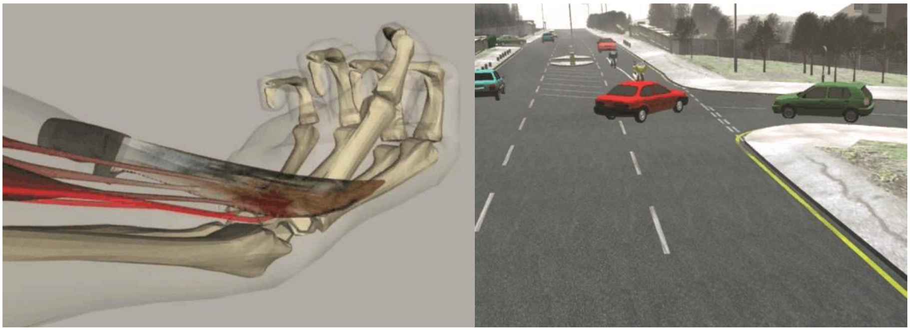


Fig. 1. Computer-generated evidential images from a suicide inquest and a road traffic accident.

The use of such computer-generated presentations in a court is sometimes viewed as nothing more than the current manifestation of the illustration of evidence and visualisation in a long history of evidential graphics used in litigation [2]. However, there are a number of fundamental implications inherent in the analysis of visual interfaces over textual/oral mediums and computer animations and interactive virtual simulations are potentially unparalleled in their capabilities for presenting complex evidence [4, 5]. The use of such enabling visualisation technology can affect the manner in which evidence is assimilated and correlated by the viewer. Their influence on human memory and behaviour cannot be underestimated. In many instances, visual media can potentially help make the evidence more relevant and easier to understand. In other cases it may be seen to be unfairly prejudicing a jury [3, 6, 7].

It is useful at this point to clarify the terms used to describe such technology. The standard form of evidence from such virtual environments usually consists of a series of still images and animations. In this context, the term 'computer animation' is often misused to describe an animation created from a virtual environment that is not based on the laws of physics, but is still represented as 'simulating' a given event.

The terms 'animation', 'scientific animation' and 'simulation' have had specific definitions in the reconstruction community for many years [8].

- ‘*Animation*’ is a general term describing ‘any presentation which consists of a series of graphical images being sequentially displayed, representing objects in different positions from one image to the next, which implies motion’ (note: all the quotes in this paragraph are from Grimes). This term may be used to describe a presentation consisting of artist renditions or illustrated moving graphics, sometimes referred to as a ‘cartoon animation’.
- The phrase ‘*Scientific Animation*’ is consequently used to describe a more technically based presentation, and is defined as ‘a computer animation that is based on the laws of physics and the appropriate equations of motion’. Velocities and positions are time integrals of the acceleration data, and the objects and environment in a scientific animation are properly and consistently scaled.
- In the reconstruction community, a ‘*Simulation*’ is often defined as being based on the laws of physics and containing specific underlying equations. A simulation goes further than a scientific animation, and can be further defined as ‘A model that predicts an outcome. The model may be a physical or a mathematical model, but the significant property is that a simulation predicts a future result’ – for example a computation fluid dynamics model to predict smoke flow through an environment.

In summary, an ‘animation’ may only be illustrative or demonstrative evidence, whereas a ‘scientific animation’ is more technical, and relies upon scientific laws, and thus might be categorised as substantive evidence. A ‘simulation’ is more predictive in nature, and consists of data or forecasts that are usually created via a computer program.

The vast majority of people called to be on a jury have grown up watching visual media on screens: cinemas, televisions, computers and even on their mobile phones. Research has shown that many people tend to believe what they see in the mass media and merge mediated fictions into their beliefs about the world [9, 10]. The cognitive default when viewing visual media is to believe what is seen, only later engaging in the effort needed to suspend or reject belief. Pictures on a screen which move tend to be even easier to believe. These are usually more engaging and entertaining, and hence decrease the mental resources of the viewer which are available for doubt [11].

However, audiences receive visual information differently when they watch it on a screen compared to when they see it in real space and these differences can affect everything they see. This difference can be described in two contexts, firstly the way the screen frames and what the viewer sees; a physical border that limits and creates new relationships between the elements displayed inside it. Secondly, the visuals presented also act as a carrier of personal and cultural associations [12].

This ability of viewers to place undue reliance on visual evidence has profound implications for the use of any form of animated visual digital technology to present evidence in courtrooms [13]. The potential life-and-death weight of the issues means that those undertaking this important civic duty by acting as jurors need to able to make objective assessment of the evidence before making their decisions. The way the evidence presented must be probative, not unfairly prejudicial.

This paper gives a brief background to the use of animated visual digital technology in courtrooms and describes past research that has been undertaken to examine the impact any form of animated visual presentation has upon members of the jury. The paper also provides an extensive discussion of the issues arising from the use of animated visual digital presentation, specifically those based on video game technology, in courtrooms. This includes an analysis of the emotional and psychological impact of the use of this technology, the creation of narrative through interaction with virtual environments and the influence of viewer perspective on the user experience. The paper concludes by comparing the advantages and disadvantages of using such a medium to present evidence.

2 Visual Evidence

Modern culture is dominated with images whose value may be simultaneously overdetermined and indeterminate, whose layers of significance can only be teased apart with difficulty. Different academic disciplines (including critical theory, psychology, education, media studies, art history, semiotics, etc.) have been developed to help explain how audiences interpret this visual imagery [2, 14]. Improvements in forensic science have led to an increasing amount of complex, technical evidence being presented in courts. The issues in question can be extremely complicated and difficult to explain without some form of graphical representation. A further survey by the American Bar Association found that members of a jury are often confused, bored, frustrated and overwhelmed by technical issues or complex facts [15]. Other research has indicated that the attention span of the average member of a jury in a court is, on average, only seven minutes (the same as the average attention span of a visitor to a website) [16–18].

Any visualisation or graphic can potentially be a valuable aid to help construe and convey a large amount of complex information. An American judge, Rubin highlighted the problem of retaining the interest of the jurors when he stated [19]:

> *'It isn't difficult to tell when jurors have lost interest … Such wandering attention is much less likely in a paperless trial, because the evidence is presented in a format jurors are used to watching … I have noticed repeatedly that when a document is displayed on the monitors, the jurors sit up and pay attention. Such attention is far greater than that given to a document which they cannot see as it is being discussed by the attorney and the witnesses …'.*

This comment illustrates the perceived need to reduce lengthy verbal explanations and increase the use of visual tools for a media-literate modern audience. This, in turn, offers a lawyer the possibility of improving the capacity of a jury to retain the evidence they present, to maintain their interest in the proceedings, and to allow the jury to understand the nature of the case more fully [20, 21].

In courtroom settings, static images such as diagrams, plans and charts have been traditionally used to explain the testimony of an expert witness. A number of modern expert witnesses now provide animated multimedia explanations illustrating their evidence. Such forensic animations or virtual reconstructions can be seen as an advance due to their unique ability to visually illustrate and animate visually the passing of time. This extra temporal dimension can be extremely useful when explaining a chronological sequence of events, such as the reconstruction of the occurrences leading up to a vehicle collision. In this case the dynamic movement of the vehicles involved in the collision may be dependent on complicated engineering or mathematical principles that are potentially difficult to explain to members of the jury – but easy to understand when visually represented in an animated, photo-realistic reconstruction [22].

A particularly relevant aspect of the technology under discussion is the ability to visualise unseen or imaginary environments. In a courtroom context this manifests itself as the ability to visualise evidential information that may not be naturally or readily visible to the naked eye. The virtual camera can break free of the physical restrictions restraining real world cameras and show processes that occur on too large or too slow a scale (from the unfolding of a storm to the replication of DNA), or processes that are occluded by other objects [5, 23–25].

The precise effect that this increasing reliance on visual media over the more traditional mechanism of oral presentation is having on members of a jury, witnesses and other viewers in the court is not currently known. Concerns are beginning to be articulated that the use of computer-generated visualisation technology can distort perceptions, memories, attitudes and decision making in the court. Some research work, previously undertaken in the USA, has examined how members of a jury retain details in their memory from different forms of evidence:

a. Research evidence has also shown that individuals are more likely to be persuaded if the arguments are supported by such visual aids [26, 27].
b. One study showed that the average person retains 87% of information presented visually, but only 10% of information presented orally [28].
c. Another study showed that the average person retains 65% of information presented visually and 15% of that presented orally [29].
d. A further survey showed that individuals will retain twice the amount of information when using a visual presentation, as distinct to an oral presentation [30].

When the evidence is animated, the improvement in memory retention is even more apparent: another survey revealed that members of a jury will retain an increase of 650% of information when presented with presentations using a form of computer animation [31]. The Visual Persuasion Project (New York School of Law) identified a number of issues and problems with the use of visual technology [32]. These issues, along with many others will be expanded upon and addressed later in this paper.

Kassin and Dunn undertook two experiments to assess the effects of computer-animated displays on mock jurors [33, 34]. In both experiments, participants watched a trial involving a dispute over whether a man who fell to his death had accidentally slipped or jumped in a suicide. They observed that when the claimant and defence used an animation to depict their own partisan theories, participants increasingly made judgments that contradicted the physical evidence, suggesting that computer-animated displays can have a greater effect than oral testimony. More recent research by Dunn and others examined the prejudicial effects of computer-generated animations in more detail [35]. This research work offered varying results, depending on the familiarity of the viewers with the scenarios depicted. These experiments also showed that the juror's expectations about the persuasiveness of animations were at odds with the animations' actual influence on jurors' verdicts.

There is little argument regarding the effectiveness of animated visual media as a tool for communication and knowledge transfer. The technology can offer significant benefits over traditional static (photographic) or moving (film) media captured in the real world. The rendered images from virtual worlds are not bound by the limitations of available lighting; they can avoid extraneous information, focusing only on salient evidential items; and they can be colourful, animated and lively enough to guarantee the attention and engagement of the viewer [24, 36–38].

3 Psychology of Images

Decades of research has shown how fragile human memory can be. Early memory experiments demonstrated how through misinformation and suggestibility we could influence and change the memory of others [39]. More recent work has led to many theories regarding behaviour, and many theories and guidelines are now available to show how human decision making can be influenced by external stimuli [40, 41]. A large volume of research output exists in this field, but the majority of the research work has focused on language and its ability to influence readers and listeners.

There are a number of concerns relating to the viewer's understanding of the visual evidence, based on the issues described above. These are identified and classified below. These are areas of concern that should be considered whenever a visual display or graphical user interface is designed.

3.1 Cognitive Biases

A cognitive bias is a systematic error in thinking that affects the decisions and judgments that people make. Some of these biases are related to memory. The way a person remembers an event may be biased for a number of reasons and that in turn can lead to biased thinking and decision-making. Other cognitive biases might be related to problems with attention. Since attention is a limited resource, people have to be selective about what they pay attention to in the world around them [42].

A cognitive bias is a type of error in thinking that occurs when people are processing and interpreting information in the world around them. The human brain is powerful but subject to limitations. Cognitive biases are often a result of the brain's attempt to simplify information processing. They are rules of thumb that help us make sense of the world and reach decisions with relative speed.

When users are making judgments and decisions about any interactive device or visual image, they like to think that they are objective, logical, and capable of taking in and evaluating all the information that is available. Unfortunately, these biases sometimes lead to poor decisions and bad judgments.

The sheer complexity of the interactive devices around the modern user, the overwhelming bombardment of visual images that surrounds everyone and the enormous amount of information in the environment, mean that users need to sometimes rely on mental shortcuts that allow them to act quickly [43].

Cognitive biases can be caused by a number of different things, but it is these mental shortcuts, known as heuristics, that often play a major contributing role. These biases are not necessarily all bad, however. Psychologists believe that many of these biases serve an adaptive purpose - they allow us to reach decisions quickly. This can be vital when facing a dangerous or threatening situation [42].

A few of the most relevant cognitive biases are described below:

- '***Framing Effect***': Involves drawing different conclusions from the same information depending on how the information is presented. For example given the following choices:
  - Option **A:** 200 of the 600 people will die
  - Option **B:** 400 of the 600 people will live

  Most people will choose Option **B,** humans choose a benefit over a loss.
- '***Anchoring***': The common human tendency to rely too heavily on the first piece of information presented. For example, a store is selling shoes at $150 and they are not selling. The store then offers them at $500; a week later the store has a sale and discounts the shoes to $200. They will now sell, as the customers were initially '*anchored*' to the $500 price and now only think in relation to that initial figure.
- '***Recognition vs Recall***': Recall is much more difficult than recognition, for example it is easier to recognise a friend than recall their face. When asked the question "*Who wrote the book Moby Dick?*" the answer can be hard to recall. But when asked "*Did Herman Melville write Moby Dick?*" the answer is easy to recognise.
- '***Suggestive Evidence***': The misinformation effect is very powerful, suggestive evidence can influence our memories. In a famous experiment two groups of people were shown the same video of a car accident. When asked what speed the cars **HIT** each other, the average speed was 30 miles per hour. When asked what speed the cars **SMASH** each other, the average speed was 40 miles per hour [20, 39].

The key point is that if we can affect user's memory, behaviour and decision making so much with text and voice information, imagine how much they are affected by visual images and interfaces.

3.2 Memory

Loftus has demonstrated that the memory of a viewer of an event can be biased by a wide variety of seemingly inconsequential factors [44, 45]. The results of Loftus's work can be extrapolated to predict that images and visualisations can possibly to lead to similar biases. Critical variables in how we view such images and interfaces may include the representation of depth, size, shape and colour. The question of how much detail or realism is needed in order for a image of diagram to be effective (i.e. believable) is considered crucial. Object recognition studies have shown that outline drawings can often be just as effective as colour photographs [46], but in other circumstances the interpretation of small details can be critical. Many researcher contend that media displays can occasionally create false memories [5, 7, 12, 37]. Studies have indicated that visual images can create pseudo-memories of an event and the memorability of what was seen can have no validity in fact [47].

3.3 Attitudes

Research has found that when people believe they have a sufficient volume of evidence, they feel more confident about making judgments, even when the information is irrelevant [48]. Images can provide just such an illusion of sufficiency. users are often more comfortable with visuals over oral presentation and discourse, and hence the visuals may be considered more believable. Many factors also influence the credibility of information presented, such as the gender of the presenter, their race, appearance, and socioeconomic circumstances. Visuals visualisation based on information presented have the potential to be more objective and to cause users to discount such factors [49–51]. The anonymous and abstract nature of a well-made visual presentation or interface (one which takes into account the issues discussed later in this paper) may help to remove any such bias or prejudice. On the other hand, a poorly made one may serve to emphasise any such differences.

3.4 Decision Making

Research on group decision making has found that once a group starts a communal discussion, many social and linguistic biases are exhibited, such as group polarisation, production losses and Grice's maxims (which are a way to explain the link between utterances and what is understood from them) [52]. Images can provide a shared memory or representation for users or a group of decision makers. Although this has the potential to reduce a number of social and linguistic biases, it is likely to increase others (for example, production loss). It is necessary to determine if the technology being used undermines critical reasoning; in other words, whether the display that is to be used supports or hinders decision making, and whether it affects the way in which users interact. A reconstruction often contains uncertain or inferred data, which may need to be represented in order for it to be understood by the viewer [53]. The communication and collaborative process between individuals will also be affected by the type and extent of the display and will also determine content, in as much as it might affect the way groups reach decisions [54, 55].

4 Recommendations for the Use of Evidential Images

The previous sections have described how information presented using computer generated visuals can be extremely advantageous, providing they are used appropriately. However, potential difficulties can occur from the use of this technology, and when the use of these images are examined in further detail, a number of issues and questions can arise. In a courtroom setting, the consequences of these problems cannot be underestimated, since errors, inaccuracies, misuse, tampering or bias within visual and graphical evidence are capable of leading to miscarriages of justice.

On reflection, many of the issues regarding the use of technology to generate explanatory visuals and graphical interfaces can be expressed as a list of advantages and disadvantages, as shown in Table 1.

Table 1. Advantages and disadvantages of using evidentiary visuals.

Advantages	Disadvantages
Such displays can provide an effective means of conveying complex evidence to the user. Visual memory has been found to be highly detailed and almost limitless, in contrast with memory for verbal material [56]. Images and diagrams have the potential to improve a viewer's ability to retain complex spatial and temporal data and hence increase the potential comprehension of complex information	The very fact that images and visuals impress themselves on the memory, and are persuasive and convincing, is also their greatest disadvantage: they can leave a strong impression on viewers. Moving images tend to mesmerize, and they can relax an individual's natural critical nature. This means that viewers are inclined towards a 'seeing is believing' attitude, as they do with television, potentially reducing the standards expected when critically assessing the information [20, 61]. Simulations can assume a 'hyper-real' character that eclipses the significance of the reality [62–64]
Visual media can provide an increase in the attention span of the viewer, since human attention is naturally drawn to animated images. Moving objects rank top on the hierarchy of methods to draw attention, which covers actions, objects, pictures, diagrams, the written word, and the spoken word [57]. A modern audience will more readily engage with audio-visual forms of communication, rather than relying solely on verbal modes of discourse	It is often difficult to represent uncertainty in images. Viewers often wrongly believe there is little or no margin of error in information presented in a visual form [53, 65]. Research studies have examined how to visualise uncertainty and provide non-prejudicial representations of uncertain information [5, 24, 53]
Computer displays can also act to help persuade users and inspire confidence. Studies comparing oral, textual, and static visual presentations to computer-generated animated presentations containing the same information found the animations to be more memorable [28–31]. This has implications not only for the retention of information, but also the weight given to the information viewed by the user [5, 58]. Also, visual, rather than verbal information, more readily activates the formation of an impression [59]	The flexibility of a computer-generated image also implies that they inherently contain the potential for tampering and alteration. Image quality does not equate with sufficiency, and the public's general knowledge that filmmakers can use computers to resurrect dinosaurs and create alien landscapes make allegations of digital alteration a potentially major issue when it comes to the weight placed on visual information [66]. The image creator must also be able to prove the accuracy of their reconstruction, both with reference to the original data, and to validate the development stages of the image itself [24]
Interactive displays also have the ability to provide the creator with an improved illustration of their arguments; the user can retrieve information instantly, and the display can be manipulated for better viewing. The user can zoom into an area of interest, examine a specific piece of information in more depth or create overviews of the whole display [24]	A party may intentionally create an image or virtual reconstruction that provides a favourable perspective to support a particular hypothesis, or unintentionally choose a viewpoint, perspective, illumination model or colour scheme that alters the appearance of the image or animation to work against the same hypothesis. This could create bias in the viewer, whether that is conscious bias (a form of tampering) or subconscious bias [34]. Designers and the creators of virtual reconstructions can learn much from the work of film and media theorists who continually strive to define the nature and functions of the media in which they work, particularly in relation to viewer perception and engagement [13, 67]
Increasing using of graphics and visuals can improve the efficiency of displays, thus time, as arguments and complex information are understood at a faster pace. The increase in efficiency because of the use of graphical display technology is a factor of the potential improvements in the speed with which complex information can be imparted to an audience, which therefore may shorten the length of a task [17, 60]	

A number of these potential issues along with recommendations to overcome them are discussed below. By their very nature, any recommendations and guidelines formulated are likely to be broadly defined and generic. Many of the recommendations offered below are little more than general suggestions that users of the technology be aware of these issues when involved in developing the types of images and reconstructions to be used in critical situations such as courtrooms [68]. Unfortunately, many of these recommendations have been ignored in the past when such technology has been used, and this may have been a contributing factor to the admissibility problems encountered when using this technology in certain jurisdictions [2].

4.1 Viewpoint (Field of View)

One issue with images used in courtrooms is how to correlate the viewpoint of a witness in a 'virtual' environment with the view from their physical position at the scene. For example, compare the problem of accurately replicating the 'physical world' view of the driver of the vehicle involved in a road traffic accident with the field of view of a camera in a virtual reconstruction. The driver has a wide field of vision through two eyes with differing levels of visual acuity across the field of view (for example there will be lower resolution vision at the periphery of the field of view compared to the current focus of attention), and the driver may also move their head around within the car to gain a better view. Whereas, animated driving simulations often rely on a fixed camera viewpoint within the vehicle [69].

Popular computer game titles provide a good example of distinct viewing configurations through various playing styles. Unreal Tournament (Epic Games) and the Halo Series (Bungee Studios) are examples that belong to a genre known as the First Person Shooter (FPS); distinguished by a first person perspective (egocentric) that renders the game world from the visual perspective of the player character. Grand Theft Auto (Rockstar Games) and Tomb Raider (Core Design) are games that belong to a genre known as the Third Person Shooter (TPS); this is a genre of video game in which an avatar of the player character is seen at a distance from a number of different possible perspective angles (exocentric). In any forensic reconstruction (as in any computer game), the choice of the viewing perspective may have significant effect on the way an image is interpreted by the viewer. Changing the viewing perspective can potentially alter which 'character' in an evidence presentation a viewer identifies with, or aligns themselves with [5, 22, 70].

However, research has shown that in positioning the virtual camera to represent a specific subject's viewpoint can actually incline the viewer to attribute less responsibility to the actor whose point of view the simulation leads them to adopt and more responsibility to other actors or to the circumstances. Cognitive psychologists call this actor-observer bias, and it is a bias since this point of view ought to be irrelevant to judgements of responsibility. This actor-observer effect is well established in the social psychology literature [71].

Designers of images, interfaces and reconstructions ought to study film-making techniques for two reasons. First, to be able to achieve the same effects as a film-maker; perhaps getting the viewer to identify emotively with a particular character in a reconstruction to enhance the power of the message. More importantly, a designer may wish to eliminate these effects and to remove the emotive content to provide an objective, understandable view of a forensic data set, with no distracting emotive attachment. An awareness of the ways in which the viewer can be manipulated (for example, through the use of egocentric and exocentric viewpoints) is essential [5, 20].

4.2 Realism and Level of Abstraction

The environment surrounding any particular scene that is to be reconstructed may be included within image presented in court. For example, a model may not only show the location of items or objects that form part of the evidence, but also the position of such items in relation to nearby buildings or other environment features, and these items may be placed and animated within a chronology of events or a time frame.

The realism of 'virtual' environments continues to improve. A number of researchers have noted an interesting observable fact relating to the realism in animated imagery, where many viewers become 'unnerved' by images of humans which are close to, but not quite real. This phenomenon has become known as the 'uncanny valley', because of the sharp dip seen in a graph of familiarity against the perception of reality [72]. As computer-processing power increases and software tools develop, it is natural to assume that it will be possible to achieve a similar level of realism to that used in photorealistic animated Hollywood movies within the computer-generated images used in courtrooms and other environments where critical decisions are made.

Virtual objects in a court reconstruction can be modeled with varying degrees of accuracy to explain and visualise the certainty, believability and veracity of the information related to that object. For example, trajectories of bullets are often displayed as cones or wedges within shooting reconstructions to show a range of possible positions of the weapon, instead of showing a single definitive line trajectory [53].

However, the mixing of visual metaphors and modes may be potentially disorientating to some viewers. Combining abstract data representations in photo-realistic environments may provide an unnatural experience for the viewer. Fielder has commented on the way members of juries may be misled by the use of visual metaphors and abstract representations in forensic animations. Combining different degrees of photorealism and expecting the viewer to draw additional information from a number of abstract representations in the virtual environment may overload the viewer and potentially add to their confusion, rather than increasing their comprehension of the evidence that is presented. In a forensic graphics context, many presentations used in court currently rely on fairly abstract representations [38]. However, as technology develops, the development of increasingly photorealistic evidence reconstructions becomes ever more likely. Increasing use of the rendering of photorealistic components into images may lead to instances where viewers may be lulled into a 'seeing is believing' attitude, causing a potential relaxation of their critical faculties [48, 73, 74].

Hence, careful use of visual metaphors is essential. Thought needs to be given to each abstract data representation in the environment and how that will be perceived by the potential audience. Experience and literature from disciplines such as psychology, cultural and critical theory, visual media, art history, education and such like can inform how abstract (and realist) representations are interpreted by the viewer. This in turn provides for what the viewer remembers and understands from the images presented to them [5, 11, 43]. An example of a realism issue is shown in Fig. 2.

Fig. 2. A photo from a murder scene is on the left, a computer-generated image is on the right.

4.3 Media Mode

It is rare that one form of media will be sufficient to explain fully every facet of a complex process or case to a viewer. It is important to choose an appropriate representation mode (photographs, text, video, graphics etc.) for the information that needs to be presented. Additional data may be included and displayed within any virtual environment; for example, location based statistical or analytical data may be displayed, calculation and test results may be presented in a visual format, and original documents and photographs can be linked to three-dimensional virtual objects.

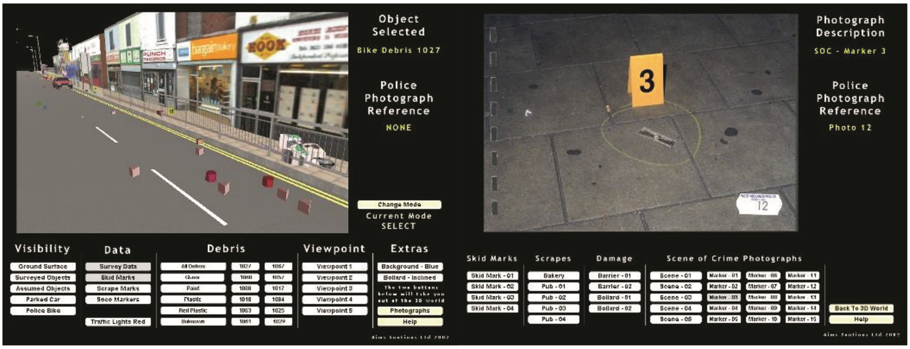


Fig. 3. A real time evidence display system (a murder trial) showing virtual and real imagery.

A reconstruction developed for the West Midlands Police in the United Kingdom (by the author) uses real-time virtual reality technology (Fig. 3). The user can pass the mouse over any relevant piece of evidence and view textual data about that item, and by clicking on any particular object in the virtual world (in this case, mainly items of vehicle debris), relevant crime scene photographs and evidential data will be displayed. The linking of 'real' evidence to spatially contextualised hotspots in a virtual environment has the potential to provide an effective mechanism to help the viewer understand the spatial relationship of the evidence. Such a multi-modal approach can be very effective, and different media may also be used as a device to help to retain the attention of the viewer and thereby increase understanding [22, 24].

It is important to be aware of the effect that the particular form of media being displayed will have on the viewers, and also to have an appreciation of the context in which it will be experienced by the user. The pedagogical effect of transitions between the forms of media should be considered. For example, switching between a virtual, rendered image of a crime scene and a real crime scene photograph may cause confusion in the viewer as they attempt to correlate evidence between the different forms of media [11, 20, 24].

4.4 Resolution

One difficulty is to correlate the resolution of an image with that subjectively perceived by the viewer in the physical world. In this instance, resolution not only refers to screen image dimensions (the pixel count), but also to the level of photorealism of the virtual environment that is created [5]. This also relates to the display mechanisms used: viewing an image on a mobile device such as a mobile telephone or an iPod is a very different experience to watching it on a cinema screen. In addition, seeing an image or animation on a screen may not have the same experience (depth of field, motion parallax, peripheral vision etc.) as a viewer watching the actual event [75, 76].

Careful thought needs to be given to the enabling technology; it is necessary to consider how the user will interact with any visualisations created. For example, the best mechanism for a particular case may be to deliver a spatially contextualised evidence visualisation to a user's personal device (a mobile telephone or iPod screen) as they traverse the actual scene. Alternatively, a complex forensic data set with many spatially interlinked evidential items may be best utilised as a shared viewing experience on a large screen in the court [5, 20, 24].

4.5 Audio

The integration of physical-world audio evidence with a forensic animation has been used in the United States for many years. One of the first recorded applications of such a forensic animation was the reconstruction of the Delta 191 airplane crash in 1985 [77, 78]. In the court, the animated evidence showing the movement of the airplane was played simultaneously with an audio recording from the cockpit voice recorder. Research suggests that adding audio to a computer-generated visual can have a significant effect on the level of engagement of the viewer, and hence may potentially affect their understanding and interpretation of the evidence viewed [79].

The integration of sound into the virtual world is often overlooked or added as an afterthought. Very few designers or developers are also qualified as or competent at being sound engineers. Effective audio soundtracks can add new dimensions to the viewer's media experience. The addition of an audio track can be a positive alteration to the visuals, providing an increased understanding of events or it can be distracting, adding unnecessary emotional context [5, 79].

4.6 Accuracy

If the images and virtual reconstructions are created to a sufficient level of accuracy, then they may potentially be used to test hypotheses in a courtroom. Examples include verifying the location of a witness (especially where lines of sight around obstructions or hazards that are present in the environment may call into question the physical location of the witness) or perhaps to evaluate potential alternative bullet trajectories through the environment [80].

The use of generic computer models allows the recreation of dynamic events. Such reconstructions are, by their very nature, often dependent on the knowledge, expertise and opinion of experts [81]. Hence, in many of these cases, the advice of the expert is seen as crucial in creating a graphical representation that accurately matches the medical opinion. However, the potential inaccuracies involved mean that these reconstructions must be viewed cautiously and the uncertainty associated with the exact position of virtual objects must be explained to the viewer [82].

4.7 Simulation

It should never be forgotten that most computer-generated visuals are by their very definition a simulation of reality. In the context of the courtroom, it is necessary to understand the nature of the simulation and the veracity of the representation - that is how close it is to the original evidence from which it was derived.

However, questions that arise include whether visual applies scientific theory in the same way as the expert; whether the simulation works to the same level of accuracy; whether the simulation makes the same assumptions as the expert; and whether the visual representation provides a realistic portrayal of the simulation data [8, 69].

It is important that the developers of these images have an understanding of the processes and events being simulated (whether this is vehicle movement, bullet trajectories or human anatomy). The developers must be aware of the veracity and realism of the simulation - that is, the accuracy of the model. Also, it is important that if decisions are to be made based on the images created, then it is necessary that information is made available that explains how the simulation works and details of the underlying mathematical model [5, 24, 69].

4.8 Narrative

The ability to move through time and along a chronology of events in the virtual environment may be potentially disorientating to many viewers. Most are used to linear narratives (such as those in books or films), and may struggle to follow multiple narrative threads when faced with such a non-linear approach [83].

In an interactive simulation, the user may often take control of the narrative, altering the chronological presentation of information, and choosing which information they see at which time. This can easily become confusing to the viewer, particularly to those used to linear narratives in other media (for example, novels and films). Developers should produce a guide to the interactions in their environments and be aware (through user testing) of how the users are able to interact with the data and any possible unexpected interpretations that may result [24, 83].

5 Empirical Evidence of Effects

Australia currently has a number of projects underway in this thematic area. In Western Australia, rare permission has been given by the Attorney General for a researcher to interview jurors after criminal trials in which a range of expert evidence was presented. While the data showed statistically significant findings that jurors are clearly influenced in their treatment of some forensic evidence by the manner of presentation, reassuringly, no support was found for the operation of a detrimental effect of the technology. In fact the study found support for the proposition that most jurors assess forensic evidence in a balanced and thoughtful manner, whatever the mode of presentation [84, 85].

Fig. 4. Computer-generated graphical images from the JIVE terrorism trial.

The author was a member of a large international research project based in Australia, the Juries and Visual Evidence Project (JIVE), which also examined some of these issues. The project measured the effect of interactive displays on the trial process; specifically whether forensic animation and virtual reconstruction technology better informs juries or potentially increases prejudice against defendants [86].

In January 2008, the JIVE project team ran a number of mock trials in the Supreme Court in Sydney, Australia. A range of forensic animations and interactive reconstructions of evidence relating to a terrorist bombing were shown to a number of different groups of jurors (Fig. 4).

Each jury deliberation was filmed and recorded. A major theme emerging from the analysis of the project data is that the main experimental effects (interactive visual evidence and judicial instructions) have relatively modest influence overall. However, they do show stronger effects in some groups of people, particularly those who are most prone to convict. The JIVE data has so far shown that fear of terrorism may be a better predictor of a verdict than either the method of presentation, experimental interventions, deliberation or any demographic characteristics. The research team intends to publish a book on the data from this project which will focus on issues of juries and trials in terrorism cases [22, 24, 86].

6 Conclusions

In many courtrooms, the crux of any case is the presentation of information to the finder of fact, whether in the form of an opening statement, evidence or closing argument. The need for a clear presentation is summed up by Burns, who states [87]:

> *'The presentation typically takes the form of a report, and the scientist must be prepared to explain this report in such a way that a typically science-phobic judge and jury are able to comprehend it. Presentation is everything.'*

The unavoidable future for courtrooms across the world is the introduction of technology; this technology could be merely electronic filing and teleconferencing, but is likely to encompass many forms of computer-generated imagery. As computer-graphics based technologies continue to evolve, this will inevitably lead to improvements in the realism of evidential graphics and virtual simulations. This could, in turn, result in jurors and triers of fact experiencing a greater depth of immersion when viewing and experiencing the incident within a virtual world. This could also potentially lead to a corresponding increase in their acceptance or belief in the hypotheses being presented; and conversely also result in a rise in any associated possible prejudice caused by the visual media.

In conclusion, designers should endeavour to ensure that any images produced accurately reflects the scientific data available and augment the testimony of the witnesses. However, to be effective, the evidence must not only tell 'the story' but also be understood easily. To that end, designers must strive continuously to develop new and creative ways to present complex information. Graphics based technologies have the potential to have an important effect on many future cases.

References

1. Powers, D.: Vision in HCI: embodiment, multimodality and information capacity. In: Proceedings of the Workshop on the Use of Vision in HCI (VisHCI 2006), Canberra, Australia, vol. 56 (2006)
2. Schofield, D., Mason, S.: Using graphical technology to present evidence. In: Mason, S. (ed.) Electronic Evidence, 3rd edn, pp. 217–253. LexisNexis Butterworths, London (2012)

3. Burton, A., Schofield, D., Goodwin, L.M.: Gates of global perception: forensic graphics for evidence presentation. In: Proceedings of ACM Symposium on Virtual Reality Software and Technology, Singapore, pp. 103–111 (2005)
4. Joseph, G.P.: Modern Visual Evidence. Law Journal Seminars Press, New York (2009)
5. Feigenson, N., Spiesel, C.O.: Law on Display: The Digital Transformation of Legal Persuasion and Judgment. NYU Press, New York (2009)
6. Mervis, J.: Court views engineers as scientists. Science **284**(5411), 21 (1999)
7. Tufte, E.R.: The visual display of quantitative information. Am. J. Phys. **53**(11), 1117–1118 (1983)
8. Grimes, W.D.: Classifying the Elements in a Scientific Animation, Accident Reconstruction: Technology and Animation, USA Society of Automotive Engineers, SAE Paper no. 940919, pp. 39–404 (1994)
9. Burgoon, J.K.: Interactivity in human–computer interaction: a study of credibility, understanding and influence. Comput. Hum. Behav. **16**(6), 553–574 (2000)
10. Shapiro, M.A., McDonald, D.G.: I'm not a doctor, but I play one in virtual reality: implications of judgments about reality. J. Commun. **42**(4), 94–114 (1992)
11. Gilbert, D.T.: How mental systems behave. Am. Psychol. **46**(2), 107–119 (1991)
12. Hopkins, R.: Picture, Image and Experience. Cambridge University Press, Cambridge (1998)
13. Schofield, D.: Animating and interacting with graphical evidence: bringing courtrooms to life with virtual reconstructions. In: Proceedings of IEEE Conference on Computer Graphics, Imaging and Visualisation, Bangkok, Thailand, pp. 321–328 (2007)
14. Spiesel, C.O., Sherwin, R.K., Feigenson, N.: Law in the age of images: the challenges of visual literacy. In: Wagner, A., Summerfield, T., Vanegas, F.S.B. (eds.) Contemporary Issues of the Semiotics of Law, Oñati International Series in Law and Society, vol. 13 (2005)
15. Kuehn, P.F.: Maximising your persuasiveness: effective computer generated exhibits. Journal of the DuPage Country Bar Association. http://www.dcba.org/mpage/vol121099art4. Accessed 21 Jan 2018
16. Schroder, K.J.: Computer Animation: The Litigator's Legal Ally, Computers and Law. University of Buffalo, Buffalo (1997)
17. Devine, D.J., Clayton, L.D., Dunford, B.B., Seying, R., Pryce, J.: Jury decision making, 45 years of empirical research on deliberating groups. Psychol. Pub. Policy Law **7**(3), 622–727 (2001)
18. Durkin, K.P., Dunn, C.H.: Building your case for the jury, Litig. J. **36**(3) (2010)
19. Rubin, C.B.: A paperless trial. Litig. Magaz. **19**(3), 66–88 (1992)
20. Loftus, E.R., Loftus, G.R.: On the permanence of stored information in the human brain. Am. Psychol. **35**(5), 409–420 (1980)
21. Leader, L., Schofield, D.: Madness in the method? Potential pitfalls in handling expert evidence. J. Pers. Inj. Law **6**(1), 68–86 (2006)
22. Schofield, D.: The future of evidence: new applications of digital technologies, forensic science: classroom to courtroom In: Proceedings of 18th International Symposium of the Forensic Sciences, Fremantle, Western Australia (2006)
23. Richter, E.M., Humke, A.M.: Demonstrative evidence: evidence and technology in the courtroom. In: Weiner, R.L., Bornstein, B.H. (eds.) Handbook of Trial Consulting, pp. 187–201. Springer, Heidelberg (2011). https://doi.org/10.1007/978-1-4419-7569-0_9
24. Schofield, D.: Playing with evidence: using video games in the courtroom. J. Entertainment Comput. **2**(1), 47–58 (2011). (Special Issue: Video Games as Research Instruments)
25. Jones, I.S., Muir, D.W., Groo, S.W.: Computer animation – admissibility in the courtroom, accident reconstruction: technology and animation. Soc.Autom. Eng. **1**, 143–151 (1991)

26. Lederer, F.I., Solomon, S.H.: Courtroom technology – an introduction to the onrushing future. In: Proceedings of Fifth National Court Technology Conference (CTC5). National Centre for State Courts, Detroit (1997)
27. Lederer, F.I.: Courtroom technology: for trial lawyers, the future is now. Crim. Justice **19**(1), 14–21 (2004)
28. Seltzer, R.F.: Evidence and Exhibits at Trial, 387 PLI/Lit 371 (1990)
29. Cobo, M.E.: A Strategic Approach to Demonstrative Exhibits and Effective Jury Presentations, 3 PLI/Lit 359 (1990)
30. Krieger, R.: Sophisticated computer graphics come of age-and evidence will never be the same. J. Am. Bar Assoc. **78**, 93–95 (1992)
31. Thomas, R.D.: Computer Re-enactment. http://www.pimall.com/nais/n.reenact.html. Accessed 21 Jan 2018
32. Sherwin, R.K.: When Law Goes Pop: The Vanishing Line between Law and Popular Culture, 2nd edn. University Of Chicago Press, Chicago (2002)
33. Kassin, S.M., Dunn, M.A.: Computer-animated displays and the jury: facilitative and prejudicial effects. Law Hum. Behav. **21**(3), 269–281 (1997)
34. Selbak, J.: Digital litigation: the prejudicial effects of computer-generated animation in the courtroom. Berkley Technol. Law J. **19**(9:2), 337 (1994)
35. Dunn, M.A., Feigenson, N., Salovey, P.: The jury persuaded (and not): computer animation in the courtroom. Law Policy **28**(2), 228–248 (2006)
36. Tait, D.: Rethinking the role of the image in justice: visual evidence and science in the trial process. Law Probab. Risk **6**(1–4), 311–318 (2007)
37. Tait, D., Goodman-Delahunty, J., Schofield, D., Jones, D.: Evidence on the holodeck: jury responses to computer simulations. In: Proceedings of the 4th Law and Technology Conference. Australian Institute of Judicial Administration, Sydney (2008)
38. Schofield, D.: Animating evidence: computer game technology in the courtroom. J. Inf. Law Technol. (JILT) **1**, 1–20 (2009)
39. Loftus, G.R., Loftus, E.F.: The influence of one memory retrieval on a subsequent memory retrieval. Mem. Cognit. **2**(3), 467–471 (1974)
40. Young, W., Davis, M., McNeill, I.M., Malhotra, B., Russell, S., Unsworth, K., Clegg, C.W.: Changing behaviour: successful environmental programmes in the workplace. Bus. Strategy Environ. **24**(8), 689–703 (2015)
41. Davis, R., Campbell, R., Hildon, Z., Hobbs, L., Michie, S.: Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol. Rev. **9**(3), 323–344 (2015)
42. Harding, E.J., Paul, E.S., Mendl, M.: Animal behaviour: cognitive bias and affective state. Nature **427**(6972), 312 (2004)
43. Haselton, M.G., Nettle, D., Murray, D.R.: The Evolution of Cognitive Bias. The Handbook of Evolutionary Psychology. Wiley, New York (2005)
44. Loftus, E.F.: Eyewitness Testimony. Harvard University Press, Cambridge (1996)
45. Loftus, E.F.: Our changeable memories: legal and practical implications. Nat. Rev. Neurosci. **4**, 231–234 (2003)
46. Biederman, I., Ju, G.: Surface vs. edge-based determinants of visual recognition. Cognit. Psychol. **20**, 38–64 (1998)
47. Cooper, C.: Computer Animation on Trial, The San Diego Union Tribune (27th January 1993), E1 (1993)
48. Darley, J.M., Gross, P.H.: A hypothesis-confirming bias in labeling effects. J. Pers. Soc. Psychol. **44**(1), 20–33 (1983)

49. Nunez, N., McCoy, M.L., Clark, H.L., Shaw, L.A.: The testimony of elderly victim/witnesses and their impact on juror decisions: the importance of examining multiple stereotypes. Law Hum. Behav. **23**(4), 413–423 (1999)
50. Schuller, R.A., Terry, D., McKimmie, B.: The impact of expert testimony on jurors' decisions: gender of the expert and testimony complexity. J. Appl. Soc. Psychol. **35**(6), 1266–1280 (2005)
51. Wuensch, K.L., Campbell, M.W., Kesler, F.C., Moore, C.H.: Racial bias in decisions made by mock jurors evaluating a case of sexual harassment. J. Soc. Psychol. **142**(5), 587–600 (2002)
52. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Speech Acts Syntax and Semantics, vol. 3, pp. 41–58. Academic Press, London (1975)
53. Goodwin, L., Schofield, D.: Visualising uncertainty: combining evidence with statistics. In: Proceedings of Conference on Expert Evidence: Causation, Proof and Presentation, Florence, Italy (2002)
54. Klein, G.: Recognition-primed decisions. In: Rouse, W.B. (ed.) Advances in Man-Machine Systems Research, vol. 5, pp. 47–92. JAI Press, Greenwich (1989)
55. Lurie, N.H., Mason, C.: Visual representation: implications for decision making. J. Market. **71**, 160–177 (2007)
56. Chapman, P.: Remembering what we've seen: predicting recollective experience from eye movements when viewing everyday scenes. In: Underwood, G.D.M. (ed.) Cognitive Processes in Eye Guidance, pp. 237–258. Oxford University Press, Oxford (2005)
57. Drummond, M.A.: Courtroom Persuasion: Eight Keys to Success. Loyola University Press, Institute for Continuing Legal Education, Chicago (1999)
58. Yale, D.: Computers on the witness stand: expert testimony that relies on data generated by computers in the age of Daubert. http://www.dcyale.com/law_papers/daubert.html. Accessed 21 Jan 2018
59. Brewer, M.B.: A dual process model of impression formation. In: Wyer, R.S., Srull, T.K. (eds.) Advances in Social Cognition, vol. 1, pp. 1–36. Psychology Press (1998)
60. Morkes, J., Kernal, H.K., Nass, C.: Task-oriented computer-mediated communication and human-computer interaction. In: Proceedings of CHI 98 Conference Summary on Human Factors in Computing Systems, pp. 215–216. ACM (1998)
61. O'Flaherty, D.: Computer-generated displays in the courtroom: for better or worse? Web J. Curr. Legal Issues **2**(4), 1–14 (1996)
62. Habermas, J.: The Structural Transformation of the Public Sphere. MIT Press, Boston (1996)
63. Baudrillard, J.: Simulacra and Simulation. University of Michigan, Michigan (1994)
64. Nichols, S., Haldane, C., Wilson, J.R.: Measurement of presence and side effects in virtual environments. Int. J. Hum. Comput. Stud. **52**(3), 471–491 (2000)
65. Hussin, N., Schofield, D., Shalaby, T.M.: Visualising information: evidence analysis for computer-generated animation (CGA). In: Proceedings of the 8th International Conference on Information Visualisation (IV), London, UK, pp. 903–908 (2004)
66. Rossner, M., Yamada, K.M.: What's in a picture? The temptation of image manipulation. J. Cell Biol. **166**(1), 11–15 (2004)
67. Porter, G.: Visual culture in forensic science. Aust. J. Forensic Sci. **39**(2), 81–91 (2007)
68. Worring, M., Cucchiara, R.: Multimedia in forensics. In: Proceedings of the 17th ACM International Conference on Multimedia, New York, USA, pp. 1153–1154 (2009)
69. Noond, J., Schofield, D., March, J., Evison, M.: Visualising the scene: computer graphics and evidence presentation. Sci. Justice **42**(2), 89–95 (2002)

70. Bryce, J., Rutter, J.: Spectacle of the deathmatch: character and narrative in first-person shooters. In: King, G., Krzywinska, T. (eds.) ScreenPlay: Cinema/Videogames/Interfaces, pp. 66–80. Wallflower Press, Boston (2002)
71. Jones, E.E., Nesbett, R.E.: The actor and the observer: divergent perceptions of the causes of behavior. In: Jones, E.E. (ed.) Attribution: Perceiving the Causes of Behavior, pp. 79–94. General Learning Press, Morristown (1971)
72. MacDorman, K.F.: Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: an exploration of the uncanny valley. In: Proceedings of the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science, pp. 26–29 (2006)
73. Machado, L.S., Morales, R.M., Souza, D.L., Souza, L., Cunha, I.L.: A framework for development of virtual reality-based training simulators. In: Westewood, J.D. (ed.) Medicine Meets Virtual Reality. IOS Press, New York (2009)
74. Tromp, J., Schofield, D.: Practical experiences of building virtual reality systems. In: Proceedings of Designing and Evaluating Virtual Reality Systems Symposium. University of Nottingham, UK (2004)
75. Fielder, B.S.: Are your eyes deceiving you? The evidential crisis regarding the admissibility of computer-generated evidence. New York Law Sch. Law Rev. **48**(1–2), 295–321 (2003)
76. Galves, F.: Where the not so wild things are: computers in the courtroom, the federal rules of evidence, and the need for institutional reform and more judicial acceptance. Harvard J. Law Technol. **13**(2), 161–302 (2002)
77. Gold, S.: Forensic Animation – Its Origins, Creation, Limitations and Future, Expert Law. http://www.expertlaw.com/library/animation/forensic_animation.html. Accessed 21 Jan 2018
78. Marcotte, P.: Animated evidence: delta 191 crash re-created through computer simulations at trial. J. Am. Bar Assoc. **75**, 52–57 (1989)
79. Porschmann, C., Pellegrini, R.S.: 3-D audio in mobile communication devices: effects of self-created and external sounds on presence in auditory virtual environments. J. Virtual Real. Broadcast. **7**, 1–9 (2010)
80. Buck, U., Naether, S., Räss, B., Jackowski, C., Thali, M.J.: Accident or homicide–virtual crime scene reconstruction using 3D methods. Forensic Sci. Int. **225**(1), 75–84 (2013)
81. March, J., Schofield, D., Evison, M., Woodford, N.: Three-dimensional computer visualisation of forensic pathology data. Am. J. Forensic Med. Pathol. **25**(1), 60–70 (2004)
82. Maksymowicz, K., Tunikowski, W., Kościuk, J.: Crime event 3D reconstruction based on incomplete or fragmentary evidence material–case report. Forensic Sci. Int. **242**, e6–e11 (2014)
83. Craven, M., Taylor, I., Drozd, A., Purbrick, J., Greenhalgh, C., Benford, S.: Exploiting interactivity, influence, space and time to explore non-linear drama in virtual worlds. In: Proceedings of the SIGCHI, Human Factors in Computing Systems, pp. 30–37 (2001)
84. Fordham, J.G.: Muddying the waters with red herrings: a progress report on Western Australian jury research. In: Brooks-Gordon, B., Freeman, M. (eds.) Law and Psychology Current Legal Issues, vol. 9, pp. 338–360. Oxford University Press, Oxford (2006)
85. Holmgren, J.A., Fordham, J.G.: The CSI effect and the Canadian and Australian jury. J. Forensic Sci. **56**(S1), S63–S71 (2011)
86. Schofield, D.: Displaying the bomb on the train: the challenge of preparing visual evidence. In: Tait, D., Goodman-Delahunty, J. (eds.) Juries, Science and Popular Culture in the Age of Terror, pp. 123–143. Palgrave Macmillan, Basingstoke (2016)
87. Burns, D.C.: When used in the criminal legal process forensic science shows a bias in favour of the prosecution. Sci. Justice **41**(4), 271–277 (2001)

Impact of the Size of Chinese Characters on the Visual Search Performance Under Vibration

Lifen Tan[1], Chunhui Wang[1(✉)], and Hongting Li[2]

[1] National Key Laboratory of Human Factors Engineering, China Astronaut Research and Training Center, Beijing 100094, China
chunhui_891@163.com

[2] Department of Psychology, Zhejiang Sci-Tech University, Hangzhou 310018, China

Abstract. Display of characters is one of the important and widely applied display mode in display design. This study mainly aims to examine the impact of different sizes of characters on visual search under simulated vibration and real running vehicle vibration through the performance of visual search and ratings of subjective difficulties. To complete the visual search task of the target words, the study aims to analyze the accuracy of the visual search, the reaction time and the grading results of the search difficulties. The experimental results show that with the increase of the size of characters under vibration, the visual search performance will improve, and the rating of participants' subjective difficulties will reduce. Especially when the size is 20 pt, the search performance and subjective evaluation will be dropped significantly. The result of this study provides an advice for the design of the display of the characters that the character size should be greater than 20 pt under vibration.

Keywords: Chinese characters · Vibration environment · Visual search

1 Introduction

The effect of character display on visual search performance has been widely concerned by many researchers at home and abroad. The study found that the display ergonomics such as character font, size, character density, line spacing and alignment affect visual search performance. Ling (2006) compares the visual search performance of the Times New Roman font and the Arial font design in a web page. For Chinese fonts, Gongkun (2009) used visual search procedure and experimental introspection methods to examine the influence of Chinese character fonts on the visual search response. The experimental results show that when the font size is the same, the visual search of Song-style is very significantly less than that of the Regular Script. Shen (1990) also conducted systematic research on the factors that affect the ergonomics of Chinese characters display early. Text information density is one of the important factors affecting the visual search performance of the visual display interface. For example, Ling et al. (Ling and Schaik 2007) studied the effects of different line spacings and text alignment methods on visual

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 345–354, 2018.
https://doi.org/10.1007/978-3-319-91238-7_28

search performance for web-based web interfaces. The results show that wider line spacing and left-aligned text can significantly improve search performance. In terms of the density of Chinese textual information, Jing et al. (2010) studied experimentally the relationship between web page character information density and recognition response time. Their experimental results show that the information display density has a significant effect on search efficiency, and there is an approximate inverted U-shaped curve relationship between the two. Environment may also affect visual search. Yuan et al. (2011) and Underwood et al. (2002) studied the impact of vehicle speed and text height on visual search patterns in a simulated traffic environment.

With the wide application of touch screen display, more and more physical operation interfaces are replaced by touch screen interfaces, which also makes the working environment of the touch screen interface more complicated. For example, the touch screen operation in a vibrating environment has a high requirement for the operator's visual interpretation performance to avoid operational errors. Wu et al. (2015) analyzed the vibration environment of armored vehicle cabins, explained the impact of vibration on the ergonomics of occupant operations, and gave the effect of vibration on visual recognition.

However, so far, no study has examined the effect of text size on visual search performance in a vibration environment. Therefore, this study will mainly examine the visual interpretation performance of different character sizes in a vibration environment to analyze the effect of character size on interpretation performance.

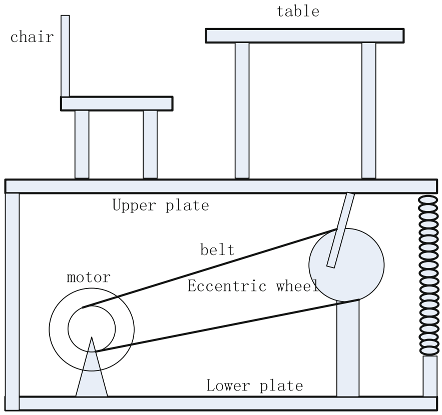
2 Method

2.1 Participants

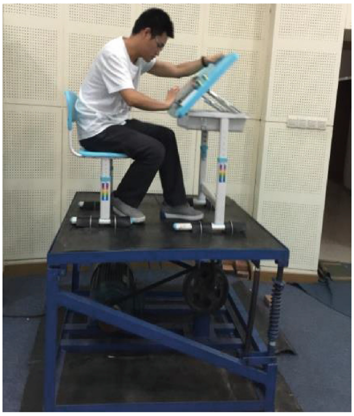
A total of 20 college students participated in the experiment, and all of them are males with age between 22 to 26 years old with normal visual ability and all with right-handedness.

2.2 Equipment

Simulated Vibration. The factors that cause vehicle vibration during vehicle driving include road surface excitation, traction load, vehicle motion device, power device, etc. The road surface irregularity is the main cause of vehicle vibration. The vibration of the vehicle is mainly three-axis vibration. The positive X1 axis is defined from the left to the right, the Y1 axis is positive from the back to the chest, and the Z1 axis is directed from the foot to the head. Vehicle vibration is mostly low-frequency vibration, and the intensity of vibration in the Z1 direction is generally larger than that of the X1 and Y1 directions, which is the main factor affecting the ergonomics of the crew. The research needs to establish an environment simulation model to simulate different vibration environments, mainly including the vibration amplitude and vibration frequency, and the vibration amplitude and frequency are required to be adjustable. A self-oscillating platform was developed, as shown in Fig. 1. An eccentric was driven by a motor to make the platform vibrate. The initial state of the table and chairs is that the upper plate is substantially parallel to the ground, and the last state is that the upper plate vibrates to the maximum amplitude position. Taking into account the common installation position and ease of use, the desktop on the shake table is adjusted to a tilt angle of about 70°.



a) design sketch of simulated vibration



b) physical simulated vibration

Fig. 1. Simulated vibration

Because the vibration simulation model drives the vibration of the upper plate through the eccentric wheel, the amplitude of the vibration needs to be changed by changing the eccentric diameter and eccentricity. Two types of eccentric wheels with different diameters are designed for this project, there are two amplitudes. The vibration frequency was 2.5 Hz and 5 Hz.

Vibration Measuring Equipment. According to the characteristics of the measuring equipment, vibration acceleration was used to characterize the vibration intensity. The vibration acceleration was measured using the MEMS triaxial accelerometer MPU6050 in the Mini IMU AHRS.

Character Search Device. Experiments are performed on Lenovo Yoga3 computer with 14 inch touch screen, 1920 × 1080 screen resolution, 16:9 aspect ratio in simulated vibration environment and running real vehicle environment.

Visual Search Test Software. The main purpose of the visual search test software development is to obtain the impact data of font size and spacing on visual search task performance in simulated vibration environment and running real vehicle environment. The visual search test software support text display to visual interpretation performance research.

2.3 The Experimental Design

Simulated Vibration. This experiment is a 2*2 *4 within-subjects design. The independent variables include frequency, amplitude, and text size. The amplitude includes two levels of 2.5 Hz and 5 Hz. The amplitude includes two sizes. The text size includes four levels of 20 pt, 30 pt, 40 pt, and 50 pt. The dependent variable includes performance indicators and subjective evaluation indicator, and the performance indicators include the accuracy and the response time of search. A total of 20 college students have participated in the experiments.

Running Real Vehicle Vibration. Beyond the simulation vibration station, experiments have also been carried out on one certain type of real vehicle running on the training field, and the vibration intensity of the real vehicle is weaker than the simulated vibration intensity. This experiment is a within-subjects design. The independent variable of the experiments is character size, including four levels: 20 pt, 30 pt, 40 pt and 50 pt. The real vehicle running at normal speed. A total of 8 college students have participated in the experiments.

2.4 Task

The task of the experiment is to ask the subject to visually search the screen for the number of target words in a line of text under the vibration state, and record the result by touch click.

2.5 Procedure

The subject sat on the table on simulated vibration or running real vehicle vibration. After the main experimenter introduces the instructions to the participants, the practice procedure is first started. When the participants confirm to understand the experiment tasks completely, the experiment will begin. In the formal experiment, the distance between the eyes of the participants and the screen is 40 cm. Firstly, the central fixation point " + " will be presented in the central of the screen. After 2 s, the target word which will be researched by the participants will appear, with a size of 20 pt and the font of Song Typeface, and it will disappear after 2 s. At this point, one line will scroll up from the bottom of the screen. The text has a total words of 19 characters, including the number of the target words from 3 characters to 5 characters randomly, and the positions of the target words in the text are also random. Participants need to find out the number of target words contained in the line as soon as possible in the case to ensure the correct situation, then enter the number of the target words by touching and clicking in the keystrokes area. The experimental program is shown in Fig. 1. Each participant needs to complete a total of 16 tests under the combination of three character sizes, two frequencies, and two amplitudes. Each group includes 12 judgment tasks. In order to balance the influence of the sequence of each group, the Latin square balance was used in the experimental sequence. At the same time, in order to control the effect of the text length of the line under different character sizes on the visual search difficulty, the program automatically adjusts the character spacing to keep the total length of the line text unchanged.

After the completion of each test, the participants need to make 5 score on the difficulty of the search of the size of this group characters under vibration.

The experimental procedure is complied with the C # language, and the program will automatically record the right and wrong of the participants and their reaction (Fig. 2).

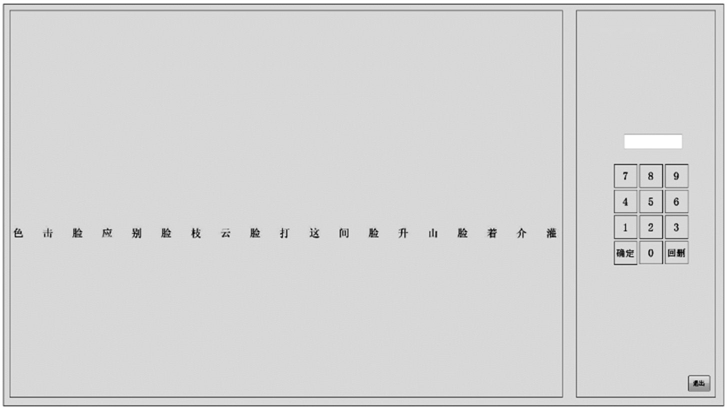


Fig. 2. Schematic diagram for the task of the experiment

3 Results

The experimental results are processed by SPSS19.0 soft package. Firstly, the data of extreme value outside the three standard deviations in all conditions are removed, which account for about 1.1% of the total data.

3.1 Visual Search Performance in Simulated Vibration Environment

Visual Search Accuracy under Different Frequency, Amplitude and Character Size Combinations. Figure 3 is the search accuracy under different character sizes, vibration frequencies and amplitudes. The repeated measure with analysis of variance show that the main effect of characters is not significant ($p > 0.05$); The main effect of the vibration frequency is significant ($p < 0.01$); The main effect of the vibration amplitude is significant ($p < 0.01$); And the interaction effect of vibration frequency and amplitude is significant ($p < 0.05$). Afterwards, the tests show that the difference between 30 pt and 50 pt is significant ($p < 0.05$), the difference between 40 pt and 50 pt is significant ($p < 0.05$), and the difference between 30 pt and 40 pt is not significant ($p > 0.05$). This shows that character size has influence on visual search accuracy under vibration state, But the main effect is not significant more relative to vibration.

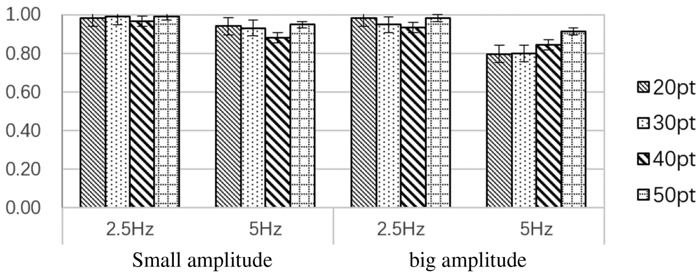


Fig. 3. Visual search accuracy of different character sizes under simulated vibration

Visual Search Reaction Time for Different Frequency, Amplitude and Character size Combinations. Figure 4 shows the results of search reaction times under different vibration and character sizes. The unit of the time is millisecond. The repeated measure with analysis of variance show that the main effect of the size of the characters is not significant ($p > 0.05$). The main effect of the vibration frequency is significant ($p < 0.01$). The main effect of the vibration amplitude is significant ($p < 0.01$). The interaction effect of vibration frequency and amplitude is significant ($p < 0.05$). Afterwards, the tests show that the difference between 20 pt and 50 pt has reached the significant level ($p < 0.05$), and the differences between other sizes are not significant ($p > 0.05$). From the Fig. 4, we can see that with the character size becoming larger, the reaction time is gradually decreasing, and the effect can be more obviously under higher frequency and larger amplitude conditions.

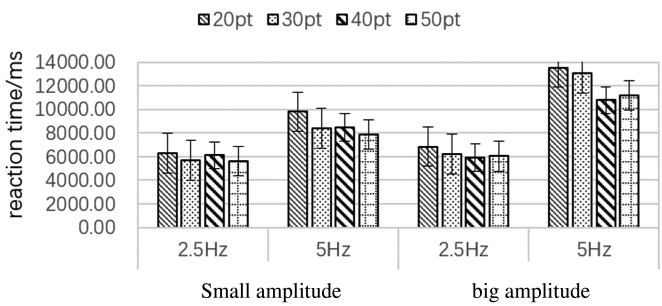


Fig. 4. Visual search reaction time of different character sizes under simulated vibration

The Subjective Ratings of the Visual Search for Different Frequency, Amplitude and Character Size Combinations. Figure 5 shows the results of ratings under different vibration and character sizes. The repeated measure with analysis of variance results of subjective ratings of the participants under different character sizes and vibrations show that the main effect of vibration frequency is not significant ($p > 0.05$). The main effect of vibration amplitude is not significant ($p > 0.05$), But the main effect of character size is significant ($p < 0.01$). There is no interaction between vibration frequency, amplitude and character size. The further analysis shows that the character size of 20 pt and the differences between 30 pt, 40 pt and 50 pt have reached the significant level ($p < 0.01$). It means that the smaller size of (20 pt) the characters is inappropriate.

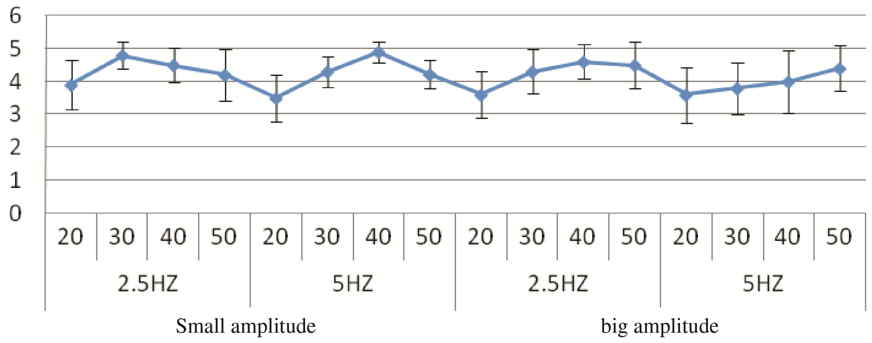


Fig. 5. Subject ratings of different character sizes under simulated vibration

3.2 Visual Search Performance in Running Real Vehicle Vibration Environment

Visual Search Accuracy under Different Frequency, Amplitude and Character Size Combinations. Figure 6 is the search accuracy under different character sizes under real vehicle running vibration. The analysis of variance show that the main effect of character size is significant ($p < 0.05$). Afterwards, the tests show that the character size of 40pt and the differences between 20 pt and 30 pt have reached the significant level ($p < 0.05$).

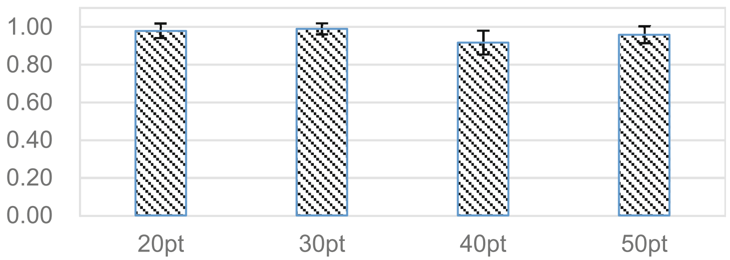


Fig. 6. Visual search accuracy of different character sizes under vehicle vibration

Visual Search Reaction Time for Different Frequency, Amplitude and Character Size Combinations. Figure 7 is the search reaction time under different character sizes under real vehicle running vibration. The analysis of variance show that the main effect of character size is not significant ($p > 0.05$). Afterwards, the tests show that there is no significant difference between various sizes.

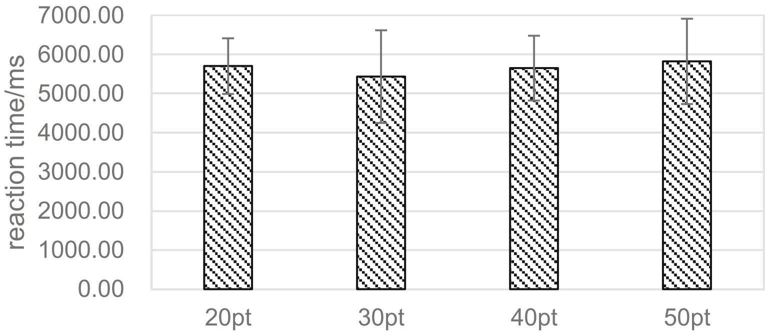


Fig. 7. Reaction time of visual search of different character sizes under vehicle vibration

The Subjective Ratings of the Visual Search for Different Frequency, Amplitude and Character Size Combinations. Figure 8 shows the results of ratings under different character sizes. The analysis of variance results of subjective ratings of the participants under different character sizes show that the main effect of character size is significant ($p < 0.01$). The further analysis shows that the character size of 20 pt and the differences between 30 pt, 40 pt and 50 pt have reached the significant level ($p < 0.01$).

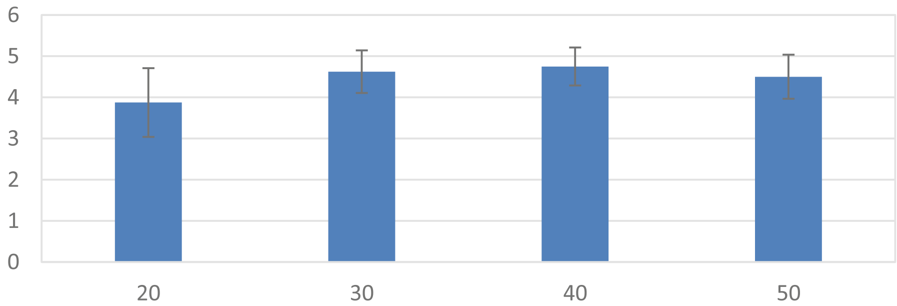


Fig. 8. Subject ratings of different character sizes under vehicle vibration

Under the vibration of running vehicle, the smaller size of (20 pt) the characters is inappropriate and even will improve subjective difficulties. But the measured performance show that the performance of character size 40 pt and 50 pt is worse than the performance of size 20 pt and 30 pt. This may because that the smaller size of character will inspire participants' motivation to become more concentrated, which has promote the performance.

4 Discussion

In this study, the influence of different character sizes on visual search under different frequency and amplitude vibration environments was investigated through the accuracy, the reaction time and subjective scores in visual search. The results show that under the vibrating environment, as the size of characters increases, the reaction time will decrease. In the case of 2.5 Hz, the decreasing trend of the reaction time at both small amplitude and large amplitude is relatively gentle; at 5 Hz, the decreasing trend of the reaction time is significant, especially between 20 pt and 50 pt. In the case of a large amplitude of 5 Hz, as the font size becomes larger, the increase in the accuracy of interpretation becomes more pronounced. The subjective evaluation of subject's difficulty in visual search also decreased significantly with the increase of font size, which was consistent with the performance results. It shows that in the vibration environment, choosing a larger font is more conducive to ensuring job performance. The stronger the vibration, the more obvious the effect of font size on performance.

The amplitude and frequency both have a significant effect on the accuracy and reaction time of Chinese characters' visual search. And the interaction between amplitude and frequency is significant.

Due to the limited experimental conditions, the range of variables selected in this study is still relatively small. This may be the reason that the difference in the character size performance between reaction time and accuracy is not obvious. In addition, previous studies have shown that different font types, arrangements, and other factors will affect the performance of character interpretation search (Ling 2006; Gongkun 2009). In this study, the font type was selected as the Song Dynasty. In the follow-up study, we can examine the influence of different fonts and other factors on the character display performance judgment under the vibration environment.

5 Conclusion

This study shows that with the increase of vibration frequency and amplitude, larger fonts can improve visual interpretation performance, and subjects are subjectively perceived to be easier to operate. In the case of a large amplitude of 5 Hz, the effect of character size on performance is even more pronounced. No matter which kind of vibration state, when the character size is 20 pt, the interpretation performance and subjective evaluation will be significantly reduced. The results of this study for character display design under vibration conditions are that the character size should be greater than 20 pt.

Acknowledgement. This research was supported by Experimental Technique Youth Fund Project (SYFD1400618) and Manned space advance research project (030602).

References

Ling, J., Schaik, P.: The influence of font type and line length on visual search and information retrieval in web pages. Int. J. Hum Comput Stud. **64**(5), 394–404 (2006)

Gongkun, Hao, C., Wang, D.: Effects of font characteristics and search methods on visual search response. Psychol. Sci. **32**(5), 1142–1145 (2009)

Shen, M., Zhu, Z., Jin, W.: The influence of the stroke width of Chinese characters on interpretation performance. Institute of Industrial Psychology, Hangzhou University, "Electro-optical display ergonomics of aircraft cockpit", pp. 101–114 (1990)

Ling, J., Schaik, P.V.: The influence of line spacing and text alignment on visual search of web pages. Displays **28**, 60–67 (2007)

Jing, L., Junhao, X., Dan, H., Yongjian, L.: Recognition efficiency and reliability of web character display density. Ind. Eng. Manag. **15**(5), 82–86 (2010)

Wu, G., Xu, X., et al.: State of armored vehicle cabin vibration environment and its effect on crew work efficiency. Chin. J. Ergon. **21**(1), 84–86 (2015)

Yuan, W., Fu, R., et al.: Effects of vehicle speed and traffic sign text height on drivers' visual search patterns. J. Traffic Transp. Eng. **11**(1), 119–126 (2011)

Underwood, G., Crundall, D., Chapman, P.: Selective searching while driving: the role of experience in hazard detection and general surveillance. Ergonomics **45**(1), 1–15 (2002)

Image Blurring Method for Enhancing Digital Content Viewing Experience

Hiroaki Yamaura(✉), Masayuki Tamura(✉), and Satoshi Nakamura(✉)

Meiji University, Nakano, Tokyo, Japan
hiro46staraki@yahoo.co.jp, ev30596@meiji.ac.jp,
satoshi@snakamura.org

Abstract. Many systems have been studied for enhancing a user's interest in digital content by using HMD, 3DTV, and so on. However, for these systems to be able to enhance a user's interest, the creator needs to elaborate the content. In this paper, we present a method that extends the experience of digital content by simply superimposing blurring effects that follow the gaze point of the user. To clarify how the experience of viewing still images and videos was enhanced by our method, we compared user impressions when viewing digital content with and without our method. We also examined physiological impressions such as visibility and discomfort. The experimental results showed that the participants' impressions of the video content were changed by superimposing the blurring effect on the peripherally viewed area. In particular, almost all psychological impression items (immersion, stereoscopic effect and so on) were scored higher when the blur was superimposed than when it was not.

Keywords: Viewing experience expansion · Impression evaluation · Gaze point Peripheral vision

1 Introduction

Many systems have been studied for enhancing a viewer's experience of and interest in digital content. Dive into the Movie [1] is a system that embeds viewers' faces into characters in a movie, so they can feel as if they were a character in the movie. This system succeeds in amplifying the viewer's interest and viewing experience. On the other hand, 3DTV and head mounted displays (HMDs) have the potential to enhance the degree of immersion offered by digital content. However, these systems are difficult to generate content for and have demanding hardware requirements. We thus aim to enhance the viewer's experience of and interest in existing digital content without him or her having to expend time and effort.

Here, a gaze point is regarded as accurately representing the viewer's interest. In addition, a person's visual field has two parts: central and peripheral vision. Central vision clearly perceives the gaze object, while peripheral vision perceives the periphery of the field of view vaguely [2]. Accordingly, we can expect that the viewing experience can be enhanced by presenting digital content in a way that takes into account the properties of central and peripheral vision. Okatani et al. [3] developed a gaze-reactive

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 355–370, 2018.
https://doi.org/10.1007/978-3-319-91238-7_29

display that changes a viewer's impression of a photograph by presenting a blur effect on it depending on the gaze point detected by an eye tracking device. However, their method requires preparing blurred images for each part of the target photograph in advance, which also requires time and effort. In addition, this method is difficult to apply to video clips because it requires a blurred image to be generated for each position in each frame. Moreover, the resolution of the gaze point is limited.

In this study, we focused on the characteristics of central vision, peripheral vision, and gaze and devised a method to enhance viewers' experience of digital content by dynamically superimposing a simple blur effect on it depending on the gaze point.

To realize this system, we expanded the versatility of the gaze-reactive display [3] by monitoring a viewer's gaze point and by superimposing blur effects surrounding the gaze point on the digital content in real time by using OpenGL Shading Language (GLSL). Here, the parts on which the user's eyes are focused remain clear, while the parts not focused on become more blurred the farther away they are from the gaze point (Fig. 1). In our method, users do not need to prepare blurred images in advance and can enjoy not only still images but also videos and video games with blurred effects. By applying blur effects, we aim to expand and amplify impressions such as presence or immersion.

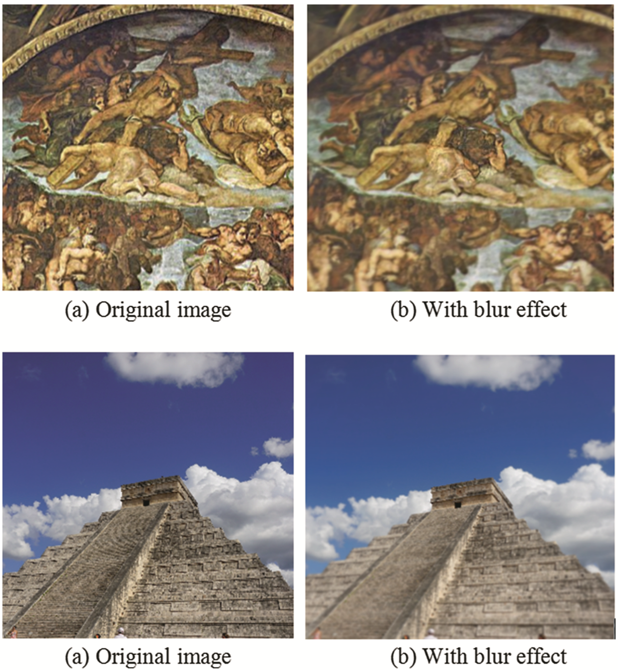


Fig. 1. Examples of adding blur effect by our method.

We implemented a prototype system that superimposes blurring using a Gaussian filter on arbitrary moving images and experimentally evaluated its usefulness. We also experimentally clarified for which types of video the proposed method is effective.

2 Related Work

Many researchers have attempted to amplify interest in content by superimposing effects on it. As mentioned above, the gaze-reactive display of Okatani et al. [3] dynamically changed the depth of the focus point of a photograph depending on the viewer's gaze point detected by an eye tracking system and its depth information. Changing the depth of focus enhances the user's experience. However, up to 256 blurred images had to be generated in advance for the user to experience content on their gaze-reactive display.

Other systems amplify interest, stereoscopic effect, and sense of reality by changing the blurring degree (depth of field) of the virtual world depending on the gaze point [4, 5]. However, these systems only focus on the virtual world and lack versatility because they need to use the data of the cameras used for rendering in the contents. In contrast, our method expands the viewing experience by superimposing simple blur effects on arbitrary video content. Furthermore, Okatani et al. [3] reported that viewers notice the artificiality when there is a slight delay between the movement of the gaze point and the movement of the blurred point of the content. To solve these problems, we use GLSL, which can process images with as short as possible delays.

Hirai et al. proposed the VRMixer system [6], which increases the fun provided by video contents by projecting real user images into them by using a depth camera. Kagawa et al. [7] developed a method that enables users to add emotion illustrations like hearts, called "intuitive emoji comment," to video content and showed its usefulness in an evaluation. However, these methods superimpose the effects directly on the video content, which impairs its visibility. Our method keeps visibility as high as possible by reproducing the blurring condition in the viewer's peripheral visual field on the display.

Hata et al. [8] revealed that the gaze point unconsciously moves from low-resolution areas to high-resolution areas in an image. They developed a gaze control system that changes the resolution of each part of content in an image. They also found that the time required for visual guidance depends on the intensity of blurring. Their method effectively uses the features of central and peripheral vision. This knowledge can also be used for our method, and blurring is thought to be able to enhance the viewers' ability to concentrate.

3 Prototype System

Our work is intended to easily enhance the experience of digital content such as the feeling of realism and immersion by adding a blur effect to the content surrounding the central visual field.

Because peripheral vision only perceives things faintly and processes visual information unconsciously, we accordingly focus on the difference between central vision and peripheral vision and develop a new method for viewing digital content. When a user views digital content by using our method, the central visual field, which is detected by the eye-tracking system, is clear and the peripheral visual field becomes more blurred the farther out from the center it goes. This effect emphasizes the centrally viewed part of the content. Figure 2 shows example images of the system in action. In the figure, the blur effect is superimposed on the light blue part and not on the clear elliptical part.

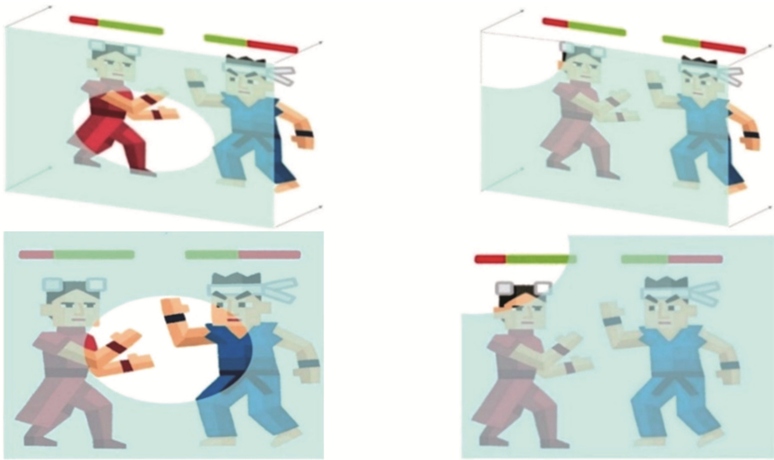


Fig. 2. Images of our method that blurs the peripherally viewed part of the content.

We use a Gaussian filter for this process. An image is processed by increasing the weight of the filter as it moves away from the central vision of the gaze point acquired by the eye-tracking device (see Fig. 3). Figure 3 illustrates the filter structure: the blurring is stronger farther from the gaze point. In particular, the weight is increased in four stages to change the blur level depending on the distance from the gaze point.

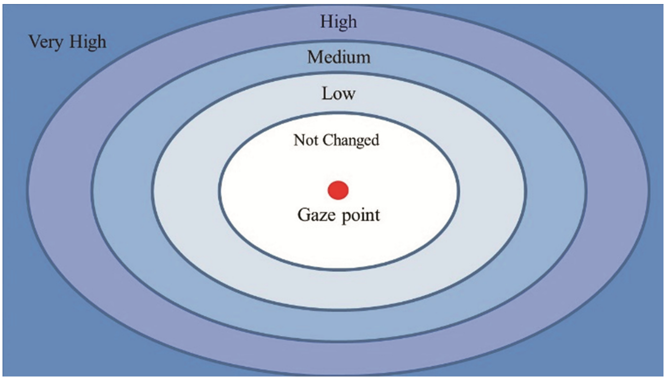


Fig. 3. Blur level is changed depending on the distance from the gaze point.

The shapes of the regions with different blur levels must be elliptical to match the human visual field, so the prototype system has to give a weight to each point of the image by using a filter [3]. Here, we let σ be the weight of the filter, (x, y) be arbitrary coordinates of the video contents, and c be the RGB value before the Gaussian filter is superimposed on the target pixel.

When the length of one side is k [pixels], the range to which the filter is applied is represented by a square of k × k pixels centered on the target pixels, and the RGB values C after superimposing the Gaussian filter on the target pixels are derived from (1) and (2), respectively.

$$f(x, y) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2}) \tag{1}$$

$$C = \sum_{y=1}^{k} \{ \sum_{x=1}^{k} c \bullet f(x, y) \} \tag{2}$$

The blur can be increased by increasing the value of **σ** in Eq. (1). Here, our method manages the color information by using the RGB value. If the value exceeds a certain threshold, the brightness of the pixel is decreased and the image turns black. Therefore, we set **σ** such that it does not exceed a predefined upper limit. In addition, k [pixels], which is the length of one side of the filter, increases the farther it is from the gaze point.

By superimposing such effects, we reproduce a "human" field of view that is close to reality on the display. The user is expected to be able to obtain a stereoscopic effect and feeling of immersion.

3.1 Implementation Method

We implemented our system by using Processing and GLSL. Here, our system consists of an eye-tracking module and content presentation module. The eye-tracking module monitors the gaze point of a user by using Tobii EyeX and transmits the detected gaze point data to the content presentation module. The content presentation module superimposes the blur effect on the currently presented content using GLSL on the basis of the gaze point data. GLSL is a programming language specialized for image processing and operates at high speed because it processes images using a graphics processing unit (GPU). Furthermore, our method succeeds in minimizing the delay in updating the video information and reproducing the human visual field with high accuracy.

Incidentally, there is a phenomenon called saccade, in which the gaze point moves frequently due to minute eye movements that the eyeball always perform involuntarily. If the system reflects the saccade when drawing effects, the effects will become an obstacle to viewing the content. Therefore, when the variation of the gaze position is less than a certain threshold, the gaze information is not updated. Since the gaze point of humans is elliptical [3], the range of effect is changed in an elliptical shape in accordance with the distance from the gaze point (Fig. 4).

(a) No effect (b) With effect (c) Effect and Boundary

(a) No effect (b) With effect (c) Effect and Boundary

(a) No effect (b) With effect (c) Effect and Boundary

(a) No effect (b) With effect (c) Effect and Boundary

Fig. 4. Examples of effect superimposition.

Figure 4 shows (a) original images, (b) original images with blur effects, and (c) original images with blur effects and boundary information where the weight changes to clarify the boundary of the blur. In (c), the central ellipse is not superimposed with the effect, but the farther out the ellipse is, the stronger the blurring becomes.

4 Preliminary Experiments

To clarify how the viewing experience of still images and videos is enhanced by superimposing blurring, we investigated the difference in subjective impressions between viewing digital content with and without our method. Here, we also investigated physiological impressions of effects such as visibility and discomfort.

4.1 Preliminary Experiment on Image Content

First, we clarified whether the viewing experience of contents is enhanced by superimposing blur using a Gaussian filter on the image content in the peripheral visual field. Images for the experiment were selected subjectively by us. The images were classified into video-game content, landscapes, paintings, illustrations, and geometric patterns. We selected 30 images for evaluation and asked ten university students aged 18–21 years old to participate in the experiment.

The experiment procedure was as follows. First, we asked each participant to sit in front of the display and adjust the Tobii EyeX. The distance at which Tobii EyeX succeeded in adjusting the gaze point was set as the distance between the display and the user. The position of the display was adjusted so that the viewing angle of the center field of view was 10° above and below and 15° to the left and right.

Next, we informed the participants that they were going to view image contents from now on and told them to look at the whole image. Here, we taught them how to turn the blur on and off. They viewed the 30 original images on a 27-inch display (Fig. 5) and turned the blur on and off arbitrarily while viewing images. After the viewing, we verbally asked them which images gave them different impressions when the blur was used.

Fig. 5. Scene of experiment.

As a result of this experiment, we found that our method was effective on the video-game content and landscape images but not on the paintings or geometric patterns. The reason for this is that the Gaussian filter does not work well in a single-color area.

4.2 Preliminary Experiment on Video Content

Since the prototype system worked effectively on some types of image content, we thought that the impressions of the user might change even for video content. Therefore, we prepared 12 videos and conducted an experiment by preparing an impression evaluation questionnaire to find out whether the impression is enhanced when the blur is superimposed on the peripheral visual field.

In this experiment, we prepared two types of video clip: video games and actual landscape videos. The video content lasted from 30 s to 2 min 30 s. As an experimental procedure, Tobii EyeX was adjusted as in Subsect. 4.1.

Next, we asked four participants (male and female university students aged 18 to 21 and divided into two groups of two) to view 12 videos on a 27-inch display. Then we asked them to answer the questionnaire after viewing each video. After they answered the questionnaire, we asked them to press the enter key to go to the next video. Here, one group viewed the videos with the blur, and the other group viewed videos without it. In addition, both experimental groups viewed the same videos in the same order.

The questionnaire consisted of eight items about emotions (pleasure, excitement, relief, pleasure, disgust, excitement, surprise, frustration, fear, and interest) selected from Plutchik's Wheel of Emotions, eight items about psychological impressions that we expected the blurring to give (stereoscopic effect, immersion, presence, tension, exhilarating feeling, feeling of freedom, sense of stagnation, and feeling of stir) and eight items about physiological impressions (discomfort, visibility, concentration, flickering, botheration, blur, motion sickness, and feeling strange). We asked the participants to score all items on a 7-level Likert scale.

The results showed that the score for each type of item increased for both types of video content. However, our system did not work well in videos where the similarity of each frame is very high, which coincides with the results of the preliminary experiment performed on image content.

Also, the emotions and impressions of participants changed depending on the length of the video contents, which suggests that videos of the same length should be used to test the system.

5 Experiment

We redesigned the experimental test on the basis of the results described in Section 4 in order to clarify the effectiveness and usefulness of our method for videos and clarify how it enhances the viewing experience.

5.1 Experimental Content

We prepared videos of retro video games (Retro-Game), new video games (New-Game), and actual landscapes (Landscape). We prepared four videos for each category. We omitted videos with few image changes on the basis of the results in Subsect. 4.2. Each video lasted 1 min 30 s. The viewing environment of the video content was the same as that described in Subsect. 4.2.

The participants were eight male and female university students aged 18–22 years old and divided into two groups of four.

The questionnaire included seven items about emotions (relief, pleasure, disgust, excitement, surprise, frustration, and interest), four items about psychological impressions (stereoscopic effect, immersion, presence, and tension), and three items about physiological impressions (comfort, visibility, and concentration).

The items were scored on a five-level Likert scale: maximum value of 2 and minimum value of –2. Taking comfort as an example, students answered whether they felt [2] comfortable, [1] somewhat comfortable, [0] neither comfortable not uncomfortable, [–1] somewhat uncomfortable, or [–2] uncomfortable. Emotion items were selected from Plutchik's Wheel of Emotions, and psychological and physiological impression items were selected on the basis of the experimental results in Subsect. 4.2. The experiment procedure was the same as that of the experiment in Subsect. 4.2.

5.2 Experimental Results

Figures 6, 7 and 8 show the average values of emotions, psychological impressions, and psychological impressions evaluated by participants for each of the three video categories.

Figure 6 shows disgust, surprise, and frustration were scored negatively in all categories. Also, disgust and frustration were more positive for videos with the blur than without it. Evaluation values of pleasure were positive in all categories and increased with the blur in Retro-Game and New-Game.

Figure 7 shows videos without blur were scored negatively in each category for every item. On the other hand, videos with blur had more positive values than videos without it for almost all items. Immersion was scored positively in all categories.

Figure 8 shows that videos without blur were scored as having higher comfort and visibility than videos without it. These results indicate that our current method sometimes gives viewers negative physiological impressions.

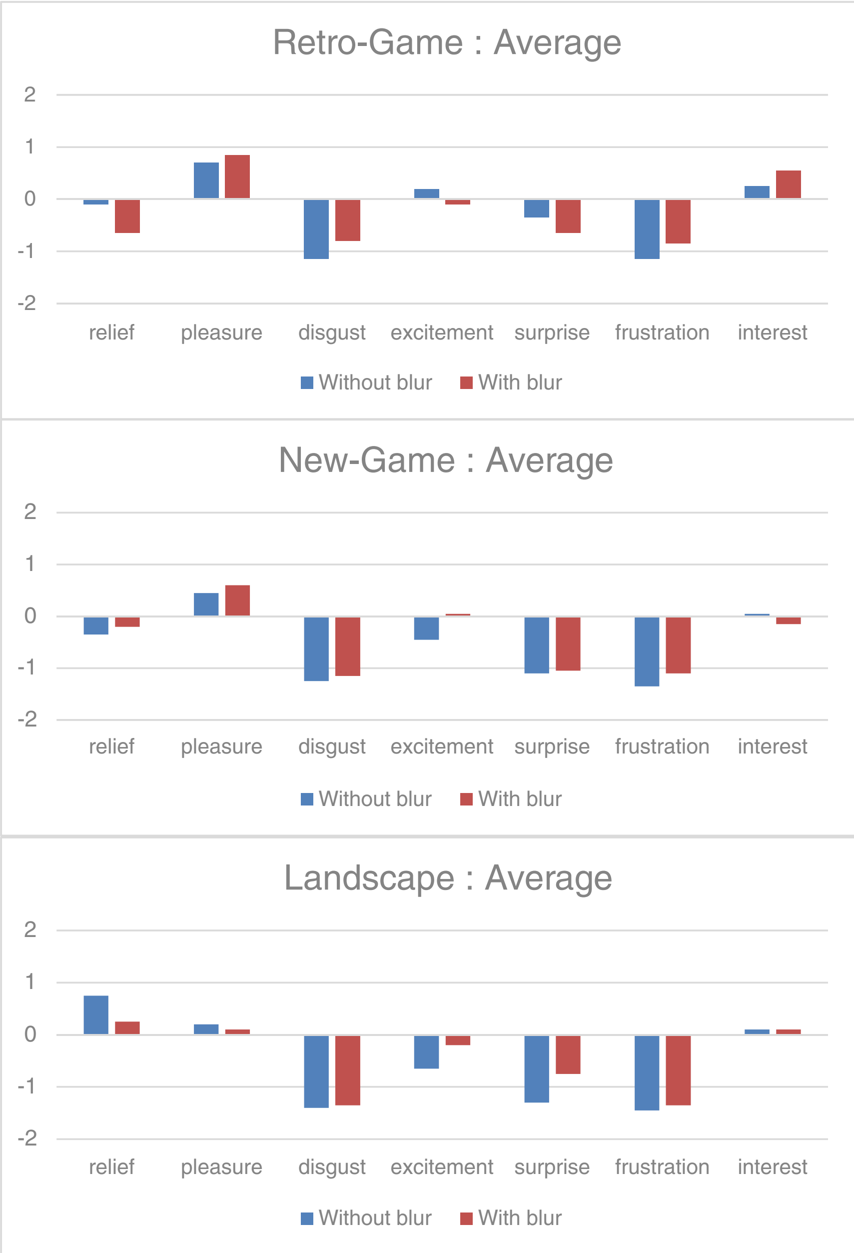


Fig. 6. Results for emotion items.

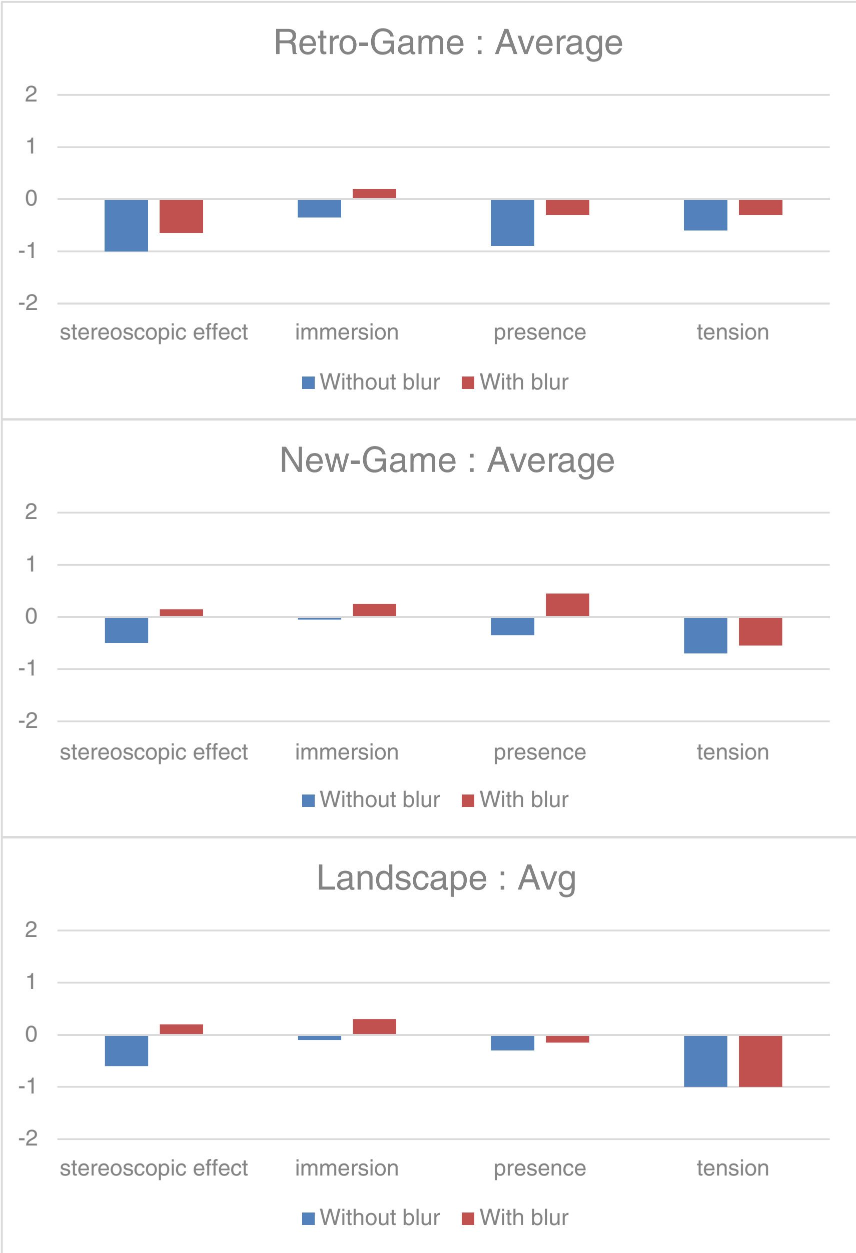


Fig. 7. Results for psychological impression items.

Fig. 8. Results for physiological impression items.

5.3 Considerations

The experimental results showed that the participants' impressions of the video content were changed by superimposing the blurring effect on the peripherally viewed area. In particular, almost all psychological impression items were scored higher when blur was superimposed than when it was not. These results clarify that our system is useful in all three categories.

In Fig. 8, landscape videos with blur scored high for concentration. This result coincides with those of Hata et al. [8]. In other words, our method achieves visual guidance because the centrally viewed area remains high resolution and the peripherally viewed area has low resolution. As a result, the gaze point is focused in this central visual field, which may increase the degree of concentration. Therefore, our method can be said to increase the degree of concentration for viewers of landscape videos.

On the other hand, since videos with blur scored lower for comfort and visibility in all the categories, our method seems to decrease the visibility of content. In this experiment, we set the range of the weight of the Gaussian filter to a range where the brightness of the image does not decrease. However, in the future, it will be necessary to verify the range of weights that do not impair visibility while enhancing the viewing experience. Also, none of the participants said that they noticed a delay, so our method seems to meet the requirement for no delays.

As well as videos of video games and landscapes, we aim to further narrow down the videos for which our method is useful by conducting experiments on animations, special effects videos, etc.

From the questionnaire on the physiological impression, it turned out that when participants view the video by superimposing blur on the landscape, the degree of concentration gets higher. Therefore, we will investigate the relationship between the degree of concentration and the gaze point by measuring the gaze point log when videos are viewed using this prototype system as an additional experiment.

5.4 Additional Experiments on Concentration and Gaze Point

To clarify whether or not the gaze point moves differently depending on the presence or absence of the blur, we prepared three new videos, one for each category (Retro-Game, New-Game, and Landscape). In addition, we implemented a logging system of viewer's the gaze point for analyzing the viewer's behaviors.

Participants were eight university students aged 18–21 and divided into two groups of four. Although the viewing environment and procedure are almost the same as those in Subsects. 4.2 and 5.1, there was no need to answer a questionnaire, so participants took a break for 10 s after viewing a video and then moved on to the next video. The measured gaze point logs are shown as heat maps for each video with and without the blur (Fig. 9).

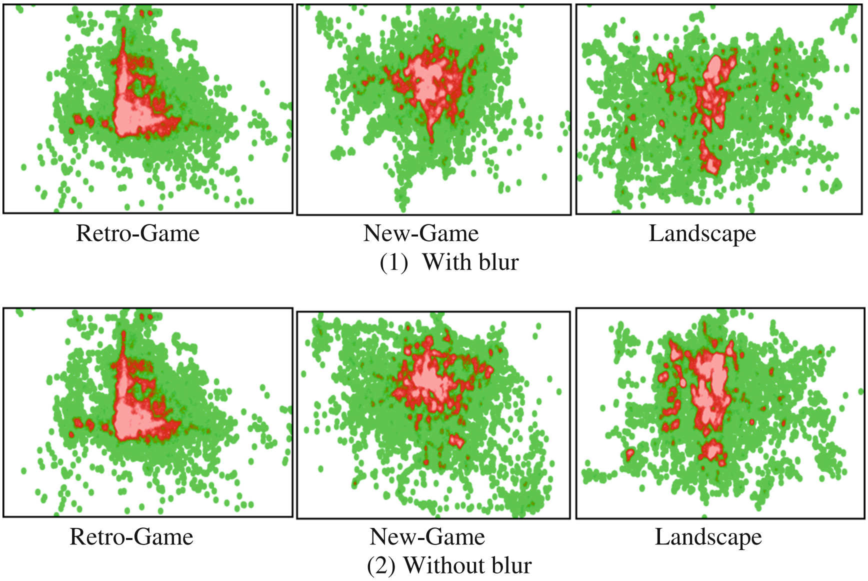


Fig. 9. Heat maps showing gaze point log.

The results show no large differences between Retro-Game and New-Game videos with and without blur. However, viewers tended to concentrate more on the center of the Landscape video without the blur (2) than that with it (1). In some cases, the gaze points tended to be slightly dispersed.

According to the visualized gaze point log data obtained in the additional experiments and the questionnaire results, the degree of concentration is increased by using this prototype system when viewing actual landscape videos, and the gaze points tend to be somewhat dispersed. In other words, the increased degree of concentration led to viewers paying attention to a wider range of the video. However, this result may not be reliable since the videos used in the additional experiment consisted of only one video per category and differed from the videos used in the main experiment. Therefore, we aim to improve the reliability by repeating the experiment of measuring the gaze point log and increasing the number of videos for each category used. Furthermore, we aim to measure the gaze point log when other types of videos are viewed and clarify the degree of concentration and its relationship with other psychological and physiological impressions.

6 Conclusion

We investigated whether a viewer's experience of video content is enhanced by superimposing blurring using a Gaussian filter on the peripherally viewed area. In addition, we evaluated the method's effectiveness in a user-based evaluation experiment and clarified its usefulness and problems.

In this work, blurring effects were superimposed on video content. In the near future, we will clarify whether the viewing experience is enhanced when blur is superimposed on the screen while the viewer is actually playing a video game. Moreover, we will investigate the threshold of the Gaussian filter at which the blur is not noticed and attempt to develop a new system that does not impair visibility while enhancing the viewing experience. Furthermore, since the questionnaire results in Subsect. 4.1 showed that our method is useful for still images, we expect that it can be usefully applied to e-books.

This system adopts a Gaussian filter as a blurring effect and superimposes it on the video content, but its usefulness is low for video content that does not change much. Therefore, we are planning to implement a new effect for decreasing saturation and brightness in areas away from the center of the acquired gaze point data. Sensitive perception of light flicker is a characteristic of peripheral vision [3]. Michael et al. [9] investigated the effect of color perception of the user in an environment where the background and the object color change as a result of movement of the gaze point and in an immutable environment. By changing the color of the background and the object, they showed that the method extended the region of a color perceivable by a person and improved viewers' ability to discriminate objects. Therefore, decreasing saturation and brightness is expected to enhance the viewing experience. Also, their system seemed to work well for videos with few changes. Finally, we plan to analyze our system in more detail by increasing the number of research participants and acquiring more data.

Acknowledgements. This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

References

1. Morishima, S.: Dive into the movie -audience-driven immersive experience in the story. IEICE Trans. Inf. Syst. **91**(6), 1594–1603 (2008). Journal on Virtual and Mixed Reality. (Japanese edition), Special Section on Human Communication III
2. Strasburger, H., Rentschler, I., Jüttner, M.: Peripheral vision and pattern recognition: a review. J. Vis. **11**(5), 13 (2011)
3. Okatani, T., Ishizawa, T., Deguchi, K.: Gaze-reactive image display for enhancing depth perception by depth-of field blur. IEICE Trans. Inf. Syst. **92**(8), 1298–1307 (2009). (Japanese edition)
4. Hillaire, S., Lécuyer, A., Cozot, R., Casiez, G.: Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments. In: Virtual Reality Conference, pp. 47–50. IEEE, Nebraska (2008)
5. Mantiuk, R., Bazyluk, B., Tomaszewska, A.: Gaze-dependent depth-of-field effect rendering in virtual environments. In: Ma, M., Fradinho Oliveira, M., Madeiras Pereira, J. (eds.) SGDA 2011. LNCS, vol. 6944, pp. 1–12. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23834-5_1
6. Hirai, T., Nakamura, S., Yumura, T., Morishima, S.: VRMixer mixing video and real world with video segmentation. In: 11th Advances in Computer Entertainment Technology Conference, pp. 1–7. ACM, Funchal, Madeira (2014)
7. Kagawa, K., Itou, J., Munemori, J.: Effects of an intuitional pictograph comment function in a video sharing web system. Int. J. Inf. Soc. **2**(3), 94–99 (2010)

8. Hata, H., Koike, H., Sato, Y.: Visual guidance with unnoticed blur effect. In: Proceedings of the International Working Conference on Advanced Visual Interfaces, pp. 28–35. ACM, New York (2016)
9. Mauderer, M., Flatla, D.R., Nacenta, M.A.: Gaze-Contingent manipulation of color perception. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 5191–5202. ACM, New York (2016)

Characterizing the EEG Features of Inspiring Designers with Functional Terms

Qian Zhang, Jia Hao(✉), Qing Xue, and Yu Yan

School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, People's Republic of China
haojia632@bit.edu.cn

Abstract. This paper constructed an inspiring database containing functional terms, which was taken as the source of external stimuli provided to designers. We obtained EEG of two groups of designers based on design experiment. One group is provided with closely related functional terms, while another group is provided without stimuli. After processing these EEG, we found that there are different characteristics in the EEG for the two groups of designers. Our experimental results provide a basis for the study of design thinking using EEG.

Keywords: EEG features · Inspiring designer · Functional term

1 Introduction

The ability of creative design lies in the ability to create novel design solutions [1]. The design solutions are derived from designers' experience and knowledge [2], which also limit designers' space for generating novel design solutions [3]. One common phenomenon is that designers are often fixed on existing experience [4]. Therefore, using external stimuli to inspire designers is one of the important methods to assist concept design.

The key to inspire design is to reveal the influence mechanism of external stimuli on design thinking [5], and to provide designers with appropriate external stimuli based on this mechanism. To this end, many scholars have explored the impact of external stimuli on innovative design through cognitive experiments. Miller [6] proved that designers exposed to physical examples would produce ideas that were less novel and less functionally focused than those who were exposed to the 2D representation. Sarkar [7] provided designers with video/animation and audio, text, explanation, and other stimuli respectively, and the results showed that the types of stimuli have major influence on the number and quality of the generated ideas. Among diverse types of stimuli, texts were proved to be capable of inspiring designers. Because of the characteristics of being ubiquitous and easy to be measured, texts were widely used in the research community. For instance, Miller [1] showed that text stimuli yield ideas that would receive higher originality grades compared to a no stimulus condition, but practicality was not affected. Linsey [8] found that the addition of a function structure or more general functions (stated as active verbs), to the sketch-based concept de-sign improved designers' ability of finding novel solutions.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 371–381, 2018.
https://doi.org/10.1007/978-3-319-91238-7_30

However, the unknown internal mechanism that governing how functional terms influence design thinking makes the above results difficult to be interpreted. At present, scholars have come up with the methods of think aloud [9] and sketching behavior [10] to indirectly obtain the state of design thinking, but these methods are more or less influenced by subjective factors. Therefore, it is very valuable to study the objective description method of design thinking. Recently, physiological signals such as eye-tracking, galvanic skin response (GSR), electrocardiograms (ECG), and electroencephalograms (EEG) have been used by design researchers to describe the conceptual design process. Among them, the process of exploring design thinking based on EEG is an important theoretical research direction of design science [11]. The biological basis of brain activity is an electrochemical process, and EEG is a charge signal recorded by electrodes placed in various regions of the brain [12]. We can extract valuable information from the EEG to describe the conceptual design process, and then explore the impact of external stimuli on conceptual design.

The goal of this work is to use EEG to describe the design thinking. In order to achieve this goal, Sect. 2.1 constructed an inspiring database represented by functional terms, which was taken as the source of external stimuli provided to the designers. Then we constructed the FCM based on the database, by which it could export functional terms that are closely related to the design problem. Section 2.2 identified the experiment scheme and process. Section 3 calculated the sample entropy of EEG data that could represent the complexity, and then analyzed them qualitatively and quantitatively. Section 4 discussed the analysis results and Sect. 5 summarized the experiment results, which provide theoretical basis for how to assist designers in product innovation design in information provision.

2 Method

2.1 Obtain External Information

2.1.1 Build Functional Database

In conceptual design, functional modeling provides a direct method for understanding and representing an overall product function without reliance on physical structure [13]. In practice, to achieve repeatable and meaningful results from functional modeling, a right size of functional inspiring database is in need. Hirtz and Stone [14] built RFB, whose functional part consists of 8 major categories and 22 minor categories. However, this database is not big enough to cover all product features. Murphy constructs a set of functional terms through natural language processing [15]. They extracted 1,700 function terms from a large number of patents and established a large-scale functional database. With the enlargement of the size of a functional database, the designers will describe the same design concept in different ways and ambiguities may arise.

In order to solve this problem, we filtered and reorganized the above two to construct a proper functional database. We used RFB as a basis for the similarity of function terms calculated by WordNet as a screening indicator. Then, the filtered function terms were reorganized with RFB to obtain a functional database to inspire designers. The whole process contains the following 3 steps (Fig. 1).

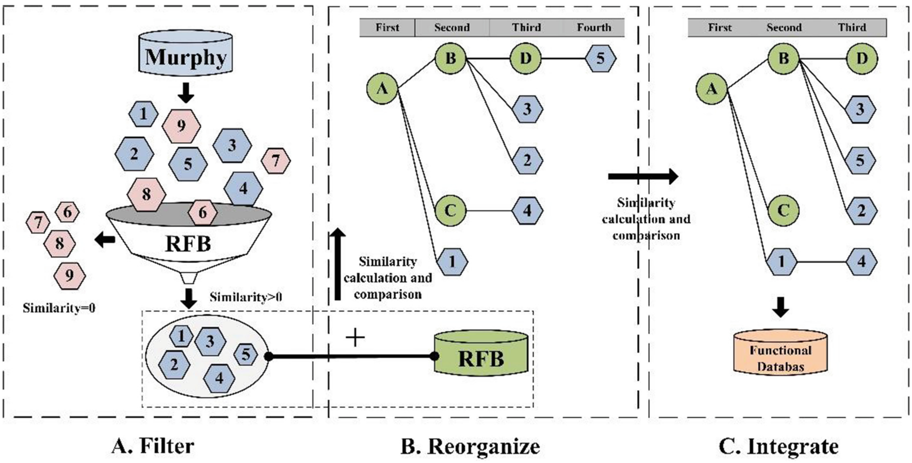


Fig. 1. The process of building Functional Database.

1. **Filter**: The terms in RFB and Murphy are all three-levels and share the same first level. Calculate the similarity between Murphy's second level, third level terms and the terms in RFB. Remove the terms of 0 similarity (Fig. 1A).
2. **Reorganize**: The filtered terms in Murphy are used as the low-level term of the most similar in RFB. The reorganized functional database has four levels (Fig. 1B).
3. **Integrate**: Find the term that has the highest similarity with the fourth level in the reorganized functional database. If the term is in the second level, it is regarded as the low-level term of it. If the term is in third level, it is regards as a term at the same level (Fig. 1C).

After the above three steps, a functional database of three levels is obtained, which includes 839 functional terms, and 8 terms in the first-level, 128 terms in the second-level and 703 terms in the third-level (Table 1). External information provided to the designers are derived from this database.

2.1.2 Construct FCM Based on Functional Database

According to the similarity between function terms in database, a Fuzzy Cognitive Map (FCM) was constructed, and the functional terms with high relevancy to the target product were output through the analysis and reasoning mechanism of the FCM. The construction of FCM contains the following 5 steps.

1. Calculate the semantic similarity between every two function terms in the database, and get Initial Matrix (IM) according to the number of function terms and the value of similarity. IM is a matrix of [n × n], where the element O_{ij} represents the similarity between the word i and the word j, and the elements O_{i1}, O_{i2}, …, O_{im} of column i form the vector V_i.

Table 1. Part of Functional Database

First level	Second level	Third level
Branch	Separate	Divide, Extract, Remove, Spray
	Distribute	
	Clean	Wash, Splash, Grind
	Split	Partition, Tear, Fracture, Rip
	Break	Dab, Scrub, Fork, Shuffle, Punch, Gore, Shave, Shear, Burst, Rupture, Breach, Tap, Cascade, Shower, Snap, Bore
	Machine	Filter, Comb, Crater, Trim, Fan, Surface
	Miss	Pop
	Cube	Crack
	Screen	
	Polish	
	Bleach	
	Buff	Sweep
	Chafe	
	Network	
	Plow	
…		
…	…	…
8	128	703

2. The largest element in V_i is assigned a value of 1, the smallest element is assigned a value of 0, and the others are scaled to [0, 1] to obtain Fuzzy Matrix (FZM), the element of which is F_{ij}.
3. Relationship Strength Matrix (FSM) is derived from FZM, the element of which is S_{ij}. In the database, relationship between the adjacent levels is unidirectional, and the direction is from lower level to higher level. The relationship between terms in the same level is bidirectional. When a relationship exists, $S_{ij} = F_{ij}$; otherwise, $S_{ij} = 0$.
4. All the elements in the SRM are arranged in an ascending order, and elements smaller than 20% of the quantiles are defined as 0, and then Adjacency Matrix (AM) is obtained, the elements of which is W_{ij}.
5. W_{ij} represents the final relationship between the term i and the term j, the direction of which is from i to j. Finally, we got the FCM.

2.1.3 Obtain Functional Terms

The function terms that can represent the product functions are taken as input, and after the calculation and reasoning of the FCM, the functional terms that are closely related to target product are obtained.

According to the function of the target product, functional terms are obtained from the database, and the Input Vector (IPV) is obtained. IPV is an n-dimensional row vector, and $I_{1j} = 1$ when the word j is selected, otherwise $I_{1j} = 0$. The IPV is multiplied by AM to obtain the Middle Vector (MV), and the MV is converted to IPV^1 to use a binary compression function. IPV^1 is multiplied by AM as the next input vector until IPV^n is stable. According to IPV^n, the corresponding terms are found from the database, that is, the function terms that are closely related to target product.

2.2 Experiment Design

2.2.1 Subjects

In this experiment, 16 students with a background of engineering design from Beijing Institute of Technology were taken as designers, and they were divided into two groups randomly. The two groups have the same number of students, and named Group A and Group B respectively.

2.2.2 Provision of Functional Terms

The functional terms provided to Group A is closely related to the target product. The experiment selected *High-rise auxiliary escape tool* as the target product, and then obtained functional terms from the database in Sect. 2.1.1 that can represent the product function. They are "*drop, stop, steady, guard, retard, ground, balance*". According to Sect. 2.1.2, the output function terms are "*hold, position, grip, grasp, clasp, seat, place, rest, center, balance, assist, order, ground, posture, fit, pose, cinch, cradle, steady, drop, retard, guard* ".

2.2.3 Data Collection

- The EEG data of conceptual design process was collected by EMOTIV Epoc+, which is with 16 electrodes and both sides of the mastoid are reference electrodes. The EEG data recorded by EEG was 128 Hz.
- Product design results was collected by using the sketch method.

2.2.4 Experiment Process

The designers first read the experimental instruction to clarify the for the product design experiment. For Group A, 12 functional words were randomly selected from Sect. 2.2.1 as external information. Group B was not provided with external information. Designers need to think about 10 min or so, draw a solution and dictate realization for this problem immediately after thinking. The EEG of designers in the entire process of conceptual design was recorded. The experiment process is shown in Fig. 2.

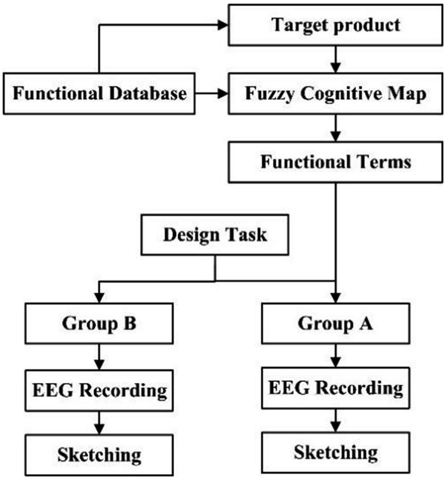


Fig. 2. Experiment process

3 Results and Data Analysis

EEG is a spontaneous, rhythmic discharge of the brain, and it is a non-linear, non-stationary electrophysiological signal [16]. Therefore, the analysis of EEG signals can use nonlinear parameters to characterize different neural activity states. In this work, sample entropy was selected as the complexity index to analyze designers' EEG, because of its good noise immunity [17].

We used a 1 s length window to calculate the sample entropy of the EEG, and select the parameters m = 2, r = 0.2SD. Based on this, the sample entropy was calculated, and then it was analyzed quantitatively and qualitatively.

Table 2. The mean EEG sample entropy of the designer at electrode F7, F8 and the T-test result

F7	A	Designer	1	2	3	4	5	6	7	8	P = 0.0356
		Sample entropy	0.5660	1.0899	0.6152	0.6057	1.1070	1.1730	0.8293	1.2387	
	B	Designer	9	10	11	12	13	14	15	16	
		Sample entropy	1.0985	1.3576	0.7619	1.2940	1.4065	1.3165	1.2836	0.9265	
F8	A	Designer	1	2	3	4	5	6	7	8	P = 0.0091
		Sample entropy	0.4769	1.0673	0.4229	0.5744	1.2944	1.6383	1.1836	0.6927	
	B	Designer	9	10	11	12	13	14	15	16	
		Sample entropy	2.0449	2.6683	0.5311	1.1635	1.9783	1.8364	1.5293	1.3826	

3.1 Quantitative Analysis

The sample entropy of EEG recorded by 14 electrodes in each group is shown in Table 2. The result was that the averaged sample entropy of group B (1.5741) was larger than that of group A (1.3488). We further made *T-test* on all 14 electrodes.

The *T-test* showed that the sample entropies of group B at F7 (P = 0.0356, left anterior temporal) and F8 (P = 0.0091, right anterior temporal) were higher than that of group A. There was no significant difference at other electrodes.

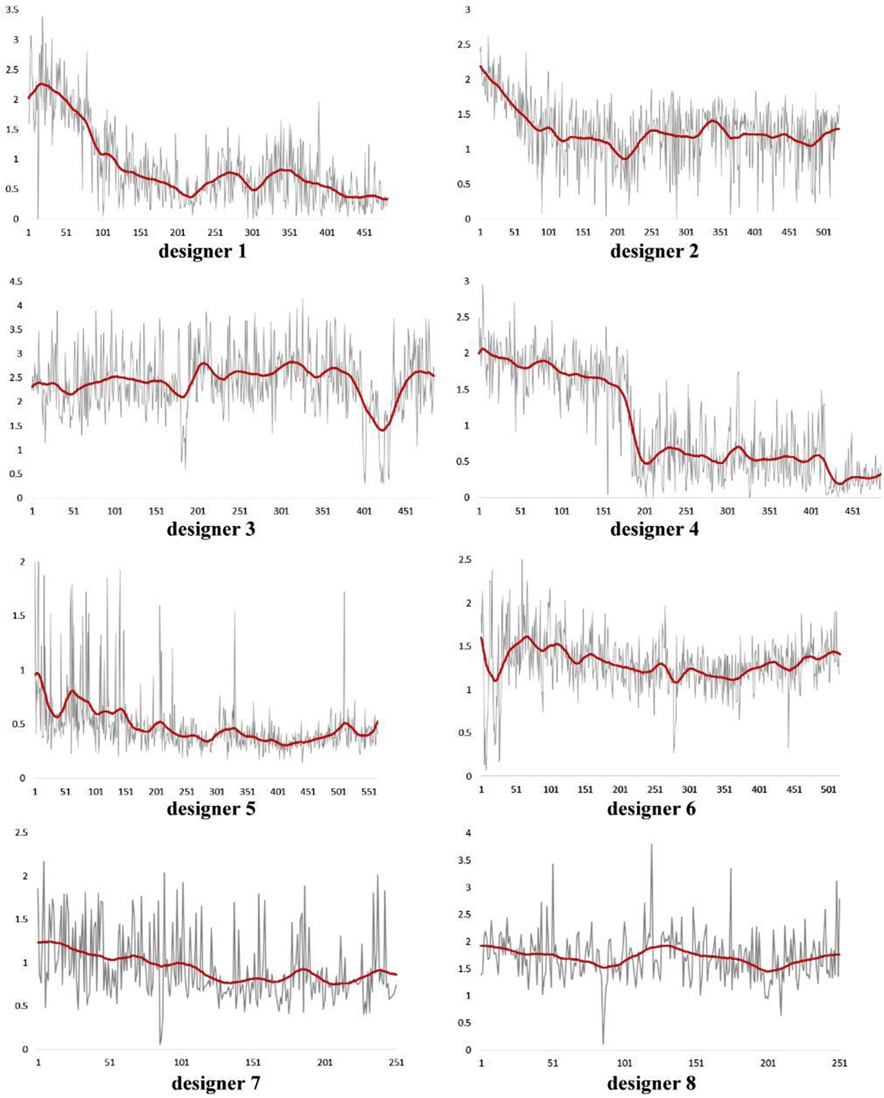


Fig. 3. Designers' trend of EEG sample entropy at electrode T7 in group A

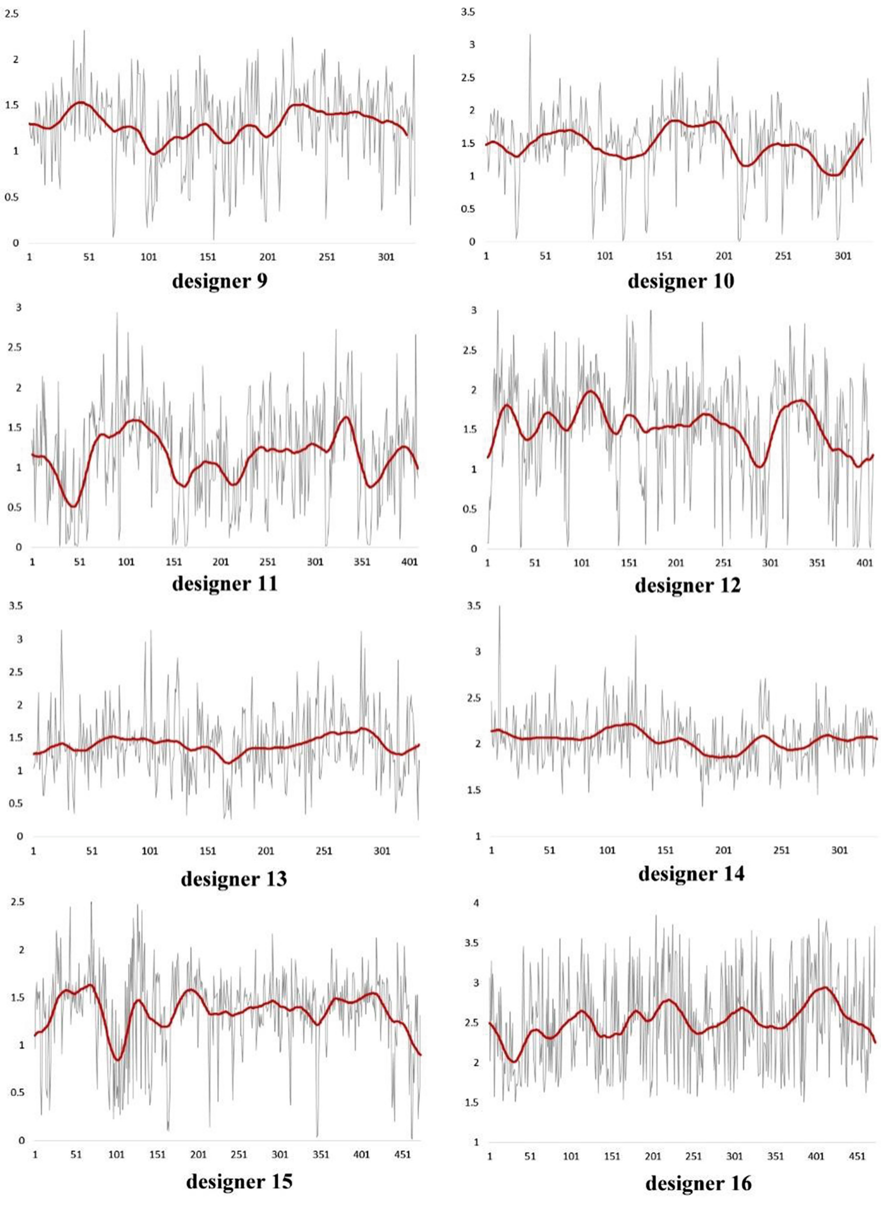


Fig. 4. Designers' trend of EEG sample entropy at electrode T7 in group B

3.2 Qualitative Analysis

We drew the trend of each window sample entropy. To reduce the impact of data fluctuations, we took a moving average method to process the data. The results were that the sample entropy of EEG in Group A was a shape of lower in the middle and higher on both sides at electrode T7 (left temporal lobe), T8 (right temporal lobe), P7 (left parietal lobe), P8 (right parietal lobe) and F8 (right anterior temporal) (Fig. 3). The sample entropy of Group B showed a stable or disorderly fluctuation (Fig. 4).

4 Discussion

Further, according to the position of the electrodes, we find the brain areas corresponding to the obtained brain electrical signal. Ratiocinate about how external stimuli affects design thinking is based on the function of brain partitioning and the characteristics of EEG. In 1909, Brodmann divided the cerebral cortex into 52 areas according to the type of cortical cells and the density of the fibers [18]. Currently, the study on the function of Brodmann Areas has been comprehensive [19].

The EEG is more complex when no functional terms were provided. Especially in the temporal polar, which is related to memory region. This finding may imply that without external stimuli, the memory of the designer awakens, which prompts the designer to associate other information and leads to complex brain activity. When provided with function terms closely related to the design problem, de-signers can obtain information from these function words for conceptual design, and the complexity of brain activity is relatively low.

When provided with external stimuli, the EEG complexity of the designer's inferior temporal gyrus, the temporal gyrus, the angular gyrus and the temporal pole region roughly follows a trend of first decreasing and then increasing. The inferior temporal gyrus is responsible for a high level of information processing, i.e. answering the question of what the information is. The temporal gyrus is the language processing area. Cornerback is part of the Wernicke District, located in the parietal lobe, with the function of semantic processing. The temporal polar regions are related to human memory.

When provided with external stimuli, the designers' brain are in a state of complex activity for information processing, language processing, and semantic processing. The designers first ponders over what the information provided is, and then manipulates the information. This process would awaken the designers' memory, which leads to complex brain activity. As the designer performs the conceptual design, this information has been transformed into working memory stored in the brain. In the process, the complexity of brain activity gradually decreases. Later in the conceptual design, the designer needs to perform semantic processing to output program-related information and generate a conceptual design. Once again, this complicates the designer's brain activity.

The complexity of the designer's brain activity remains stable or disorderly when no external stimuli is provided.

5 Conclusion and Future Work

The main purpose of this paper is to explore the impact of external stimuli on the designers' EEG features, and to ratiocinate the inner cognitive process of conceptual design based on EEG features. This paper selected functional terms as the external stimuli. We screened current functional terms and built a well-adapted inspiring database. Combining with FCM, we can deduce the functional terms that are closely related to target product. Obtained the designer's EEG and combined with the complexity of the analysis results and functional division of the brain, we can ratiocinate the process of design thinking.

Conclusions are as below:

1. Designers without any external stimuli have more complex brain activities than designers with external stimuli. This promotes the awakening of associations and memories, and the brain activity in this process is fluctuating.
2. When provided with external stimuli closely related to design problem, the designers awaken the memory during information processing and then it changes to working memory. Then designers output the conceptual design after the semantic processing. In the process, the complexity of brain activity first decreases and then increases.

Future work is to partition the conceptual design process according to the characteristics of EEG and select relevant indicators to summarize the EEG features and thinking characteristics of each area. Moreover provide different external stimuli according to the different thinking features that the designer presents in the conceptual design areas to assist the designers.

Acknowledgments. The authors would like to thank the anonymous reviewers for their valuable comments and thank the strong support provided by National Natural Science Foundation of China (NSFC 51505032) and Beijing Natural Science Foundation (BJNSF 3172028).

References

1. Howard, T.J.: Information management for creative stimuli in engineering design. University of Bath (2008)
2. Goldschmidt, G., Sever, A.L.: Inspiring design ideas with texts. Des. Stud. **32**(2), 139–155 (2011)
3. Gonçalves, M., Cardoso, C., Badke-Schaub, P.: What inspires designers? Preferences on inspirational approaches during idea generation. Des. Stud. **35**(1), 29–53 (2014)
4. Perttula, M., Sipilä, P.: The idea exposure paradigm in design idea generation. J. Eng. Des. **18**(1), 93–102 (2007)
5. Li, Y., Wang, J., Li, X., et al.: Creative thinking and computer aided product innovation design and research. Comput. Integr. Manuf. Syst. **9**(12), 1092–1096 (2003)
6. Miller, S.R., Toh, C.A.: The impact of example modality and physical interactions on design creativity. J. Mech. Des. **136**(136), 543–552 (2014)
7. Sarkar, P., Chakrabarti, A.: The effect of representation of triggers on design outcomes. AI EDAM: Artif. Intell. Eng. Des. Anal. Manuf. **22**(2), 101–116 (2008)
8. Linsey, J.S., Wood, K.L., Markman, A.B.: Modality and representation in analogy. AI EDAM **22**(2), 85–100 (2008)
9. Coughlan, P., Suri, J.F., Canales, K.: Prototypes as (design) tools for behavioral and organizational change: a design-based approach to help organizations change work behaviors. J. Appl. Behav. Sci. **43**(1), 122–134 (2007)

10. Jonson, B.: Design ideation: the conceptual sketch in the digital age. Des. Stud. **26**(6), 613–624 (2005)
11. Papalambros, P.Y.: Design science: why, what and how. Des. Sci. **1**, 38 (2015)
12. Wang, N.: Brain Cognition and Network Based on EEG Signal. Beijing University of Science and Technology Information (2015)
13. Wood, K.L.: Functional analysis: a fundamental empirical study for reverse engineering, benchmarking and redesign. ASME DETC (1997)
14. Hirtz, J., Stone, R.B., Mcadams, D.A., et al.: A functional basis for engineering design: reconciling and evolving previous efforts. Res. Eng. Design **13**(2), 65–82 (2002)
15. Murphy, J., Fu, K., Otto, K., et al.: Function based design-by-analogy: a functional vector approach to analogical search. J. Mech. Des. **136**(10), 470–495 (2014)
16. Meng, X., Ouyang, K.: Several nonlinear dynamic analysis methods of EEG signal. Beijing Biomed. Eng. **3**, 135–140 (1997)
17. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circ. Physiol. **278**(6), H2039 (2000)
18. Guillery, R.W.: Brodmann's 'localisation in the cerebral cortex'. J. Anat. **196**(Pt 3), 493 (2000)
19. Geyer, S., Dinse, J.: Brodmann's Areas. Reference Module in Neuroscience and Biobehavioral Psychology (2017)

Emotion and Attention Recognition

Speech Emotion Recognition Integrating Paralinguistic Features and Auto-encoders in a Deep Learning Model

Rubén D. Fonnegra and Gloria M. Díaz(✉)

Instituto Tecnológico Metropolitano, Medellín, Colombia
{rubenfonnegra,gloriadiaz}@itm.edu.co
http://www.itm.edu.co/

Abstract. Emotions play an extremely important role in human decisions and interactions with both other humans and machines. This fact had promoted development of methods that aim to recognize emotions from different physiological signals. Particularly, emotion recognition from speech signals is still a research challenge due to the large voice variability between subjects. In this work, paralinguistic features and deep learning models are used to perform speech emotion classification. A set of 1582 INTERSPEECH 2010 features is initially extracted from the speech signals, which are then used to feed a deep convolutional stack auto-encoder network that transform those features in a higher level representation. Then, a multilayer perceptron is trained to classify the utterances in one of six emotions: anger, fear, disgust, happiness, surprise and sadness. The size of the auto-encoders was evaluated for 4 different architectures, in terms of performance, computational cost and execution time for obtaining the most suitable configuration model. Thus, the proposed approach was twofold evaluated. First, a 5-fold cross-validation strategy was performed using 70% of the samples. Then, the best network architecture was used to evaluate the classification in a validation set, composed of the remaining 30% of samples. Results report an overall accuracy of 91.4 in the 5-fold testing stage and 61, 1 in the validation set.

Keywords: Speech emotion recognition · Deep learning
Paralinguistic features · Auto-encoders

1 Introduction

Emotion recognition is a growing field of research, which has been explored from different physiological sources such as facial expressions, electroencephalography signals, blood volume pulse, among others [1]. From them, the voice signals continue being one of the most used due to the feasibility of implementing end-user

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 385–396, 2018.
https://doi.org/10.1007/978-3-319-91238-7_31

systems that only requires a microphone for capturing the input data [1]. However, voice signal based emotion recognition is still a challenging task, due to the natural intra and inter-variability of the voices expressing a specific emotion. For leading with those challenges several approaches for characterizing voice signals have been proposed [2]. Among of them, INTERSPEECH 2010 paralinguistic challenge features have been considered as representative descriptors reflecting paralinguistic information assessment [3]. Those are a set of quantitative features that describe intensity, speaking variations, envelope and vocalization changes; which had shown to be one of the most feasible approaches for characterizing audio signals [4–6].

On the other hand, deep learning techniques have been recently proposed as a feasible strategy to perform end-to-end classification models, in which preprocessing and feature extraction steps can be leave out. In particular, novel methods for analyzing speech signals using deep learning approaches have been proposed, either as feature extraction approaches [7,8] or as classification models that use conventional acoustic descriptors [9]. Thus, Deep Belief Network (DBN) and auto-encoders architectures have been used to create higher representations from low-level descriptors [10]; while extreme learning machines, multilayer perceptrons and Long Short Term Memory (LSTM) architectures have been used as classification models [11–13].

In this paper, a deep learning classification model able to recognize the emotion of an utterance in voice signals is presented. The proposed approach integrates the INTERSPEECH 2010 paralinguistic features in a deep neural network composed of convolutional auto-encoders that transform input features in representations at a higher level. Four auto-encoder architectures were evaluated for achieving the best separability between six kind of emotions, i.e., anger, fear, disgust, happiness, surprise and sadness. In order to provide a comparison baseline, proposed approach was evaluated using voice signals extracted from a public emotional dataset named eNTERFACE [14].

The remainder of the paper is organized as follows: in Sect. 2 are presented the relevant state-of-the-art approaches, with a special focus on methods that had used the same database to perform model evaluation. Then, the proposed approach and the eNTERFACE database are described in the Sect. 3. The Sect. 4 presents both, the experimental results reported in terms of computational cost, execution time and classification accuracy, and a performance comparison with state-of-the-art methods. Finally, conclusions and future work are discussed in the Sect. 5.

2 Previous Works

As we mentioned above, development of emotion recognition systems based on speech signals has been a widely explored field in the last years. Several strategies had been proposed, either using audio speech signals alone or fusing it with other information such as provided by visual information of facial or body expressions and other physiological signals, as can be found in some recent literature reviews [15,16].

With the aim to provide a performance comparison baseline with the state-of-the-art techniques, this section describes those works, recently published, which report results using the same public dataset used for evaluating the proposed approach, i.e., eNTERFACE'05 database (see Sect. 3.1). In [9] a transfer learning model is proposed, in which 16 low-level descriptors (LLDs) and 12 functionals audio features are extracted using the openEAR toolkit. A transfer learning model, which includes auto-encoders based technique for feature transfer, maps a general structure of input characteristics by moving them from source to target to train a support vector machine (SVM). The main drawback of this model is that it is highly dependent on training reconstruction of data for knowledge transfer.

An SVM learning model was also used in [17] to differentiate between the six different emotions included in the eNTERFACE database, but the feature vector was composed of 7 short time-based features and three long-time based features extracted from the speech audio signals using JAudio toolkit. Reported results achieve an accuracy of 0.7857. Likewise, in [8] an audio emotion recognition system based on extreme learning machine (ELM) is proposed. Initially, a signal processing stage extracts multi-directional regression features (MDR) by pre-emphasize audio signals and frame them using hamming windows. Then, Fourier transform based spectral analysis and filter is applied using 24 Mel-scale Frequencies, obtaining 24 values per frame. At this point, a four-directional three-point linear regression is carried out to extract 96 features. The features are processed by an ELM classifier that achieves an overall accuracy of 0.6404.

Additionally, in [7] a sparse local discriminant canonical correlation analysis for multimodal information fusion was presented. In the case of audio emotion recognition, authors propose to apply a feature extraction stage to convert time domain signals into spectrograms with a 20 ms window and 10 ms overlap. The spectrograms are processed using the Principal Components Analysis (PCA) method to obtain 60 components, which are then considered as inputs of a sparse auto-encoder (400 units) to create a subspace representation, which is also used to train an SVM model. This approach improves previous results, achieving an accuracy of 0.74.

Finally, other works such as [13] has proposed SER systems based on a recurrent deep learning strategy. However, the feature description model is focused on the analysis of utterances where emotions can reach their highest expression peak, besides that suppress silence in sentences, or non-expressive words. That is to say, that only verbal features are used for characterizing the signal. This approach reaches an accuracy of 63.5% using an imbalanced database, by which results are not comparable.

In this work, we analyze the pertinence of the combination of convolution and encoding layers, with the aim of creating a deeper representation space for emotion classification task in audio signals.

3 Materials and Methods

In this work, it is investigated the discriminative capabilities of convolutional auto-encoders to perform human emotion analysis from low-level acoustic descriptors extracted from speech signals. Figure 1 illustrates the proposed approach, which is composed of three main stages: data preparation, training and validation. In the preparation stage, speech signals are extracted from the public emotional video database eNTERFACE [14]. Then, samples are split into two subsets: a training set, composed of the 70% of subjects, and a validation set, composed by the remaining 30% samples. As a large amount of data is required to train a deep learning model, training data were significantly augmented by extracting overlapping subsamples of two seconds long. Speech samples are then represented by a set of features, corresponding to INTERSPEECH 2010 Paralinguistic Challenge [3], extracted using the OpenSmile toolbox [18]. These features are then used as inputs of the convolutional auto-encoding neural network. The auto-encoder uses convolution to encode the low level paralinguistic features extracted in the previous step. Figure 2 shows details of the composition of a convolutional auto-encoder layer implemented in this work, which consists of a convolutional filter layer, which map the input features in a higher representation level. The convolutional

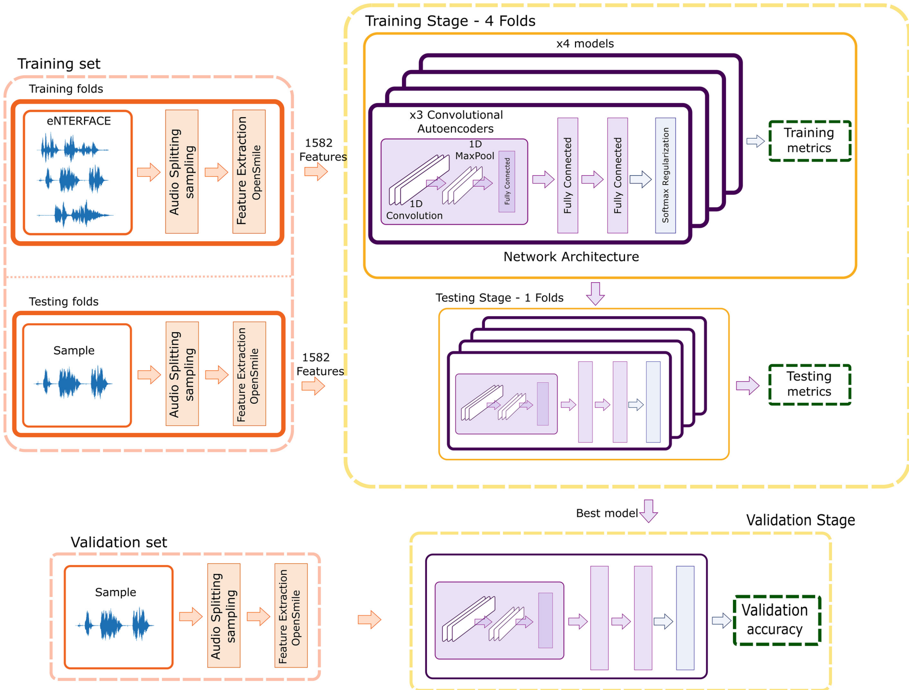


Fig. 1. Overview of the proposed approach for identifying emotions from speech signals.

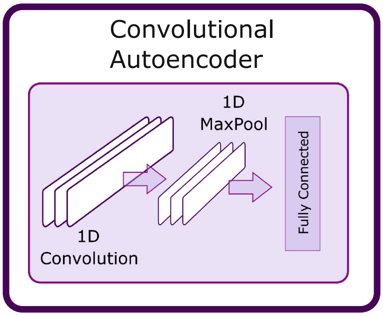


Fig. 2. Layers distribution for the convolutional auto-encoder used in this work

Fig. 3. Samples of audio signals for different emotional content, taken from eNTERFACE database

auto-encoder neural network is implemented as a classification model able to recognizing between six different emotions: anger, fear, disgust, happiness, surprise, or sadness. So, the output will be the predicted probability distribution for each emotion, according to an input data; a max pooling layer and a fully connected layer. In this stage, four different auto-encoders architecture are evaluated in a 5-fold cross validation strategy. The best auto-encoders architecture then used for validating the results in a completely unobserved data, the validation set.

3.1 The eNTERFACE'05 Database

The proposed approach was evaluated with a public, available and widely used dataset i.e. the eNTERFACE'05 database [14]. This is a bimodal database that contains video and audio signals of subjects expressing affective sentences in English language. 43 different non-professional actors (35 men and 13 women) coming from 14 different countries were asked to express emotions through specific sentences (5 sentences) for six different emotions (anger, fear, disgust, happiness, surprise, and sadness). So, the database is composed of a total of 1290 bimodal samples, corresponding to five videos (one per sentence) for six different emotion (five samples per emotion) taken from 43 subjects. Data were herein preprocessed separating audio signals from video frames, to remain only the audio signals. Figure 3 shows graphics of some sample signals from the database, in which the different emotional content is illustrated.

3.2 Data Augmentation Strategy

One of the disadvantages of Deep Learning based strategies is the amount of the data required to train the models for avoiding overfitting. Thinking of this, a data augmentation technique was considered for this work. The strategy for data augmentation consists of taking a multiple fix size windows (2 s) with 100 ms windows overlapping from every original signal. The size of the windows was chosen with the purpose of taking enough information to capture significant changes in the signals to identify emotional content on them. The signals (input samples) were all stored as individual files, to be processed in the feature extraction stage. The from the dataset we obtained 6163 samples, distributed in 4795 samples to train the model (training set) and 1368 for validation (validation set).

3.3 Feature Extraction

Due to the main characteristic of speech signals, which is data changes across time; a temporal modeling of data was considered as processing and feature extraction stage. Each input samples is firstly processed by the OpenSmile extractor, which generates a vector of 1582 low-level features, corresponding to the well known INTERSPEECH 2010 Paralinguistic challenge features [3]. Thus, the whole signal is described by 1582 features. The idea of using INTERSPEECH 2010 features is considering not only linguistic information from audio signals but also identify non-verbal patterns which could tell an emotion (Table 1).

The 1582 extracted features correspond to a set of 34 low-level descriptors (LLDs), with its corresponding delta coefficients namely: loudness raised to a power of 0.3, Mel Frequency Cepstral coefficients (MFCCs), logarithmic power of Mel-frequency bands, 8 line spectral pair frequencies from 8 linear prediction coding (LPC) coefficients, envelope of fundamental frequency contour, voicing probability of fundamental frequency, maximum and minimum value absolute positions, contour mean, slope of the contour linear approximation and, offset of the contour linear approximation. Besides, a set of 21 functionals were applied to 68 LLDs (1428), and 19 additional functionals were applied to the 4 pitch-based LLDs (152), such as standard deviation of the values in the contour, skewness, kurtosis, the smoothed fundamental frequency contour, among others. A complete description of those features can be found in OpenSmile documentation in [18].

3.4 Deep Learning Strategy for Audio Emotion Recognition

Modeling of signal characteristics was performed by a deep learning strategy with encoding instances; i.e. a convolutional auto-encoder - based neural network. Every encoder unit is a structure included in a neural network, that use convolution operations to encode inputs for creating a higher - level nonlinear combination of the audio data. In this way, we preserve most relevant patterns from a chain of events in the audio data. The encoding network implementation allows to determine an output sequence according to an input in the network, and the structure of the units allows to store information from the context of each sample.

Table 1. INTERSPEECH challenge 2010 descriptors and functionals extracted with the OpenSmile toolbox. LSP = Linear spectral pairs, DDP = Double delta of jitter

Descriptor	Functional
PCM loudness	Max/Min (position)
Mel Freq Cepstral Coefficients [0-14]	Arith, mean, std dev
Log. Mel Freq. Band [0-7]	Skewness, kurtosis
LSP [0-7]	Lin. regression slope, offset
F0 sub-harmonics	Lin.regression error
F0 envelope	Quartile 1/2/3
Voicing prob	Quartile range 2-1/3-2/3-1
Jitter local	Percentile 1/99
Jitter DDP	Percentile range 99-1
Shimmer local	Up-level time 75/90

The proposed network is composed of multiple convolutional encoding layers which allow to encode the inputs as a combination of the components using convolution. With the convolution, we can take advantage of the operation to express the inputs as the combination of multiple values, in order to create a higher lever model for recognizing emotions. The convolution for the input ($f(x)$) with a filter ($h(x)$) is described in the Eq. (1). Additionally, each convolutional auto-encoder unit is composed of a convolutional layer, a max - pooling layer and a fully connected layer. We use 3 convolutional auto-encoder layers and included a fully connected network to perform classification between the six emotions. A complete description of the model and the composition of a single convolutional auto-encoder is shown in Fig. 1.

$$f(x) * h(x) = \sum f(u)h(u-x) \tag{1}$$

The convolutional auto-encoder network is optimized during the training stage using the stochastic gradient descendant based algorithm named Adaptive Moment Estimation (Adam) optimizer [19]. Adam optimization is a momentum based learning algorithm (mean and variance), which allows single parameter tuning (such as Adagrad and RMSprop) considering gradients initialization and small decaying rates. These conditions significantly improve parameters optimization to increase accuracy and to avoid divergence during the training stage. Adam moment estimation and optimization rules are described by the Eqs. (2), (3) and (4) respectively, where the optimizer constant values were set to $\beta_1 = 0.9, \beta_2 = 0.999, \alpha = 0.001, \varepsilon = 10^{-8}$ (being α = Learning rate) during training.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t \tag{2}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2 \tag{3}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \epsilon} m_t \tag{4}$$

4 Results

4.1 Experimental Setup

Four auto-encoders based models were evaluated, each containing three convolutional encoding layers and two fully connected dense layers. Every model uses the same learning rate (0.001), Max pooling size (2), activation function (linear), weights initialization (according to initialization described in [20] with uniform distribution), Adam optimizer and preserving the same linear relationship between the number of units in the layers. Thus, the first model contains 32 convolutional encoding units in the first layer, 16 in the second layer and 32 in the third layer. The second model contains 64 units, 32 and 64 respectively; the third model contains 128 units, 64 and 128; and the four model contains 256 units, 128 and 256. The fully dense layers contains 384 units each, before getting to the output, which contains six units (one unit per emotion).

4.2 Performance Evaluation

In order to perform a comprehensive evaluation, we split the database into two segments: a training set of samples (70% of subjects) and a validation set of samples (remaining 30% of subjects). The training set was used for evaluating the performance of the four auto-encoder architectures, in order to select the one that reported the best performance as the model to be used to predicting emotion in the validation set. A 5-fold cross-validation strategy was implemented in the training stage, from which overall accuracy, execution time and memory consumption were computed. Overall accuracy was computed as the mean of the testing stage of five folds, the execution time during training is the one taken for the network to optimize all parameters for one epoch, and the memory consumption is the memory used for the algorithm to store network parameters.

All models were trained and tested using a NVIDIA Quadro K4000 under CUDA v7.5 and cuDNN 5.10; running using an Ubuntu 14.04 operative system along with the Theano + Lasagne Deep Learning framework. It is necessary to emphasize that we have not been carried out any other experiment with a greater number of units, given the hardware limitations to run the model (not enough memory requirements).

4.3 Experimental Results

Auto-encoder Architecture Selection. Figure 4 shows the obtained results for every auto-encoder architecture using a different amount of units per layer. Overall training and testing accuracies, computed as the mean of results in a 5-folding cross-validation strategy, are plotted in the Performance (a) graphic. As can be observed, the model with $128 - 64 - 128$ neurons in every layer reported the best accuracy. It is remarkable the significant accuracy lost reported by the model with $256 - 128 - 256$ neurons, in comparison with the previous one. This could be explained by a higher overfitting during parameter optimization.

Additionally, plots of execution time in seconds (b) and computational cost in Mib (c) are also presented. As can be expected, the computational cost is increased proportionally to the density of the learning network.

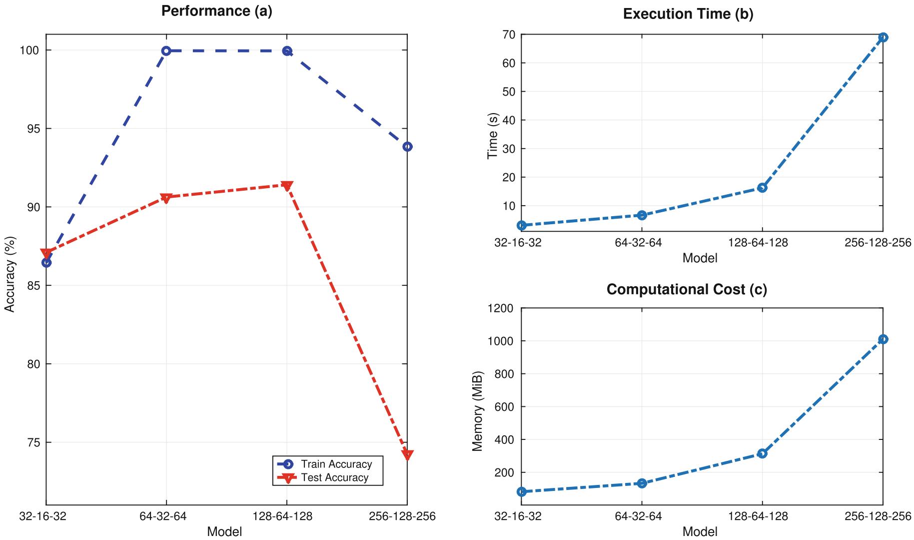


Fig. 4. Efficiency results for each model using different amount of units for convolutional auto-encoders. (a) Overall accuracy (in %) during training and testing, obtained in the 5-fold cross validation for each model. (b) Execution time (in s) for a training epoch in each model. (c) RAM Memory consumption (in MiB) for each model.

Performance Validation. Once the best auto-encoder architecture was selected in terms of overall accuracy, the model corresponding to the $128-64-128$ architecture was used to classify the emotions in samples from the validation set. The model achieved an accuracy of 61.103 in the validation set, which is a promising result, taking into account that the data were not known previously by the model. Table 2 shows the confusion matrix generated by this evaluation, in which happiness and sadness obtained the highest positive rate, with a major confusion with anger and disgust, respectively. This could be due to the similarity of the movements related to facial muscles during the expressiveness. On the other hand, the larger confusion is reported to fear and surprise emotions. In both cases, the model confuses with sadness; additionally, in other cases, sadness get a confusion rate higher in most cases. This might indicate a higher fitting from this class performed by the model.

Performance Comparison. Table 3 presents a comparison of some of the most relevant state-of-the-art works, which had used the eNTERFACE database for

Table 2. Confusion Matrix for audio emotion recognition using the data validation set

Emotion	Predictions					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**238**	29	8	30	41	11
Disgust	39	**123**	7	23	42	2
Fear	16	18	**86**	15	47	12
Happiness	18	11	5	**131**	9	7
Sadness	9	22	17	2	**209**	9
Surprise	18	6	0	18	38	**44**

evaluation purposes. The overall accuracy obtained by the 5-fold cross validation in testing stage is here reported due that a similar approach is the used in those works. As can be observed, the proposed approach outperforms previous work in more than 15%.

Table 3. Performance comparison of the proposed strategy with previous state-of-the-art works

Reference	Accuracy rate
Deng et al. (2013) [9]	0.591
Hossain et al. (2016) [8]	0.640
Dobrivsek et al. (2016) [21]	0.725
Fu et al. (2017) [7]	0.740
Yan et al. (2016)[22]	0.763
Porial et al. (2015) [17]	0.785
Proposed approach	**0.914**

5 Conclusions and Future Work

This paper presented a speech-based emotion recognition model, which implements a deep convolutional auto-encoder architecture as a classification model able to distinguish between six different emotions (anger, disgust, fear, happiness, sadness, and surprise) from speech signals featured by the INTERSPEECH 2010 Paralinguistic challenge using the OpenSmile software. Performance evaluation was carried out using a public and well-known database, which was split into a training set of data (70% of subjects) and a validation dataset (remaining 30%). The proposed approach outperforms previous works in a 5-fold cross validation model, showing that the combination of acoustic features and deep learning auto-encoders are a feasible approach for recognizing emotions of an utterance produced by a speaker. Besides, a second performance evaluation, using a validation set of data (30% of total samples) is presented, in which we obtained promising results, thinking into the implementation of end-user emotion recognition interfaces.

Auto-encoders configuration was also evaluated, showing that these configurations affect performance results but also the computational cost of training the model. A trade-off between computational cost and detection accuracy must be found. For doing so, parameter optimization strategies should be implemented.

As future work, performance evaluation with more larger datasets could be carried out. Additionally, given the difference between the voice corpora of male-female participants, we will evaluate the proposed strategy to a gender-dependent case for improving the performance metrics of the general speech emotion recognition system. Finally, we suggest combining the proposed approach with another kind of information (such as video) in order to improve emotion recognition accuracy from bimodal signals.

References

1. Ali, M., Mosa, A.H., Machot, F.A., Kyamakya, K.: Emotion recognition involving physiological and speech signals: a comprehensive review. In: Kyamakya, K., Mathis, W., Stoop, R., Chedjou, J.C., Li, Z. (eds.) Recent Advances in Nonlinear Dynamics and Synchronization. SSDC, vol. 109, pp. 287–302. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-58996-1_13
2. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. **43**(2), 155–177 (2015)
3. Schuller, B.W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S.S., et al.: The interspeech 2010 paralinguistic challenge. In: Interspeech, vol. 2010, pp. 2795–2798 (2010)
4. Kaya, H., Salah, A.A., Karpov, A., Frolova, O., Grigorev, A., Lyakso, E.: Emotion, age, and gender classification in children's speech by humans and machines. Comput. Speech Lang. **46**(Supplement C), 268–283 (2017)
5. Zhang, Y., Liu, J., Hu, J., Xie, X., Huang, S.: Social personality evaluation based on prosodic and acoustic features. In: Proceedings of the 2017 International Conference on Machine Learning and Soft Computing, pp. 214–218. ACM (2017)
6. Jassim, W.A., Paramesran, R., Harte, N.: Speech emotion classification using combined neurogram and interspeech 2010 paralinguistic challenge features. IET Signal Process. (2017)
7. Fu, J., Mao, Q., Tu, J., Zhan, Y.: Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. Multimed. Syst., 1–11 (2017)
8. Hossain, M.S., Muhammad, G.: Audio-visual emotion recognition using multidirectional regression and ridgelet transform. J. Multimodal User Interfaces **10**(4), 325–333 (2016)
9. Deng, J., Zhang, Z., Marchi, E., Schuller, B.: Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 511–516. IEEE (2013)

10. Schmidt, E.M., Kim, Y.E.: Learning emotion-based acoustic features with deep belief networks. In: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 65–68 October 2011
11. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech, pp. 223–227 (2014)
12. Cibau, N.E., Albornoz, E.M., Rufiner, H.L.: Speech emotion recognition using a deep autoencoder. Anales de la XV Reunion de Procesamiento de la Informacion y Control **16**, 934–939 (2013)
13. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231. IEEE (2017)
14. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: 2006 22nd International Conference on Data Engineering Workshops, Proceedings, p. 8. IEEE (2006)
15. Alva, M.Y., Nachamai, M., Paulose, J.: A comprehensive survey on features and methods for speech emotion detection. In: 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–6. IEEE (2015)
16. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. **21**(1), 93–120 (2018)
17. Poria, S., Cambria, E., Hussain, A., Huang, G.B.: Towards an intelligent framework for multimodal affective data analysis. Neural Netw. **63**, 104–116 (2015)
18. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. ACM (2010)
19. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
21. Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., Štruc, V.: Towards efficient multimodal emotion recognition. Int. J. Adv. Rob. Syst. **10**(1), 53 (2013)
22. Yan, J., Zheng, W., Xu, Q., Lu, G., Li, H., Wang, B.: Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. IEEE Trans. Multimed. **18**(7), 1319–1329 (2016)

Deep Learning Based Video Spatio-Temporal Modeling for Emotion Recognition

Rubén D. Fonnegra and Gloria M. Díaz(✉)

Instituto Tecnológico Metropolitano, Medellín, Colombia
{rubenfonnegra,gloriadiaz}@itm.edu.co
http://www.itm.edu.co/

Abstract. Affective Computing is a growing research area, which aims to determine the emotional user states through their conscious and unconscious actions and use it to modify the machine interaction. This paper investigates the discriminative abilities of convolutional and recurrent neural networks to modeling spatio-temporal features from video sequences of the face region. In a deep learning architecture, dense convolutional layers are used for analyzing spatial information changes in frames during short time periods, while dense recurrent layers are used to model changes in frames as temporal sequences that change across the time. Those layers are then connected to a multilayer perceptron (MLP) to perform the classification task, which consists in to distinguish between six different emotion categories. The performance was twofold evaluated: gender independent and gender-dependent classifications. Experimental results show that the proposed approach achieves an accuracy of 81.84%, in the gender independent experiment, which outperforms previous works using the same experimental data. In the gender-dependent experiment, accuracy was 80.79% and 82.75% for male and female, respectively.

Keywords: Deep learning · Facial emotion recognition
Spatio-temporal modeling

1 Introduction

The last decade has witnessed a tremendous development of interaction devices for different applications concerning entertainment fields [1], such as interactive video games, immersive environments, virtual and augmented reality, among others [2–5]. One of the most relevant study fields regarding human-computer interactions is the Affective Computing, which is intended to determine human emotional states to improve the interaction with machines. The study of emotional states has been studied from different information sources; such as biological signals [6], speech signals [7], still images and video sequences [8,9].

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 397–408, 2018.
https://doi.org/10.1007/978-3-319-91238-7_32

However, the use visual features of facial expressions is the most explored modality in the literature [10,11].

Emotion recognition based on facial expressions attempts to identify and characterize changes in motion units (in this case, those concerning to facial muscles), which allows representing the emotional expression of a person [12]. In this sense, two main strategies have been tackled, the first one focused on the characteristics of motion units that delineate facial patterns, which are detected in still images [13], and the second one attempts to describe the way in which those expressions change across image sequences [14]. Despite the high performance reported by some works that use still images, the dynamic of facial expressions had shown that can to achieves better performance [15]. However, development of visual descriptors that represent discriminative characteristics associated with different emotions is an open research challenge [16].

On the other hand, the recent development of high-performance specialized processing units (such as graphical GPUs) has increased the use of deep learning techniques based on artificial neural networks. Since 2012, when deep learning approaches outperformed classical computer vision algorithms in a very challenging task in 2012, i.e., classification 15 million images-within 22000 categories [17], this strategy has been used to solved several problems in many applications such as speech recognition, image classification, sequence modeling, among others. Specifically, the use of convolutional layers into the network architectures had shown be useful in the extraction of high-level spatial image representations without requiring to use classical preprocessing algorithms for featuring images.

In this work, we present an effective and robust emotion recognition system based on the analysis of video sequences, which allows analyzing patterns of movements of facial expressions during short time periods. For doing so, a deep learning architecture was designed to process video sequences and to identify between six emotions to be recognized. Convolutional layers are used to extract high-level representations of facial expressions across frames, and recurrent layers are used to model changes of the spatial characteristics extracted from the convolutional layers across time. Then, a multilayer perceptron classifies the sequence into one of the six possible emotions. The performance was evaluated using the well known eNTERFACE database, which was processed to obtain a broad set of training and evaluation data. Thus, video sequences were split into sets of one second, according to frames-per-seconds acquisition parameters.

This paper is organized as follows: in the Sect. 2, a brief summary of the related works described in the literature, which use the eNTERFACE database is presented. Section 3 describes the proposed spatiotemporal modeling based on deep neural network architectures. Besides, there are described the video preprocessing stage for videos, and the specific aspects of the proposed neural network architecture, including the number of layers, their main parameters, and stochastic optimizer. In Sect. 4 the experimental results are presented, and finally, in Sect. 5, the conclusions and future work are discussed.

2 Previous Works

Strategies concerning emotion recognition for affective computing models have been an active research field in the last years, because of a large amount of application, where the emotions can affect human's behavior and actions. State of the art presents several strategies for recognizing emotions [18]. However, those based on the analysis of facial expressions have been one of the most used to achieve this aim, which still represents important challenges such as illumination changes, pose variations (in still images [19]) and spatiotemporal modeling (in video or image sequences [20]). A comprehensive review of proposed approaches can be consulted in [21,22]. In this section, the most recent published works aiming to develop methods for recognizing emotions in the eNTERFACE'05 database will be presented to provide a comparison baseline.

Dobrivsek et al. [23] present an emotion recognition framework from video sequences, composed of different subsystems: the first subsystem implements a preprocessing stage, in which face is automatically detected and the image pixels are normalized using a histogram equalization approach. In the second subsystem, a dimensional subspace from preprocessed images to construct a prototypical template for the N emotions is performed by a simple eigen-decomposition of each emotion scatter matrix. In the third subsystem, a matching stage is applied for comparing encoding representations from faces, using a canonical correlation analysis, via singular value decomposition (SVD). Performance obtained using this strategy is highly dependent on the dimensionality of subspaces created for encoding the images, which increases the computational complexity of the algorithm.

Zhalehpour et al. [24] propose a strategy to select those frames from the original video in which the emotion expressiveness achieves the highest point. Then, those frames are processed as still images. Authors achieve the task by grouping all the video frames into K clusters based on dissimilarity between the local phase quantization (LPQ) features in each frame. They use a complete agglomerative link clustering algorithm, named dendrogram clustering algorithm (DEND-CLUSTER) to group the frames and then, select a frame whose average distance from the rest of the frames in the cluster is minimum. Thus, they compute the ideal selection measure (ISM) score based on gradient of each pixel, to arrange the selected frames in descendent order. The peak frames features are used for training a SVM with a linear kernel to classify emotions. This approach did not consider dynamic motion change of the face, which it is an expected condition in a real context.

Poria et al. [25] propose a model for emotion classification from visual data using an extreme learning machine. In their strategy they first make a manual annotation of every frame in the video, to subtract those frames with no relevant emotional content (neutral). Then, they perform a facial extraction of relevant points of the face (eyes, mouth, nose, eyebrows and chin) using the Luxand FSDK software to obtain a vector of features per frame. The feature vector for each video is obtained then, using coordinate - wise averaging from feature vectors of individual frames. Finally, they use a extreme learning machine (ELM) to perform classification task of video clips. Despite authors achieved a high performance with their strategy (81.21%), the main drawback lays on the manual annotation of the frames because it might elicit a significantly loss of frames with relevant emotional content to achieve the recognition.

In other works such as [26,27] propose emotion recognition systems that achieves significant performances. However, their restriction consists in the limited amount of samples for the classifiers; which bias the generalization models. In [28] is presented a model using big data towards 5G. This model includes different datasets for emotion classification task and achieves a significant performance compared with another state of the art strategies, reaching accuracies between 40.78% (disgust) and 76.9% (happy).

3 Materials and Methods

Figure 1 illustrates a graphical overview of the proposed approach, which is mainly divided into three stages: the first one, named preparation data, concerns to applying some preprocessing steps to the original data in the database, so they can be used for training and evaluating a deep learning based classification model. The second one concerns to training the deep learning classification model based on the spatio-temporal analysis of data, which aims to recognize emotions. And the last, the testing stage, in which the trained model is used for deciding the emotion from a new video (not used in the training stage).

3.1 Data Preparation

With the aim of reaching a valid experimentation stage comparable with previous works, the proposed approach was evaluated using a public, available and widely used dataset, i.e. the eNTERFACE'05 database [29]. It is a bimodal database, which contains video and audio signals of subjects expressing affective sentences in English language. The videos were acquired using a camera with a 25 fps sampling rate. 43 different non-professional actors (35 men and 13 women) coming from 14 different countries, which were asked to express emotions through specific sentences (five sentences) for six different emotions (anger, fear, disgust, happiness, surprise, and sadness). The database is composed of a total of 1290 bimodal samples, corresponding to five sentences per emotion (six emotions) for each of the 43 subjects. At this point, we separate audio signals from video frames and then, the audio signals were discarded. Some of the frames extracted from those image sequences are shown in Fig. 2.

Thus, 1290 image sequences were extracted and processed as follows in order to prepare data for be used in the experimental stage. Initially, a grayscale transformation was applied; then, original samples were processed for obtaining a dataset in which each instance contained the same number of frames, which is required to be processed by the deep learning model. Here, subsamples of one second were generated by splitting the original videos, without overlap. Additionally, because deep learning strategies requires a big amount of data

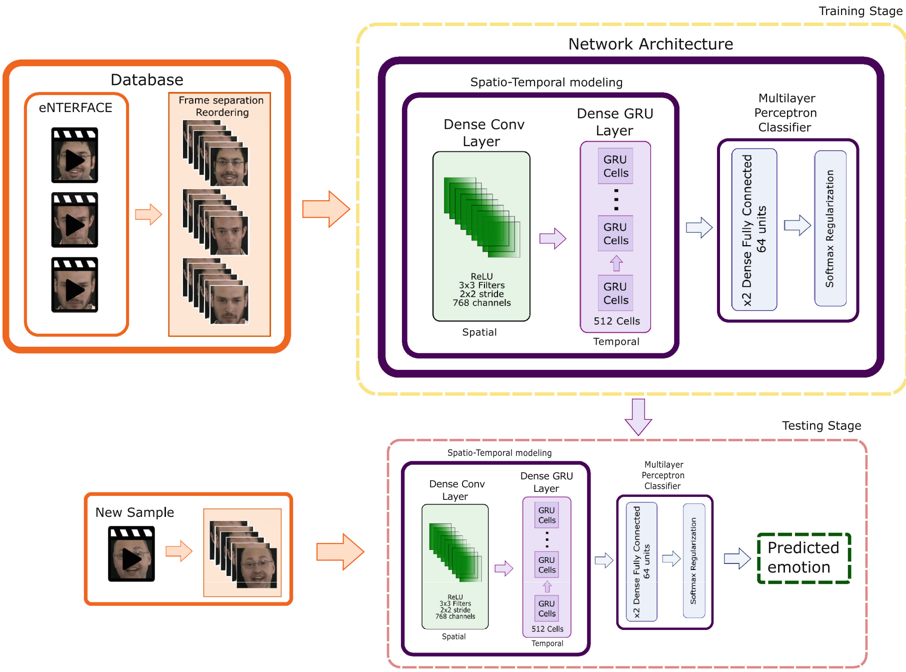


Fig. 1. Graphical overview of the proposed approach. In training stage, image sequences extracted from the database are used for trining a deep learning model composed of a dense convolutional layer, a dense GRU layer and a multilayer perceptron. In testing stage, the trained model is used for predicting the emotion oa a new image sequence.

for achieving optimal training parameters, a data augmentation strategy was applied by flipping every single frame in the sample, vertically. Additionally, four subjects in the database were discarded, due to lack of samples. Thus, a new experimental dataset was generated, which was composed of 5444 image sequences of 25 frames each.

3.2 Deep Learning Based Emotion Recognition from Image Sequences

Spatial and temporal relations from video samples described in Subsect. 3.1 were modeled using a deep learning architecture, which combines dense convolutional layers (for spatial information analysis) and recurrent units (for temporal modeling of data). The dense convolutional layer allows to extract spatial features from every sample. Given, each sample is compared with a sequence of still images, the convolution layers will learn main patterns of movements across the frames when emotions are expressed, and propagate them through the network layers. The dense convolutional layer contains 768 convolutional filters with size 3×3,

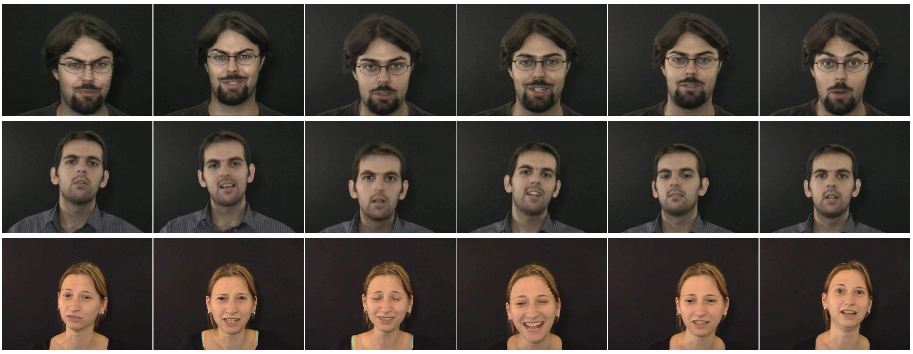


Fig. 2. Frame samples for different emotional states of three subjects, taken from videos of the eNTERFACE database

a full padding and stride size of 2×2. To maximize extraction of spatial information in the network units during the training stage, the selected activation function for the convolutional layer was the rectified linear units (ReLU) [30], which had demonstrated to generate more expressive models than classical activation function ([31,32]).

After convolutional layer, a dense recurrent layer was included to model the time dimension of video frames. So, spatial information extracted from the convolutional layers was considered, to be modeled as a sequence of temporal data which changes across the time. For doing so, a gated recurrent units (GRU cells) layer was implemented. GRU layer was used over Long-Short Term Memory units (LSTM cells - widely used in sequential models) due to they modulate the flow of information inside the unit, without having separate memory cells. This characteristic makes the GRU cell computationally more efficient. Besides, these units avoid vanishing gradient problem since the cell bypass multiple time steps, allowing the error to back-propagate easily [33]. The dense GRU layer contains 512 units with backward sequence processing (this means the cell take the sequence backward and then, reverse the output again), and a gradient clipping of 1 (to minimize computational cost).

Finally, a multilayer perceptron (MLP) was implemented in order to perform classification task. In this sense, the MLP was composed of 2 layers, with 64 units each, including a ReLU activation function. The output layer contains as many neurons as the number of emotions, with a softmax activation function, which allows to create a probabilistic density function of classes. The proposed approach is graphically illustrated in Fig. 1.

3.3 Adaptive Moment Estimation (Adam) Optimizer

The network parameters were optimized during the training stage with the stochastic gradient descent based algorithm, Adaptive Moment Estimation

(Adam) [34]. Adam optimizer consists of gradients updating based on momentums (first moment - mean, and second moment - variance). This technique allows adaptive single parameter tuning (such as Adagrad [35] and RMSprop [36]), but considering gradients initialization and small decaying rates during the training stage. These both considerations significantly improve parameters optimization, increase performance and avoiding divergence in the network. Adam moment estimation (mean and variance) and optimization rules are detailed described in Eqs. (1), (2) and (3) respectively, where constant values for optimizer during training stage were set to $\beta_1 = 0.9, \beta_2 = 0.999, \alpha = 0.001, \varepsilon = 10^{-8}$ (being α = Learning rate).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{1}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{2}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \epsilon} m_t \tag{3}$$

4 Performance Evaluation

4.1 Experimental Setup

Proposed approach was twofold evaluated; a gender-dependent case and a gender-independent case. In the gender-dependent case, the learning model was trained using samples from subjects of the same gender, i.e. males and females, separately. In the gender independent case, whole samples were merged in a unique dataset.

According to the last, in the gender-independent case, the experimental dataset was composed of 5444 samples (as was described in Sect. 3.1). So, for each fold, 4355 samples were used for training and 1089 for testing the model. In the gender-dependent case, two experiments were performed male-dependent and female-dependent. In the male-dependent were considered 31 subjects found in the database, from which were extracted 4366 samples, allowing to take 3492 for training and 874 for testing. In the female-dependent, only 8 subjects were found in the database, from which were extracted 1220 samples; 976 for the training and 244 for the testing.

In both cases, the evaluation was performed using an iterative 5-fold cross-validation strategy. For each fold, a learning model was trained in 50 epochs, obtaining a testing accuracy (Eq. 4) and a confusion matrix. Then, a the mean of the five folds were computed and reported as overall accuracy and overall confusion matrix.

$$Accuracy = \frac{Correct\ predicted\ emotions\ (hits)}{Amount\ of\ samples} \tag{4}$$

4.2 Experimental Results

Gender Dependent Case. Overall accuracies of 0.8275 and 0.8079 were obtained for female-dependent and male-dependent cases. Performance differences should be read with caution because the number of samples used for training each learning model. Likewise, Tables 1 and 2 report the confusion matrices for female-dependent and male-dependent cases, respectively. In the male-dependent case, similar results were reported for all the emotions, only the surprise reports a small performance (0.7086). On the other hand, in the female-dependent case, large differences were found. Sadness and surprise are clearly more distinguishable than disgust and fear. In this case, fear is confused with disgust, while disgust is more confused with the anger emotion. This could be expected, considering that female and male facial characteristics differ one from another.

Table 1. Confusion Matrix for emotion recognition in female.

Predictions	Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**0.8372**	0.1162	0.0232	0	0.0232	0
Disgust	0.1666	**0.7666**	0	0.0666	0	0
Fear	0	0.1555	0.6888	0.0444	0.0444	0.0666
Happiness	0.0512	0.0769	0.0512	**0.7692**	0	0.0512
Sadness	0.0222	0	0.0444	0	0.9333	0
Surprise	0	0	0	0	0.0789	**0.9210**

Table 2. Confusion Matrix for emotion recognition in male.

Predictions	Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**0.8154**	0.0952	0.0297	0.0476	0	0.0119
Disgust	0.0647	**0.8273**	0.0215	0.0431	0	0.0431
Fear	0.0447	0.0223	**0.8432**	0.0447	0.0149	0.0298
Happiness	0.0186	0.0373	0.0373	**0.8224**	0.0186	0.0654
Sadness	0.0233	0.0292	0.0584	0.0350	**0.8304**	0.0233
Surprise	0.0662	0.0198	0.0728	0.0529	0.0794	**0.7086**

Gender Independent Case. In the experiment where the gender of the samples was not considered; the proposed approach obtained an overall accuracy of 0.8184. The confusion matrix is shown in the Table 3. In this case, emotions with biggest recognition rate were disgust and surprise, obtaining 0.8620 and 0.8410 respectively, while fear and anger were most confused, showing a recognition rate of 0.7796 and 0.7881, respectively.

Table 3. Confusion Matrix for emotion recognition in gender independent case.

Predictions	Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**0.7881**	0.0508	0.0508	0.0677	0.0127	0.0296
Disgust	0.0632	**0.8620**	0.0344	0.0114	0.0057	0.0229
Fear	0.0395	0.0282	**0.7796**	0.0621	0.0338	0.0564
Happiness	0.0378	0.0378	0.0454	**0.8257**	0.0303	0.0227
Sadness	0.0186	0.0139	0.0372	0.0604	**0.8139**	0.0558
Surprise	0.0066	0.0198	0.0397	0.0463	0.0463	**0.8410**

In order to compare experimental results with those found in state of the art, Table 4 shows the most relevant studies for facial emotion recognition, using the eNTERFACE database as was specified in Sect. 2. As can be observed the proposed approach outperforms reported results in emotion recognition in a gender independent classification.

Table 4. Performance comparison with related works in the state of the art.

Authors	Recognition rate
Zhalehpour et al. [24]	40.00%
Hossain et al. [28]	40.78%–76.9%
Wang et al. [19]	72.47%
Rashid et al. [26]	74.15%
Poria et al. [25]	81.21%
Proposal Approach	**81.84%**

5 Conclusions and Future Work

In this work, we propose an effective framework for emotion recognition from video, which exploits convolutional recurrent layers in a deep learning architecture. We proposed a video emotion recognition using dynamic spatio-temporal modeling based on deep neural networks that allows to distinguish between six different emotions (anger, disgust, fear, happiness, sadness, and surprise). Learning model was designed with three main components, the first component was a deep learning strategy specially designed to perform spatial analysis from the video frames through dense convolutional layer; the second was a dense GRU layer, which aims to model time changes of the spatial information extracted in the first one. The last one was a dense multilayer perceptron, which performed the classification task. The performance was evaluated in gender-dependent and gender-independent cases, using the eNTERFACE database, which enables a comparison with state-of-the-art methods. According with the results, the proposed approach outperform previous works in the gender-independent case. Previous works reporting gender-dependent evaluation were not found.

As future work, it is required to evaluate the real effect of training specific classification models according to subject gender. For doing so, an homogeneous database is required. Additionally, other aspects such as Language, country origin and age could be also considered. On the other hand, we pretend to explore the use of our strategy in real time models using virtual gadgets and avatars for emotion recognition. Additionally, we pretend to continue exploring spatio-temporal modeling using deep learning approaches for multimodal information fusion analysis.

References

1. Blascovich, J., Bailenson, J.: Infinite reality: Avatars, Eternal Life, New Worlds, and the Dawn of the Virtual Revolution. William Morrow & Co., New York (2011)
2. Balducci, F., Grana, C., Cucchiara, R.: Affective level design for a role-playing videogame evaluated by a brain-computer interface and machine learning methods. Visual Comput. **33**(4), 413–427 (2017)
3. Bartsch, A., Hartmann, T.: The role of cognitive and affective challenge in entertainment experience. Commun. Res. **44**(1), 29–53 (2017)
4. Corneanu, C.A., Simón, M.O., Cohn, J.F., Guerrero, S.E.: Survey on rgb, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1548–1568 (2016)
5. Zhou, X., Shen, W.: Research on interactive device ergonomics designed for elderly users in the human-computer interaction. Int. J. Smart Home **10**(2), 49–62 (2016)
6. Bernal, G., Maes, P.: Emotional beasts: visually expressing emotions through avatars in VR. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2395–2402. ACM (2017)
7. Yan, J., Zheng, W., Xu, Q., Lu, G., Li, H., Wang, B.: Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. IEEE Trans. Multimed. **18**(7), 1319–1329 (2016)
8. Mavridou, I., McGhee, J.T., Hamedi, M., Fatoorechi, M., Cleal, A., Ballaguer-Balester, E., Seiss, E., Cox, G., Nduka, C.: FACETEQ interface demo for emotion expression in VR. In: 2017 IEEE Virtual Reality (VR), pp. 441–442. IEEE (2017)
9. Bekele, E., Bian, D., Peterman, J., Park, S., Sarkar, N.: Design of a virtual reality system for affect analysis in facial expressions (VR-saafe); application to schizophrenia. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(6), 739–749 (2017)

10. Marrero-Fernández, P., Montoya-Padrón, A., i Capó, A.J., Rubio, J.M.B.: Evaluating the research in automatic emotion recognition. IETE Tech. Rev. **31**(3), 220–232 (2014)
11. Goyal, S.J., Upadhyay, A.K., Jadon, R.S., Goyal, R.: Real-life facial expression recognition systems: a review. In: Satapathy, S.C., Bhateja, V., Das, S. (eds.) Smart Computing and Informatics. SIST, vol. 77, pp. 311–331. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-5544-7_31
12. Lien, J.J., Kanade, T., Cohn, J.F., Li, C.C.: Automated facial expression recognition based on FACS action units. In: 1998 Third IEEE International Conference on Automatic Face and Gesture Recognition, Proceedings, pp. 390–395. IEEE (1998)
13. Cheng, F., Yu, J., Xiong, H.: Facial expression recognition in jaffe dataset based on gaussian process classification. IEEE Trans. Neural Netw. **21**(10), 1685–1690 (2010)
14. Ji, Q., Moeslund, T.B., Hua, G., Nasrollahi, K. (eds.): FFER 2014. LNCS, vol. 8912. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-13737-7
15. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recogn. **36**(1), 259–275 (2003)
16. Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P., Quadrana, M.: Content-based video recommendation system based on stylistic visual features. J. Data Seman. **5**(2), 99–113 (2016)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
18. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
19. Wang, S.H., Phillips, P., Dong, Z.C., Zhang, Y.D.: Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. Neurocomputing (2017)
20. Yan, H.: Collaborative discriminative multi-metric learning for facial expression recognition in video. Pattern Recogn. (2017)
21. Mühl, C., Allison, B., Nijholt, A., Chanel, G.: A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. Brain-Comput. Interfaces **1**(2), 66–84 (2014)
22. Wang, S., Ji, Q.: Video affective content analysis: a survey of state-of-the-art methods. IEEE Trans. Affect. Comput. **6**(4), 410–430 (2015)
23. Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., Štruc, V.: Towards efficient multimodal emotion recognition. Int. J. Adv. Rob. Syst. **10**(1), 53 (2013)
24. Zhalehpour, S., Akhtar, Z., Erdem, C.E.: Multimodal emotion recognition based on peak frame selection from video. Signal Image Video Process. **10**(5), 827–834 (2016)
25. Poria, S., Cambria, E., Hussain, A., Huang, G.B.: Towards an intelligent framework for multimodal affective data analysis. Neural Netw. **63**, 104–116 (2015)
26. Rashid, M., Abu-Bakar, S., Mokji, M.: Human emotion recognition from videos using spatio-temporal and audio features. Visual Comput. **29**(12), 1269–1275 (2013)
27. Huang, K.C., Huang, S.Y., Kuo, Y.H.: Emotion recognition based on a novel triangular facial feature extraction method. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2010)
28. Hossain, M.S., Muhammad, G., Alhamid, M.F., Song, B., Al-Mutib, K.: Audio-visual emotion recognition using big data towards 5G. Mobile Netw. Appl. **21**(5), 753–763 (2016)

29. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: 2006 22nd International Conference on Data Engineering Workshops, Proceedings, p. 8. IEEE (2006)
30. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
31. Jarrett, K., Kavukcuoglu, K., LeCun, Y., et al.: What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153. IEEE (2009)
32. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8609–8613. IEEE (2013)
33. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
34. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(Jul), 2121–2159 (2011)
36. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw Mach. Learn. **4**(2), 26–31 (2012)

A Flexible, Low Power, Compact, Mobile Sensor for Emotion Monitoring in Human Computer Interaction

Lawrence K. Lam(✉) and Austen J. Szypula

University of Washington Bothell, Bothell, WA 98011, USA
llam@uw.edu

Abstract. Emotion-sensing is an important part in human computer interaction. Electrodermal activity measurement has been employed for a while to sense emotion. In this paper, the authors combine EDA, mobile technology, and flexible PCB substrate to implement a low power, compact emotion sensor which can be displayed on a smartphone. The flexible PCB circuit and sensing electrodes can be conformably wrapped around the user's fingers to form a compact integrated unit. The 1st generation sensor has overall size 60×60 mm and consumes 90 mW under regular Bluetooth recording. In the newer version, we house the unit in a silicone mold to form a robust package for versatile sensing environment. With the new package, we were able to record 8-h sleeping patterns.

Keywords: Wearable technology · Flexible substrate applications
Human computer interaction · Emotion sensors

1 Background

Lately, there is increasing research activity integrating artificial intelligence and robotics to make machine-robots more human. Sensing the emotion of a human is one of the essential elements to bridge that gap. Emotion-sensing of a human can be done by monitoring the physiological responses such as blood pressure, heart rate, and skin conductivity by using biosensors [1]. In this paper, we measure the skin conductivity; sometimes people call it electrodermal activity, EDA, to account for the emotional status of a person.

Researchers have come up with different kinds of devices to measure skin conductance. These devices vary in how they connect the sensors and display EDA signals [2–4]. Table 1 shows the comparison of some of these emotion sensors along with the 1st generation sensor we reported [5] and the improved version here, the 2nd generation sensor. All these devices are either use a wire or wireless connection from the emotion sensing board to a PC or smartphone. However, they are either bulky or require large batteries for power, limiting the devices' wearability and mobility.

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

© This is a U.S. government work and its text is not subject to copyright protection in the United States; however, its text may be subject to foreign copyright protection 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 409–418, 2018.
https://doi.org/10.1007/978-3-319-91238-7_33

Table 1. Comparison of emotion sensors

	HandWave 	**Shimmer3** 	**NUL-217** 	**Flexible BT Emotion sensor 1st Gen.**	**Flexible BT Emotion sensor in Silicone**
Smartphone	No	No	No	Yes	Yes
Conformable	Rigid PCB	Rigid PCB	Rigid PCB	Flexible PCB	Flexible PCB
Integrated Electrodes	No	No	No	No	Yes
Power Consumption	360mW	1.65W	52.25mW	90mW	78.1mW
Footprint (dim. and weight)	100mm x 33mm x 21mm 90 grams	65mm x 32mm x 12mm 28 grams	100mm x 50mm x 24mm 9.07 grams	72mm x 30mm x 10mm 11 grams	64mm x 50mm x 32mm 28 grams

In this paper, the authors present an improved version of emotion sensor with a smaller foot-print, lower power consumption, new built-in automatic calibration module and integrated sensing electrodes [5]. All these are housed in a silicon mold to form a robust sensing package system for longer run-time and more versatile environment recording.

2 The Sensor

2.1 Hardware

Like in the 1st generation sensor [5], the core of the 2nd generation sensor system is the same. It consists of a hardware component and a software component. The hardware circuitry consists of a sensing circuit based on a Bluno Beetle BLE microcontroller equipped with a 10-bit ADC and CC2540 Smart Bluetooth module and an INA333 instrumentation amplifier, to which the electrodes are connected. The electrodes are secured using 3.5 mm snap connectors to the flexible PCB circuit.

Construction of the 2nd generation sensor begins with the construction of a small, flexible PCB board. Next, copper tape traces were laid out with some resin epoxy holding the wires it would need to connect to the Bluno Beetle microcontroller and electrodes. Once all the traces passed the continuity test, all components and connectors were soldered on. Finally, the sensor is placed in a mold made from the same material as the silicone casting material (Figs. 4 and 5). The electrodes are the last parts to be added and simply snap into place.

The result was a working Bluetooth emotion sensor on a flexible PCB protoboard. Under normal operation, the current load placed on the battery is 21.11 mA at 3.7 V for the average power consumption of 78.1 mW.

Figures 1 and 2 show the 1st generation sensor and its circuit schematic. However, for the sensor reading to display properly, an initial voltage calibration needs to be performed. This is done manually in the 1st generation sensor, but for the newer version sensor, we incorporate a built-in, automatic voltage calibration using the PWM feature of our microcontroller. The 8-bit PWM channel on pin D5 is filtered, forming a DC equivalent analog voltage that can be used to shift the node voltage associated with the left half of our Wheatstone bridge circuit until it matches the node voltage on the right half of our circuit. Once this condition is met, the output is said to be 'zero-balanced', with optimum sensitivity to conductance changes as we've eliminated most of the common-mode voltage signal fed into the inputs of our gain amplifier [6].

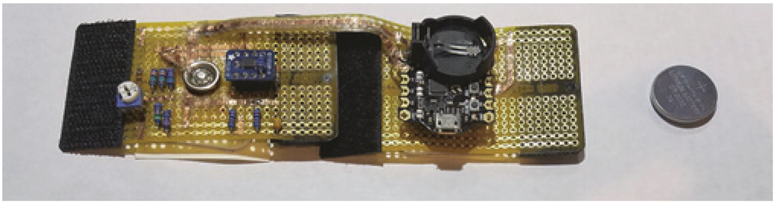


Fig. 1. 1st generation flexible sensor

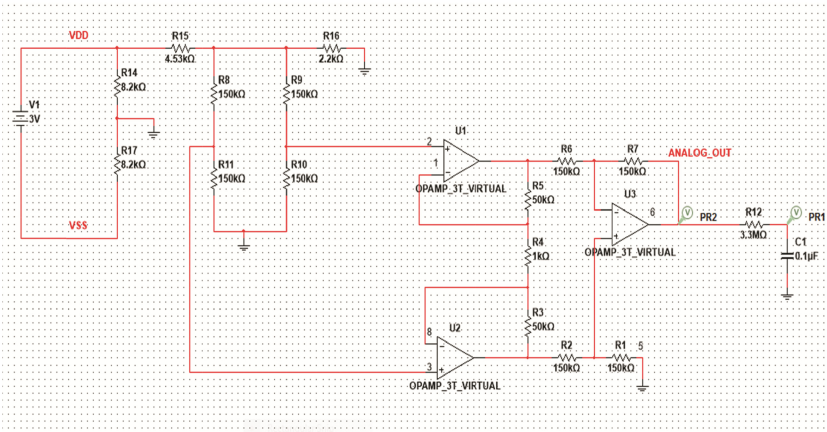


Fig. 2. 1st generation flexible emotion sensor schematic

Also, we house the circuit and electrodes inside a silicon mold to form a robust package for versatile environment usage such as sleeping. Figure 3 shows the newer generation packaged sensor. The four LEDs in Fig. 3 indicate power on, ready to sense, paired with phone, and battery charging status.

The newer generation silicone package also features an updated, 600 mAh Lithium Polymer rechargeable battery capable of powering the emotion sensor for a full 24 h under continuous-recording operation. The charge LED indicates the charge status of the battery when it is plugged-in but does not have the capability of alerting the user when it's almost drained. The next generation sensor will likely incorporate an even smaller rechargeable coin or Lithium Polymer battery and use one of the analog input pins of a microcontroller to monitor the sensor battery's percentage, generating a software interrupt when it gets too low.

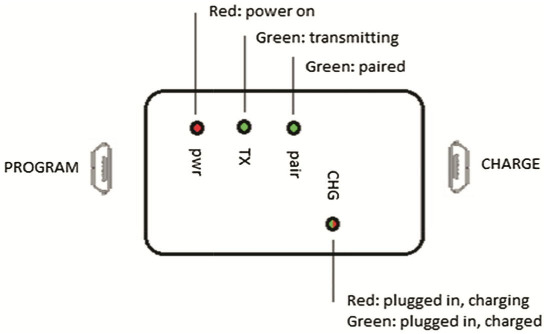


Fig. 3. Newer (2nd) generation device LED map

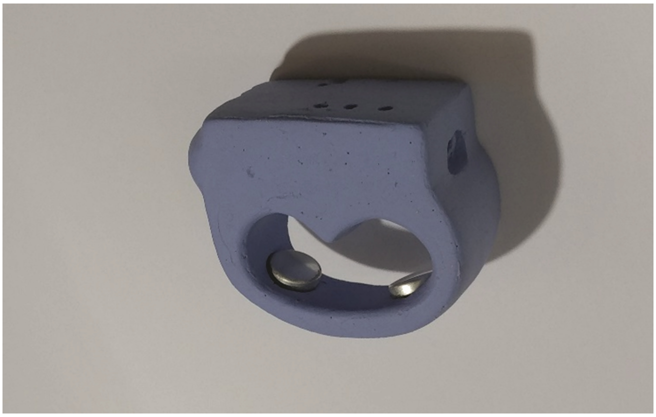


Fig. 4. Newer (2nd) generation sensor in a silicon mold package

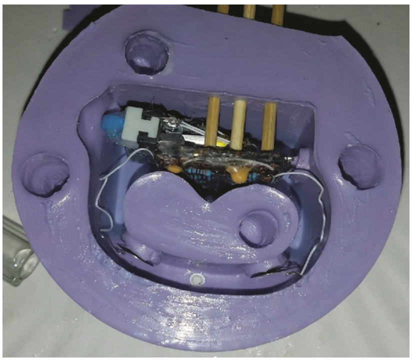


Fig. 5. Preparation for silicone casting

2.2 Software

The software portion is divided into two modules: the transmitting and the receiving modules. The Bluno Beetle microcontroller of the software represents the transmitting module, and the android (or any other BT enabled device) of the software represents the receiving module. Figure 6 shows the block diagram of the software management in the Bluno microcontroller and the Android Device.

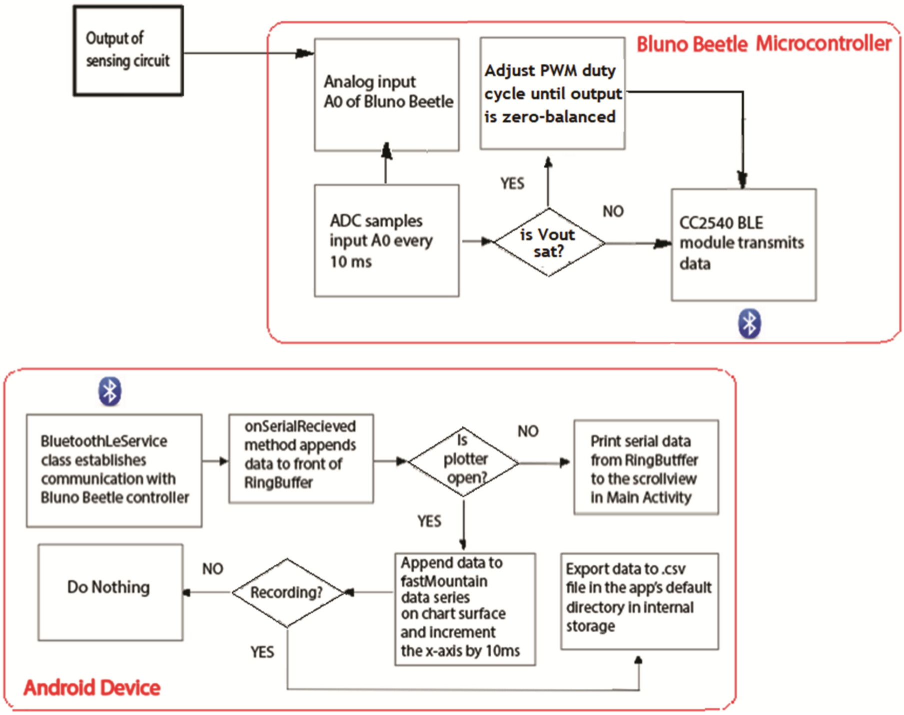


Fig. 6. Software management flowchart: Transmitting (Bluno Beetle) and receiving (Android) modules

The transmitting module samples, converts, and broadcasts the analog signal at a sampling period of 10 ms. The receiving module is programmed as a Bluetooth App. It receives the broadcast stream from the transmitting module and interprets the data from the transmitter.

The flexible Bluetooth emotion sensor prototype app incorporates only two Activities: the first includes buttons for connecting to the device and a ‘scrollview’ for displaying the incoming serial data once a connection has been established. The second Activity shows the plot and analysis window for the incoming data stream displayed on an android phone. Figure 7 shows the sample code of the Arduino sketch.

```
// Code and comments written by Austen Szypula for
// BEE 499 Flexible Emotion Sensing Device

// Sketch purpose:
// Reading analog signal fromm pin A0
// and sending the data to Android app.
// Digital pin D5 used as pwm for auto-
// balancing the wheatstone bridge.

// Hardware setup:
// Bluno Beetle <---Bluetooth---> Android 4.4+ phone/tablet

// input ----- A0

//Global Variables
int sensorPin = 0;                //pin number to use the ADC
int sensorValue = 0;              //initialization of sensor variable, equivalent to EMA Y

float EMA_alpha = 0.2;            //initialization of EMA alpha - choose between 0 and 1
                                  //effect is more pronounced with smaller EMA alpha
int EMA_result = 0;               //initialization of EMA result, or the result of the EMA at some value t
int pwm = 145;                    //default pwm output
int binCalIndex = 0;
int clockDivider = 220;
int clockDividerIndex = 0;

boolean calibration = true;       //calibrate on startup and when button is pressed.
boolean calibrationBackup = true;  //prevents saturation

void setup() {
  Serial.begin(115200); //Establishes connection with onchip bluetooth module and begins sampling pin A0 at 115200 baud
  EMA_result = analogRead(A0); // set EMA_result for t=1
  pinMode(5, OUTPUT);

//--------------------------------------------- Set PWM frequency for D5 ------------------------------

  TCCR0B = TCCR0B & B11111000 | B00000001;    // set timer 0 divisor to     1 for PWM frequency of 62500.00 Hz
//TCCR0B = TCCR0B & B11111000 | B00000010;    // set timer 0 divisor to     8 for PWM frequency of  7812.50 Hz
//TCCR0B = TCCR0B & B11111000 | B00000011;    // set timer 0 divisor to    64 for PWM frequency of   976.56 Hz (The DEFAULT)
//TCCR0B = TCCR0B & B11111000 | B00000100;    // set timer 0 divisor to   256 for PWM frequency of   244.14 Hz
//TCCR0B = TCCR0B & B11111000 | B00000101;    // set timer 0 divisor to  1024 for PWM frequency of    61.04 Hz


//--------------------------------------------- Set PWM frequency for D3 -----------------------------

//TCCR1B = TCCR1B & B11111000 | B00000001;    // set timer 1 divisor to     1 for PWM frequency of 31372.55 Hz
//TCCR1B = TCCR1B & B11111000 | B00000010;    // set timer 1 divisor to     8 for PWM frequency of  3921.16 Hz
  TCCR1B = TCCR1B & B11111000 | B00000011;    // set timer 1 divisor to    64 for PWM frequency of   490.20 Hz (The DEFAULT)
//TCCR1B = TCCR1B & B11111000 | B00000100;    // set timer 1 divisor to   256 for PWM frequency of   122.55 Hz
//TCCR1B = TCCR1B & B11111000 | B00000101;    // set timer 1 divisor to  1024 for PWM frequency of    30.64 Hz

}

void loop() {
  sensorValue = analogRead(A0);
  //Serial.print(300); //To set the lower limit on y-axis
  //Serial.print(",");
  //Serial.print(900); //To set the upper limit on y-axis
  //Serial.print(",");
  EMA_result = (EMA_alpha*sensorValue) + ((1-EMA_alpha)*EMA_result);

  Serial.print(EMA_result);  //To send the actual sensor data
  Serial.print("~");
  Serial.println(pwm);
  delay(4); //changes speed of scrolling x-axis

  if(clockDividerIndex == clockDivider){
    if(calibration){
      if(EMA_result > 600 && pwm > 0){                    // decrease pwm value until output is close to zero
           pwm = pwm - (20.0/pow(2.0, (double)binCalIndex));
           binCalIndex = (binCalIndex >= 4 ? binCalIndex : binCalIndex + 1);
      }
      else if(EMA_result < 500 && pwm < 255){            // increase pwm value until output is close to zero
           pwm = pwm + (20.0/pow(2.0, (double)binCalIndex));
           binCalIndex = (binCalIndex >= 4 ? binCalIndex + (double)clockDivider*0.002: binCalIndex + 1);
      }
      else if(EMA_result < 600 && EMA_result > 400 && binCalIndex > 5){
        calibration = false;
      }
      else{
        calibration = false;                              // set initial calibration flag to false
      }
    }

    if(calibrationBackup){                                // if input is close to saturating, zero the output
         if(EMA_result > 900 || EMA_result < 100){
              binCalIndex = 2;
              calibration = true;
         }
    }
    clockDividerIndex = 0;
  }
  else{
    clockDividerIndex++;
  }

  analogWrite(5, pwm);
}
```

Fig. 7. Sample Arduino code

3 Test Result

The emotion sensing system was tested with our user under different kinds of thought stimuli and sleep. The data presented in Fig. 9 differentiates arousal and a normal emotional state. The neutral stimulus, personal self-reflection, generates a skin conductance change of about 0.18 μS over a period of two seconds with a 3.5-s delay, whereas the emotional stimulus, excitement, in this case, produced a 0.75 μS increase in skin conductance over a period of three seconds. These results are consistent with typical EDA response times [7, 8].

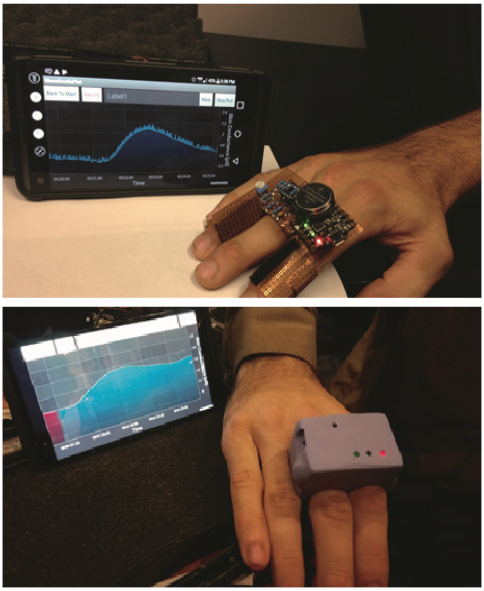


Fig. 8. Sensor data is recorded on a smartphone (a) 1st gen (b) 2nd gen robust version

A continuous signal recording was done on gaming and a sleep-pattern study. The recorded signals reveal arousal moments and REM eye movement pattern, which are useful in the study of human computer interaction and health applications. Figure 8(a) shows the 1st generation sensor in sync with a smartphone through a Bluetooth module. Figure 8(b) shows the 2nd generation, robust version sensor which can be worn in sleep. The sleeping data is shown in Fig. 10. In Fig. 10, the first 2.5 h contain some noticeable skin conductance responses (SCRs, also called EDA storms [9]), but they are relatively small (due to NREM) in comparison to the two peaks that occur at times 02:55:01.00 and 03:11:22:00 (due to REM). The graph shows a clear distinction between NREM and REM sleep if you consider only the spikes that change more than 2 μS in five minutes or less. This confirms previous research on EDA sleep patterns which found most people experience about 0.5 times their EDA level during NREM sleep compared to a wakeful state, and 1.5 times their EDA level during REM sleep compared to a wakeful state [9].

Fig. 9. Response to a neutral and emotional context stimulus

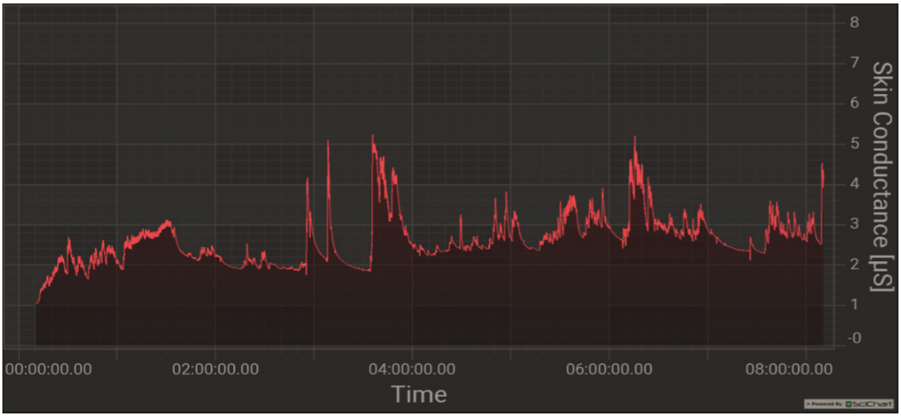


Fig. 10. 8-h sleeping data using the newer (2nd generation) robust sensor

In Fig. 11, the results of two gaming experiments are shown. In the first experiment, the user participated in a ‘Meditation Game’ (Top) Fig. 11(a), meant to help facilitate a relaxed state-of-mind, while at the same time maintaining a high level of concentration. Following a short tutorial, the user was prompted to participate in a controlled breathing exercise (10 s of breathing in followed by 10 s of slow exhalation) while focusing on only their breath. When the user became distracted, they were asked to mark the plot and make a conscious effort to focus back on their breathing. The results were a little unexpected but also made sense given the physiological implications of the unnatural breathing. As time progressed, the user's SCL slowly increased. The data show an elevated sympathetic response in the concentration game.

Fig. 11. (a) Gameplay recordings of meditation and (b) FPS games

For the second experiment, the user tested a first-person-shooter (FPS) style game, Fig. 11(b), for 15 min or the length of one online match. However, an unexpected curveball was thrown at the user 10 s into loading the match when the WiFi connection was interrupted, booting the user from the game. The sensor data indicated in the plot at time 00:06:15.00, shortly after getting kicked out from the match because of losing WiFi connection. There is a large spike of over 1 µS change, representing a real-time emotional response. The games ended with the user was killed eight times with labels 1 to 8 stated in the plot.

4 Future Direction

Currently, the authors are working on a next (3rd) generation sensor. It will use surface-mount, very low power consumption components, resulting in an eight times smaller size and a µW power conformal emotion sensor. Because of its compact size, low power, and flexible feature, the sensor can be operated stand-alone for a long time and can be easily incorporated in another sensing platform such as a smart textile to expand its sensing and application horizons.

5 Conclusion

In this paper, we have demonstrated a flexible, low power, compact, mobile emotion sensor which can be used to detect the emotional state of its users. We have fabricated a robust version (2nd gen) of the sensor which is all embedded in silicone for protection and can be worn comfortably during sleep to monitor the sleeping pattern. The next generation of the sensor will be capable of μW operation with an even smaller footprint.

Acknowledgment. The authors gratefully acknowledge the facility of UWB STEM EE lab equipment usage.

References

1. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: first steps towards an automatic system. In: ADS 2004: Affective Dialogue Systems, pp. 36–48 (2004)
2. Strauss, M., Reynolds, C., Hughes, S., Park, K., McDarby, G., Picard, R.W.: The HandWave bluetooth skin conductance sensor. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 699–706. Springer, Heidelberg (2005). https://doi.org/10.1007/11573548_90
3. Shimmer3 GSR+ Unit. Shimmer. http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor.GSR2, product on the website: http://thoughttechnology.com/index.php/gsr-2.html. Accessed 12 Feb 2017
4. NeuLog Sensors. http://www.neulog.com/gsr/. Accessed Dec 01
5. Published somewhere else
6. Nguyen, N.-V., Tu Nguyen, T.-K., Lee, H.-H.: A reduced switching loss PWM strategy to eliminate common-mode voltage in multilevel inverters. IEEE Trans. Power Electron. **30**, 5425–5438 (2015). ISSN 0885-8993
7. Kuhmann, W., Boucsein, W., Schaefer, F., Alexander, J.: Experimental investigation of psychophysiological stress-reactions induced by different system response times in human-computer interaction. Ergonomics **30** (1987)
8. Denes, G., Pizzamiglio, L. (eds.): Handbook of Clinical and Experimental Neuropsychology, p. 33. Psychology Press (1999). ISBN 9780863775420
9. Sano, A., Picard, R., Stickgold, R.: Quantitative analysis of wrist electrodermal activity during sleep. Int. J. Psychophysiol. **94**(3), 382–389 (2014)

Multimodal Paradigm for Emotion Recognition Based on EEG Signals

Naveen Masood[1](✉) and Humera Farooq[2]

[1] Electrical Engineering Department, Bahria University, 13 National Stadium Road, Karachi, Pakistan
navinza@hotmail.com

[2] Computer Science Department, Bahria University, 13 National Stadium Road, Karachi, Pakistan
humerafarooq.bukc@bahria.edu.pk

Abstract. A large number of existing real-time emotion recognition systems use stimuli with low ecological validity to elicit emotions. Furthermore, most of the emotion-based studies have used single stimuli to evoke emotions. In order to address these issues, this paper proposes multi modal emotion elicitation paradigm to investigate if same signatures of emotions exist if induced by using different methods of elicitation. The proposed method attempts to analyse Electroencephalography (EEG) data of healthy human subjects to recognize emotions using three different ways of elicitation. Emotional imagery is one of the proposed methods. Furthermore, audio-video clips and immersive videos on Virtual reality headsets are other two stimuli proposed for evoking emotions. Initially, experiments based on emotional imagery have been conducted with eight participants. The recorded EEG data is time sampled with 1 s time window and bandpass filtered in different frequency bands of Delta, Theta, Alpha, Beta, Low and High Gamma. Linear Discriminant Analysis (LDA) classified these features in two different classes of fear vs. neutral. Findings revealed that gamma band has achieved highest classification performance amongst other bands. This is in consistent with other studies related to emotion recognition. Outcome of the proposed work is a prototype of EEG based BCI system for emotion recognition using multi-modal elicitation methods. Initially, as mentioned before EEG data recorded based on emotional imagery was analysed. In the next phase of this study, experiments using audio-video clips and immersive videos will be conducted.

Keywords: Electroencephalography (EEG) · Emotions · Virtual reality
Emotional imagery · Classification

1 Introduction

Recognition of human's emotional states in real time plays an important role in machine emotional intelligence and human-machine interaction. The ability to identify a person's emotional state based on relatively easily acquired scalp electroencephalographic (EEG) data could be of clinical importance for anger management, depression, anxiety or stress

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 419–428, 2018.
https://doi.org/10.1007/978-3-319-91238-7_34

reduction, and for relating to persons with communication disabilities. Human emotion refers to a complex psychological state comprised of three components i.e. user experience, his physiological response along with behavioral or expressive reaction [1]. Different categories of emotions are disgust, pride, satisfaction, anger etc. [2]. Various studies have been conducted to find how the EEG signals correlate to human emotions. Emotiv has also been used for this purpose. Pham et al. have used Emotiv to capture EEG data while users are watching movies to induce emotions. Oscillatory brain rhythms with different frequency bands filtered from the recorded brain signals are used as input to different machine learning classifiers [3].

Human emotion can be described as any psychological mental state comprising of three constituents that are personal experience, person's physiological reaction along with his expressive or behavioral response [1]. In literature, we find different categories of human emotions that include disgust, anger, happiness, satisfaction, pride etc. [2]. There are various studies that have been performed to find out how the EEG based brain signals correlate to emotions. In one of the studies, Pham, Duy et al. recorded EEG data while the subjects are watching videos and movies. The displayed videos invoked different emotions to the participants while EEG headset is used to record brain signals. The recorded data is used for spectral analysis in different frequency bands. Classification is performed using different machine learning classifiers [3].

Currently, most of the real-time emotion recognition systems make use of stimuli having low ecological validity for example pictures, images, sound to invoke and recognize emotions. In few studies short video clips are shown to the subjects for elicitation of emotions. Furthermore, most of the emotion-based studies have used single method to elicit emotions [4–7]. Table 1 lists some studies mentioning the stimuli used for emotion elicitation. Furthermore, the type of emotion and the apparatus used to capture and analyse data are also mentioned.

Table 1. List of studies for emotion recognition

Study	Stimuli	Emotions tested	EEG headset
Aftanas et al. [10]	Video	Neutral, positive, negative	18-Channel EEG
Doronbekov et al. [11]	Video	Extremely fearful and aversive versus neutral	PET
Giuliani et al. [12]	Pictures	Amusement	Blood pressure, ECG, skin conductance, respiration, movement (piezoelectric captors), finger and ear pulse
Tsoi et al. [13]	Video	Humor	Psychological tests (ToM, PANAS, WCST, LSP, IQ)
Kothe [6]	Emotional Imagery	Fear, Pride, Happiness	BrainAmp

Most of the BCI applications and studies are conducted using medical grade EEG sensors with large number of electrodes [8, 9]. In order to enhance the portability and flexibility of these systems, various commercial low-cost EEG devices are available. Although comparative studies have reported their poorer fidelity compared with medical grade EEG recording systems like ANT and Neuroscan, but still they can provide valuable information on EEG data.

To overcome aforementioned issues and limitations, in this paper, multi modal emotion elicitation paradigm is proposed to investigate if same signatures of emotions exist if induced by using different methods of elicitation. This study is focused to record and analyze EEG data of human subjects for differences between target positive, negative and neutral emotions using different ways of stimuli for elicitation. Medical and commercial grade both EEG sensors will be used to analyze whether low cost devices could be helpful in understanding brain patterns associated with emotional states. In the present study, along with neutral brain state we attempt to explore the EEG correlates of joy and fear as positive and negative emotions respectively.

For EEG recording, following two sensors will be used:

i. BrainAmp (medical grade)
ii. Emotiv EPOC (low-cost commercial grade)

To address the issue of low ecological validity of stimuli in emotion recognition, this research focuses towards conducting the experiments based on three different ways of stimuli presentation. Explanation of each method is as follows:

1.1 Memory Recall/Emotional Imagery

As mentioned earlier, most of the EEG based emotion recognition researches have used external stimuli presentations to evoke emotions to the subjects for example images, sounds, video clips. In this method of memory recall/emotional imagery, we are primarily dealing with inwardly visualized, imagined or felt emotions evoked by the subject's own imagination or recall of emotionally loaded memories. The participants are requested to get involved or immerse themselves in prolonged, self-paced recall of emotion imagination usually with closed eyes.

In this work we make an attempt to discover and identify brain dynamic correlates of inwardly imagined emotional feelings by suggesting to healthy male and female participants, to remind or imagine such scenarios in which the subject has felt or would likely to feel a series of emotions under consideration.

1.2 Virtual Reality

During the last two decades, virtual reality (VR) has proved its significance for mainstream psychological research [14, 15]. A large number of studies have been conducted using VR sets for emotion recognition systems. This advance technology, having unique feature to simulate complicated but real scenarios and contexts, provides researchers exceptional and unprecedented options to analyze and investigate human emotions in pretty well controlled designs in the laboratory.

Having its nature virtual, with simulation of reality as much as possible, VR mainly relies on the suitable and adequate choice of specific perceptual hints and cues to evoke emotions. VR systems strongly provide safe and secure; motivating and controlled scenarios that help in improving BCI learning. Moreover, it provides not only options for better BCI experimental designs but also the analysis and investigation of brain and neural processes under consideration becomes comparatively convenient [16, 17]. In this research work, virtual reality based simulations and videos will be shown to the subjects to elicit emotions.

1.3 Audio-Visual Film Clips

In this scenario, the participants will view short clips/excerpts of audio-visual videos to elicit each category of emotions mentioned above.

2 Research Objective

The objectives of this work are:

1. To quantify with what accuracy the three emotional states of positive, negative and neutral can be differentiated from each other based on the classification on brain signals.
2. To examine whether specific pattern of brain activity exists for these emotional states.
3. By introducing different ways to elicit emotions, it is to find out if same signatures of such feelings appear.
4. Whether these patterns are to some extent common across individuals.
5. Which oscillatory processes in the brain are mainly modulated by such feelings and in what way.

3 Experimental Setup

In order to achieve its objectives, the proposed study is comprised of multiple phases that have to be completed systematically. First of all a briefing will be given to the participants regarding the purpose of the study and informed that their brain activity (EEG) will be recorded while having three different emotional states. Joy and fear are considered as positive and negative emotions respectively along with neutral state. Each participant is requested to fill a questionnaire. The first part of the questionnaire asks to write down any two events of their life/imagination that could induce feelings of joy and fear. Participants are free to choose any scenario, and may write as little or as much detail as they like.

The second part of the questionnaire asks whether the participant feels scared while watching horror movies and describe their experiences. The third part asks to mention any phobia/s e.g. the fear of height, reptiles etc. the participant is experiencing in daily life. In the last part approval from the participants is asked whether they allow the research team to show horror movies clips and videos related to their phobia.

Once the questionnaire is filled, only those participants will be shortlisted for the experiments that have given their consent to a large enough number of items eliciting negative emotions. The EEG brain activity of the participants will be recorded during the experiments based on three different stimuli on three different days. BrainCap will be used for EEG recording for first two days while Emotiv EPOC on third and last day of the experiments. Experiments are based on following three different stimuli:

3.1 Memory Recall/Emotional Imagery

To induce emotions, the first method is to recall memory of any joyful, frightening and neutral event/imagination.

In first phase of this research, initially experiments are performed with commercial grade Emotiv EPOC headset. The experiments are conducted at Bahria University, Karachi, Pakistan. Eight subjects participated in the study. Initially, recorded EEG data for self-induced emotions based on memory recall is analysed for classification performance. Total eight sessions of emotional imagery were conducted with each participant such that four out of eight sessions are for fear evoking memories while four based on neutral state. Complete details of the experiment are explained in our work [18]. Block diagram of the experiment is show in Fig. 1.

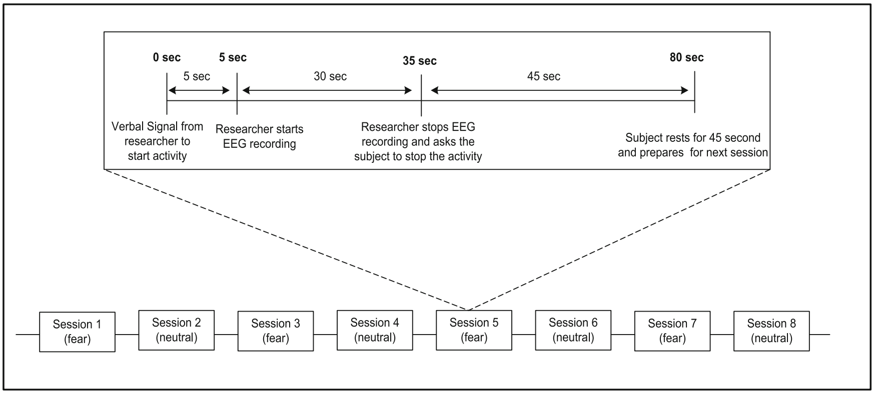


Fig. 1. Block diagram of the experiment [18]

The subject sitting on the chair will be asked to close the eyes and recall the joyful event they have mentioned in the questionnaire. The researcher signals the participant to start and stop the activity. Each session is comprised of 15 s. After the session, the participant is given a break of 30 s and asked to report the level of arousal (on a scale from 0 to 5) on the feedback form. In the same manner, sessions for fearful and neutral states are conducted.

The whole activity has to be repeated for four times on three different days.

3.2 Viewing of Immersive Videos on VR Set

The second method for elicitation of emotion is viewing of immersive videos on VR set. Each video is comprised of 1 to 5 min duration. After the session, the participants are given a break of 40 s and asked to report level of arousal (on a scale from 0 to 5) on the feedback form and specify the moments when in the video they felt specifically fear and joy. During each day, a total of ten videos will be shown to the subjects.

Selection of videos is based on two steps. Firstly, three students were selected from Bahria University who are frequent movie viewers. These students were requested to propose movies and videos that could be representative clips (at least two) for each emotion. In the next step, the recommended videos were shown to another group of eight undergraduate students and asked to report their rating for level of arousal and category of emotion elicited while watching it. Some interesting results were obtained during this activity. Immersive videos related to roller coaster ride or skydiving are categorized as joyful by three students while rest of them put these videos for fear category as these participants have phobia of height. Similarly, a video related to car driving simulation is categorized in three categories of joy, fear and neutral depending upon the personality trait of the viewer.

3.3 Viewing of Audio-Video Movie Clips

Instead of immersive videos, normal audio-visual film clips are shown to the users to elicit emotions. Each video is comprised of 1 to 5 min duration. After the session, the participants are given a break of 40 s and asked to report level of arousal (on a scale from 0 to 5) on the feedback form and specify the moments when in the video they felt fear and joy. During each day, total ten videos will be shown to the subjects.

4 Results and Discussion

As mentioned earlier, initially experiments based on scenario of emotional imagery are performed with commercial grade Emotiv EPOC headset. The experiments are conducted at Bahria University, Karachi, Pakistan. Eight subjects participated in the study. Initially, recorded EEG data for self-induced emotions based on memory recall is analysed for classification performance. Total eight sessions of emotional imagery were conducted with each participant such that four out of eight sessions are for fear evoking memories while four based on neutral state.

EEG data is time sampled with 1 s time window. Bandpass features are extracted for input to the Common Spatial Pattern (CSP) algorithm. The features then fed to the Linear Discriminant Analysis (LDA) classifier. EEG data is firstly band pass filtered in frequency range of 1–100 Hz with two pairs of spatial filters. Mention Butterworth order. Results of the classification accuracies are mentioned in Table 2. Here we observe that maximum accuracy of 96.3% and mean accuracy of 66.58% are achieved.

Table 2. Classification Accuracies in spectral band of 1–100 Hz for each subject

Subjects	Classification Accuracies in spectral band of 1-100 Hz
SUB-1	65.12
SUB-2	65.85
SUB-3	55.00
SUB-4	64.10
SUB-5	66.67
SUB-6	48.84
SUB-7	96.30
SUB-8	70.73
Average	66.58

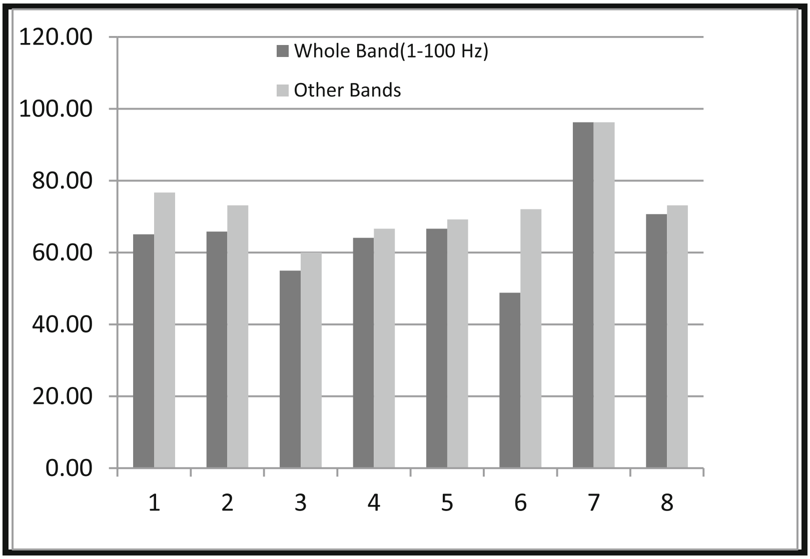


Fig. 2. Comparison between classification accuracies achieved for whole band of 1-100 Hz vs. highest accuracy for each subject in any of the spectral band other than the whole one

In order to explore the relevant spectral band for emotion recognition in case of emotional imagery, we performed bandpass filtering in different frequency bands in which the neurophysiological signals reside i.e. Delta band (1–3 Hz), Theta band (4–7 Hz), Alpha band (8–13 Hz), Beta band (14–30 Hz), Low Gamma band (31–50 Hz), High Gamma band (30–100 Hz) and the whole band of 1–100 Hz. So now we have seven filtered datasets for each of the eight subjects. Each dataset is separately provided as input to the LDA classifier to evaluate classification performance. Figure 2 plots the comparison between accuracy achieved for whole band of 1–100 Hz vs. highest accuracy for each subject in any of the spectral band other than the whole one. It is evident from the graph that for seven out of eight subjects have higher accuracy in other bands while for SUB-7, same accuracies are achieved for the two cases of spectral bands.

Now, we have attempted to find the most relevant frequency band for classification purpose. Figure 3 plots the comparison of mean classification accuracies achieved in each of the considered spectral bands of Delta, theta, alpha, beta, low and high gamma, and the whole band. Here the findings reveal that gamma band is able to achieve highest classification performance. This is consistent with the findings of studies related to emotion recognition. Li and Lu [19] have used image as stimuli for invoking emotions and they concluded that high frequency bands of EEG are highly relevant for classification. Dan Nie et al. also found that higher frequency bands provide more significant contribution to emotional response as compared to lower frequency ranges [20]. In our case, the high gamma band resulted in accuracy of 70% while low gamma is almost 68%.

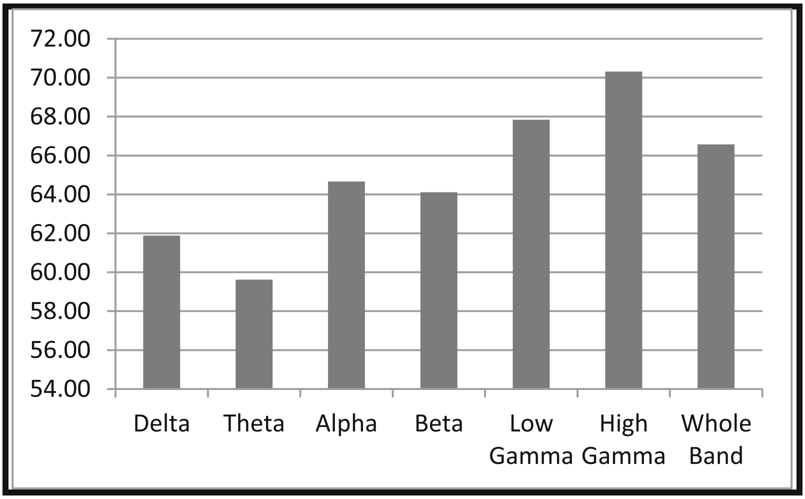


Fig. 3. Graph showing Mean accuracies achieved in each of the seven spectral band considered

5 Conclusion

In this work, we have proposed multi modal emotion elicitation paradigm for EEG based emotion recognition system. Most of the studies, use single way of invoking emotions but here in order to address the problem of low ecological validity of these studies, we proposed the paradigm with multi modal stimuli for invoking emotions. The outcome of the proposed work is a prototype of EEG based BCI system for emotion recognition. Initially, EEG data recorded based on self-induced emotions by recalling memories is analysed with Emotiv EPOC. Initial findings reveal that gamma band is able to produce highest accuracy. Furthermore, using Emotiv EPOC headset, these findings show promising result to use commercial grade EEG sensors for emotion recognition in case of emotional imagery. In the next phase of this study, experiments based on watching videos and immersive reality will be conducted.

Acknowledgments. This work is funded by Higher Education Commission (HEC), Pakistan and is being conducted and supervised under the 'Intelligent Systems and Robotics' research group at Computer Science (CS) Department, Bahria University, Karachi, Pakistan.

References

1. Hockenbury, D.H., Hockenbury, S.E.: Discovering Psychology. Macmillan (2010)
2. Ekman, P.E., Davidson, R.J.: The Nature of Emotion: Fundamental Questions. Oxford University Press (1994)
3. Pham, T.D., Tran, D.: Emotion Recognition using the Emotiv EPOC device. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012. LNCS, vol. 7667, pp. 394–399. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34500-5_47
4. Hu, X., et al.: EEG Correlates of ten positive emotions. Front. Hum. Neurosci. **11** (2017)
5. Zheng, W.-L., Lu, B.-L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
6. Kothe, C.A., Makeig, S., Onton, J.A.: Emotion recognition from EEG during self-paced emotional imagery. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII). IEEE (2013)
7. Daly, I., et al.: Neural correlates of emotional responses to music: an EEG study. Neurosci. Lett. **573**, 52–57 (2014)
8. Hettich, D.T., et al.: EEG responses to auditory stimuli for automatic affect recognition. Front. Neurosci. **10** (2016)
9. Larsen, E.A.: Classification of EEG signals in a brain-computer interface system. Institutt for datateknikk og informasjonsvitenskap (2011)
10. Aftanas, L.I., et al.: Non-linear dynamic complexity of the human EEG during evoked emotions. Int. J. Psychophysiol. **28**(1), 63–76 (1998)
11. Doronbekov, T.K., et al.: Neural basis of fear conditioning induced by video clip: positron emission tomography study. Psychiatry Clin. Neurosci. **59**(2), 155–162 (2005)
12. Giuliani, N.R., McRae, K., Gross, J.J.: The up-and down-regulation of amusement: experiential, behavioral, and autonomic consequences. Emotion **8**(5), 714 (2008)
13. Tsoi, D.-Y., et al.: Humour experience in schizophrenia: relationship with executive dysfunction and psychosocial impairment. Psychol. Med. **38**(6), 801–810 (2008)
14. Tyrrell, R., et al.: Simulator sickness in patients with neck pain and vestibular pathology during virtual reality tasks. In: Virtual Reality, pp. 1–9 (2017)
15. Yang, Y.D., et al.: Brain responses to biological motion predict treatment outcome in young adults with autism receiving virtual reality social cognition training: preliminary findings. Behav. Res. Ther. **93**, 55–66 (2017)
16. Slobounov, S.M., et al.: Modulation of cortical activity in 2D versus 3D virtual reality environments: an EEG study. Int. J. Psychophysiol. **95**(3), 254–260 (2015)
17. Simões, M., Amaral, C., Carvalho, P., Castelo-Branco, M.: Specific EEG/ERP responses to dynamic facial expressions in virtual reality environments. In: Zhang, Y.-T. (ed.) The International Conference on Health Informatics. IP, vol. 42, pp. 331–334. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-03005-0_84
18. Naveen Masood, H.F.: EEG Correlates of self-induced Human Emotions: Fear vs. Neutral. (Submitted) (2017)

19. Li, M., Lu, B.-L.: Emotion classification based on gamma-band EEG. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2009. IEEE (2009)
20. Nie, D., et al.: EEG-based emotion recognition during watching movies. In: 2011 5th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE (2011)

Automatic Low-Level Overlays on Presentations to Support Regaining an Audience's Attention

Walter Ritter[1(✉)], Guido Kempter[1], Isabella Hämmerle[1], and Andreas Wohlgenannt[2]

[1] Vorarlberg University of Applied Sciences, Hochschulstr. 1, 6850 Dornbirn, Austria
walter.ritter@fhv.at
[2] WolfVision GmbH, Oberes Ried 14, 6833 Klaus, Austria

Abstract. In a world full of distractions, keeping an audience focused on a presentation is getting increasingly difficult. In this paper, we propose a system that supports presenters in a nearly subliminal way to regain attention of the overall audience. The system uses a measure of motion complexity inside the audience area as an estimate for overall attention. It then applies low-level visual overlays over presentations if the estimated level of attention is getting too low. Ideally, these dynamically adapted visual overlays can be detected in the peripheral field of view but not in the foveal field of view. In a pilot study with 14 participants, we tested the feasibility of this approach with a simplified version of the system, limiting stimuli to red colored overlays up to an opacity of 20%. First results show that motion complexity can indeed be a good indicator of distractions and low-level visual overlays can lead to a higher perceived level of agitation. However, the visual effects used in this pilot study have been partly perceived by the audience. Further work is needed to identify visual stimuli that are best fitted for recapturing attention without irritating those already focused on the presentation.

Keywords: Presentation support · Subliminal · Adaptive systems
Visual overlays

1 Introduction

Remaining focused on presentations in today's world is getting increasingly difficult, given the countless distractions offered by smartphones and other personal electronic devices, in addition to the various external and ambient distractions. This imposes a big challenge on presenters to perfectly prepare their presentations for their audience.

In a study, Ward et al. (2017) report that even having phones in silent mode makes it harder to stay focused and reduces cognitive performance compared to having no phone at all. But also without smartphones the available attention span is limited. Reports on the typical attention span duration vary broadly. McKeachie and Svinicki (2013) mention a 10-min attention span, which they base on a study about note taking during lectures by Hartley and Davies (1978). Bunce et al. (2010) mentioned that students self-reported an attention decline already after 4–5 min into a presentation after the initial settling in period, and that they would also alternate between attention and nonattention

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 429–440, 2018.
https://doi.org/10.1007/978-3-319-91238-7_35

states in briefer and briefer cycles during a lecture. In general Wilson and Korn (2007) criticize the lack of consideration of individual differences in various attention span studies. A widely-reported study conducted by Microsoft, which described an attention span of just 8 s in 2013, has been disregarded in this context, as it basically referred to the time people typically spent on a web page before moving on, and not to actual attention span durations (Bradbury 2016).

Independent of the actual length of the attention span, it's not only a matter about attention itself, it's also about where members of the audience divert their attention to, or expressed differently, what might distract them. Pratto and Oliver (1991) describe negative social information as a powerful source to grab attention. Smith et al (2003) reported about a negativity bias in attention allocation, meaning negative stimuli can divert the attention away from the current focus more easily than positive ones.

As soon as some individuals of an audience get distracted, this state might be passed on to others subconsciously by mood transfers, which could possibly lead to an agitated audience (Tudor et al 2013). This process is caused by the same mechanisms as the ones underlying for example contagious yawning, which is described as nonverbal communication to synchronize behavior within a group (Guggisberg 2011). The mechanism of mood transfers has originally been discussed in birds' research, where one bird fleeing causes all others of the group to flee too at virtually the same time (Lorenz 1935).

There are, of course many ways presenters can prepare for these challenges in first place. Having the information perfectly selected and prepared according to the interests of the expected audience is an important first step. Obeying best practices for preparing presentations is another important step towards captivating the attention of an audience.

However, an interesting question arises. Could a presentation system possibly support the presenter in a subliminal way to regain the attention of an audience? What would be needed to distract the audience from the countless distractions back to the presentation? In this paper, we present a concept for such a system and describe a brief pilot study to test the feasibility of this approach.

Such a system should be able to judge the overall level of attention of an audience and to present stimuli in a way, that recaptures the interest of distracted people without distracting those people already focused on the presentation. In the next section, we look at some of the related work before we present our approach.

2 Related Work

Measuring the level of attention in classroom settings has been of interest to researchers since a long time. One approach is described by Helmke & Renkl in their Münchner Aufmerksamkeitsinventar (MAI). The basic idea is to observe students in a classroom during a lecture and sequentially code the attention state of individual students in fixed time intervals of 5 s. For this, attention states are classified in two layers. The first one discriminates between no-task, on-task, and off-task states. In the second layer the on-task state is further divided into compliance with task-requirements, self-initiated, or externally initiated activities. The off-task state is further divided into active (e.g. distracting others) or passive (e.g. dreaming) state (Helmke and Renkl 1992). Hommel presented an adjusted variant that's modified to better fit higher education and makes use of videography to analyze all students in 30 s intervals over the whole time (Hommel 2012). Both approaches also consider the lecture context in addition to attention state.

More advanced approaches make use of emotion recognition systems that analyze facial expressions (Magdin et al. 2016; Magdin and Prikler 2018; Heyjin 2017; Ayvaz and Gürüler 2017; Daouas and Lejmi 2018), eye movements (Krithika and Lakshmi 2016), voice characteristics, written text or a combination of different modalities (Nedji et al. 2008; Bahreini et al. 2016) to deduce the emotional state of a person. These systems are mostly used and tested in an e-learning context, where the knowledge of the emotional state should allow for better feedback by the remote lecturer. However, most of these systems are optimized for observations of individuals, but not suited as an overall measure for a whole audience. Furthermore, members of an audience (e.g. students in a lecture or in general listeners to presentations) would not feel comfortable if they would be tracked individually by cameras.

The human eye features different capabilities in the foveal and the peripheral field of view. In its periphery, the eye has a higher temporal resolution and sensitivity for light changes than in the foveal area. This higher sensitivity to light changes is caused by a higher rod density in the periphery (Curcio et al. 1990; Jonas et al. 1992; Wells-Grey et al. 2016). Rod mediated vision also seems to cause more activation in the brain area that is associated with motion, compared to cone mediated vision (Hadjikhani and Tootell 2000). The reason for the difference between cone- and rod-density could be originated in evolution: humans with this feature were able to detect dangerous situations outside of their current field of view faster and could react to the situation in time and thus survived.

These characteristics might also be exploitable by a system that targets people not directly looking at it.

Adaptive systems that react to some form of user input in an indirect/non-obtrusive way to better support users have been used in various fields. Ward et al. (2000) demonstrated a novel text input interface, which used previous input and learned language statistics to make it easier to select probable succeeding characters. This mechanism enables efficient text input even by only using an input modality with 1- or 2-dimensional input capabilities. The same principle is used for modern touch keyboards that automatically increase the sensitive area around certain keys that are more probable to be typed next than others, and therefore make them easier to reach. Kempter et al. (2003) discussed the benefits of analogue communication in the human computer interaction context, where a system's behavior can be influenced by analogue channels with no explicitly/pre-defined meanings. Ritter (2011) described different approaches where subliminal feedback loops in human computer interaction, based on these analogue communication patterns, could be used to support users in a transparent and unobtrusive way. In a widely-criticized study for its ethical aspects, Kramer et al. (2014) have shown that emotional states can be transferred over social networks. Users tend to take on the same emotions as conveyed in their news stream, even without the users being aware of this process. These studies demonstrate that systems can be used to influence people in subliminal ways.

While people might feel uncomfortable, when they hear they can be influenced in a subliminal way, the process of influencing people is one of the corner stones of advertising and is already happening all around the clock.

3 Concept and Implementation

Based on the insights of the work described in the previous section, we aimed to create a presentation system that observes the audience, deduces an estimate for overall attention and applies evolving visual stimuli to recapture the attention of people looking away from the presentation.

The system shown in Fig. 1 depicts the basic system stages. The input processing layer consists of an image analysis module that takes input from a webcam and calculates a motion complexity index. This index is then forwarded to the adaptation loop stage. The input processing layer also contains an environment module that monitors air conditions inside the room, like CO_2 level, temperature, and humidity. These measures are currently logged but not yet used as input for the adaptation loop. The adaptation loop module implements a simple genetic algorithm that uses the motion complexity input as fitness values for its candidate adaptations. Once fitness values are available for each candidate in the population, a new generation of adaptation candidates is generated using a roulette wheel algorithm for parent selection. Each candidate is then applied for a situation with exceeding motion complexity and again assigned a corresponding fitness value based on the motion complexity. In our current system, candidate solutions are encoded using a 10-bit string shown in Table 1.

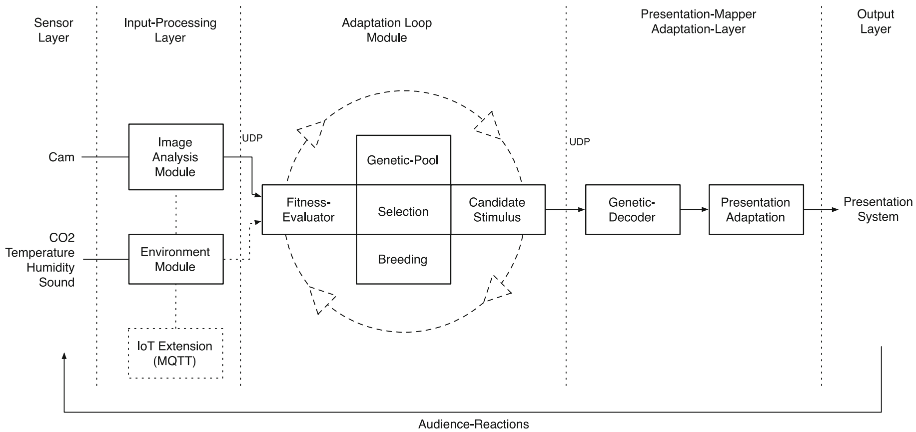


Fig. 1. System overview of an adaptive presentation system, consisting of an input processing stage, an adaptation loop module, a presentation mapping stage, and the output layer.

Table 1. Example genetic encoding of visual overlay parameters, each consisting of 2 bits

Parameter	Values
Opacity	0.0, 0.1, 0.15, 0.2
Transition time	0.025, 0.050, 0.100, 10
Fade-In curve	EaseInOut, Linear, EaseIn, EaseOut
Fade-Out curve	EaseInOut, Linear, EaseIn, EaseOut
Color	black, red, blue, white

New evaluation cycles are triggered every 60 s during situations with too much motion. The following 30 s are used as timeframe for measuring the reactions of the audience following a stimulus (see Fig. 2).

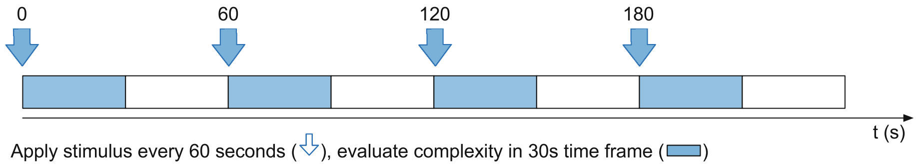


Fig. 2. Scheduled stimulus application and evaluation timeframes

The presentation mapper takes a candidate bit string as input and maps this to specific adaptation animations that are then overlaid over the presentation. The main idea of these adaptations is to be very low-level and fast, effectively producing a lighting change, and therefore being more noticeable in the peripheral field of view while hardly being visible for people looking directly at the presentation.

Motion complexity (MC) inside the audience area is derived via an image processing algorithm, taking a video stream from a webcam as input and performing grayscale conversion, background subtraction, thresholding, and morphing operators to the individual images (see Fig. 3) using OpenCV. This results in a black and white image with connected and independent motion areas.

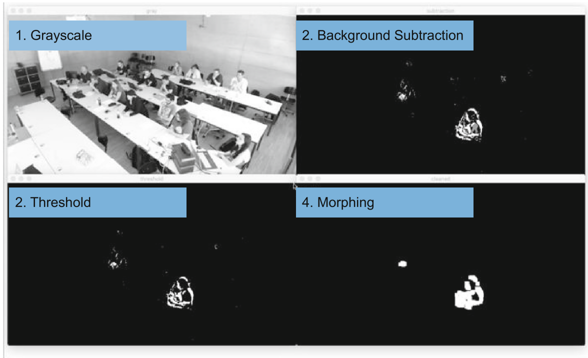


Fig. 3. Image processing steps for retrieving motion complexity

From this, two different indices are calculated (see Eqs. 1 and 2).

$$MC_{rel} = \text{image area with changes} / \text{total image area} \tag{1}$$

$$MC_{cnt} = \text{count of independent movement areas} \tag{2}$$

MC_{rel} should give an overview of the overall action taking place in the camera's field of view and can also be used to detect lighting changes inside the room (e.g. a percentage higher than 50%). MC_{cnt} reflects the number of independent disturbance areas and helps to counter the effect that moving objects closer to the camera have a bigger influence on $MC_{rel.}$

We set up a small pilot study to explore the usefulness of the two different motion complexity indices for estimating an audience's attention level, and to test the feasibility of this approach in a real-life setting.

4 Pilot Study

In a small pilot study, we used a simplified version of the system described in the previous section, to mainly investigate the plausibility of the motion complexity indices and to see if a background color overlay would cause a noticeable effect. Since the genetic algorithm adaptation loop would have added too much variability, we chose to apply a red background overlay of varying opacity with a maximum value of 20%, depending on the current motion complexity. Red was chosen because of the common signal-meaning of it. This overlay will also result in a lighting change depending on the opacity, which should mainly be recognizable from the visual periphery.

4.1 Method

The system was installed in a lecture room with the camera placed so that it was observing the audience but not covering the lecturer area. During the presentation, text-based and image-based slides were shown, both with and without dynamic overlays. Figure 4 shows the sequence of the lecture slides. The audience was not told in advance that there would be overlays. They were informed however, that they would have to fill out a brief questionnaire after some slides. Each slide was shown for exactly 60 s.

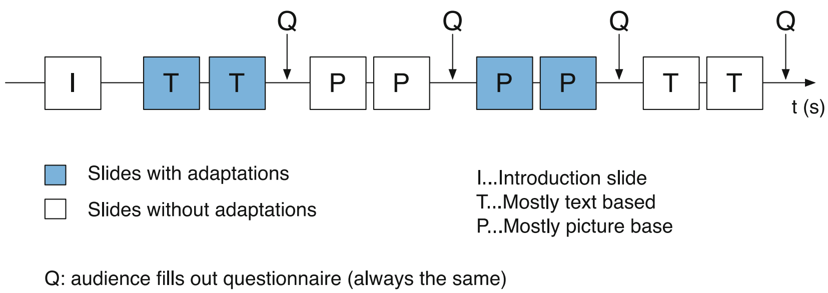


Fig. 4. Timeline of the test sequence. Duration of each slide is 60 s

In the pilot study 14 persons (6 female, 8 male) attended the test presentation. After each segment the participants had to rate the following properties on a 7-step rating scale:

Q1: "I followed the presentation…" from "exactly (1)" to "not at all (7)"
Q2: "At the moment I feel…" from "very calm (1)" to "very agitated (7)"
Q3: "The presentation content was…" from "very exciting (1)" to "very boring (7)"
Q4: "The slides were…" from "very intrusive (1)" to "very conservative (7)"

4.2 Results

Table 2 lists the descriptive statistics of the questionnaires. The text-based slides with ongoing adaptations were rated the most boring (M = 4.36, SD = 1.50). Text based slides without overlays were rated the most conservatives (M = 4.0, SD = 1.47). The reported

Table 2. Descriptive statistics of ratings

	Q1		Q2		Q3		Q4	
Stimulus	M	SD	M	SD	M	SD	M	SD
Adaptive text	3.00	2.00	3.71	1.82	4.36	1.50	3.64	1.15
Static pictures	1.93	0.92	2.71	1.54	3.29	1.49	3.71	1.14
Adaptive pictures	2.62	1.50	3.15	1.82	3.77	1.36	3.31	1.38
Static text	2.71	1.20	2.86	1.56	3.57	1.34	4.00	1.49

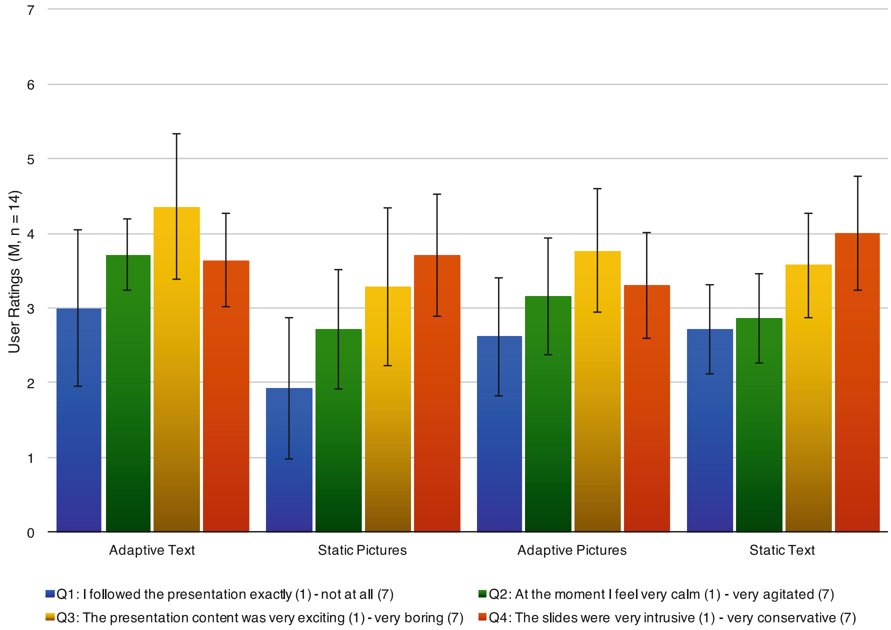


Fig. 5. Mean rating values on a scale from 1 to 7 of the different slide types (n = 14). Error bars show the 95% confidence interval.

attention (Q1) was rated highest for the static pictures (M = 1.93, SD = 0.92) and adaptive pictures (M = 2.62, SD = 1.50). Agitation was rated lowest for the two static slide types ($M_{pictures}$ = 2.71, $SD_{pictures}$ = 1.54, M_{text} = 2.86, SD_{text} = 1.56). See Fig. 5 for a graphical illustration.

T-tests were performed between the static and dynamic phases. The results show that there were significant differences between the reported agitation-level ($M_{adaptive}$ = 3.57, M_{static} = 2.79, p = .019) and the feeling-bored-level of the slides ($M_{adaptive}$ = 4.18, M_{static} = 3.43, p = .007). See Table 3 for more results and Fig. 6 for a graphical overview.

Table 3. Rating differences between adaptive and static slides

Question	$M_{adaptive}$	$SD_{adaptive}$	M_{static}	SD_{static}	$M_{paired\text{-}diff}$	p
Q1	2.96	1.90	2.32	1.12	0.64	.098
Q2*	**3.57**	**1.89**	**2.79**	**1.52**	**0.79**	**.019**
Q3*	**4.18**	**1.52**	**3.43**	**1.40**	**0.75**	**.007**
Q4	3.50	1.23	3.86	1.29	−0.36	.321

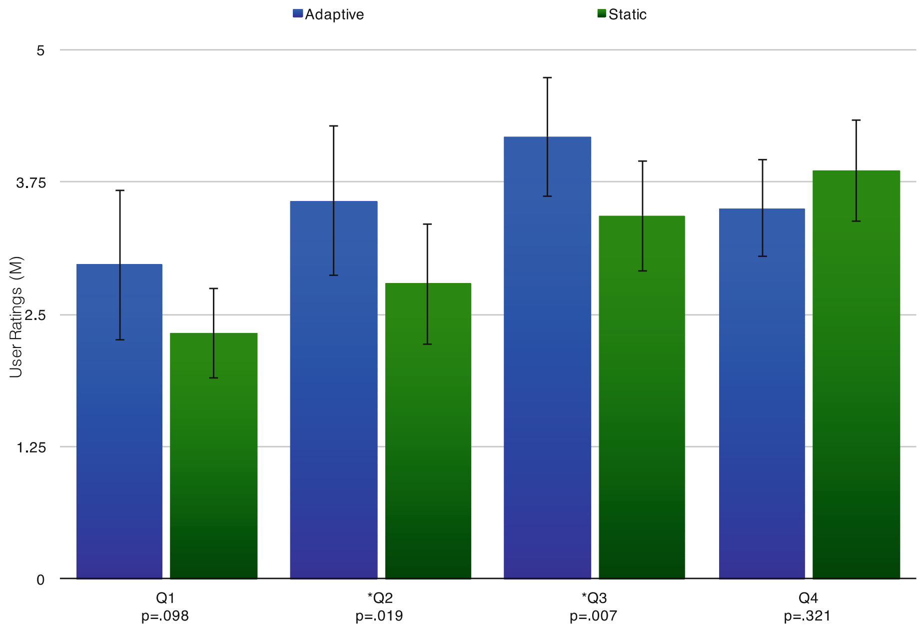


Fig. 6. Differences between adaptive and static slides. Significant differences could be found for agitation (Q2) and excitation (Q3). Error bars show the 95% confidence interval.

To also check the influence of the slide type (picture or text), a t-test has been performed between these groups. However, no significant differences could be identified.

Both motion complexity indices show a peak at the beginning of the lecture and then at the questionnaire slots 3, 6, and 9 (see Fig. 7), indicating that both measures reflect the settling in period of a presentation and react to distractions during the presentation (filling out the questionnaires). It's also interesting to note that the motion complexity levels for the picture slides (4,5 & 7,8) are above the ones for text slides (1,2 & 10,11). This could relate to the higher self-reported attention level of picture slides.

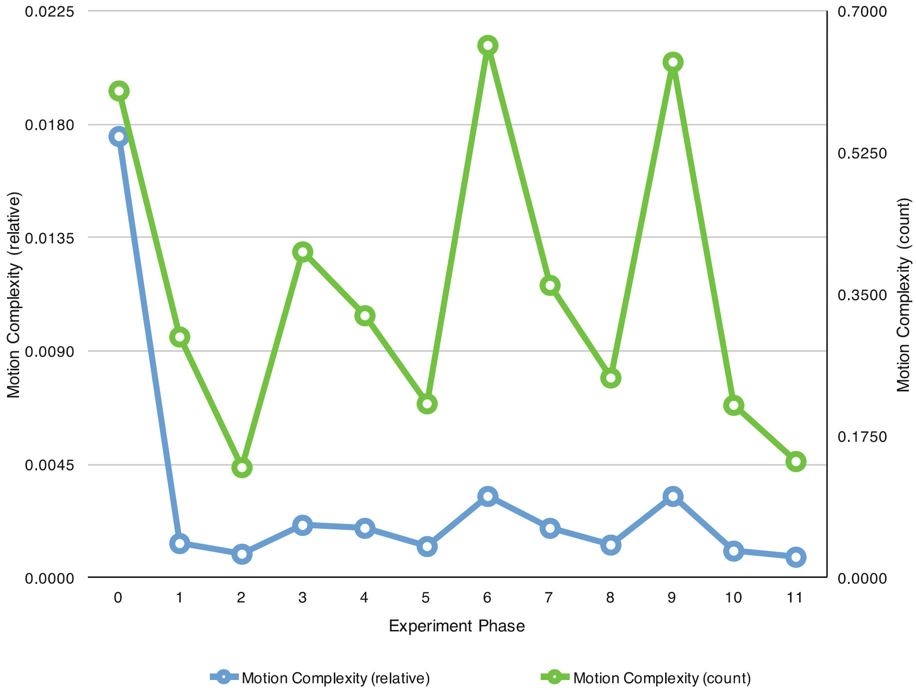


Fig. 7. Motion complexity indices through the course of the presentation. Peaks can be observed at the start, and at phases 3, 6, and 9 (questionnaire action).

5 Discussion and Conclusion

The results show that there was a significant difference between slides with and without adaptions taking place, hinting that we might indeed be able to influence the perception of slides using low-level visual overlays. However, the difficult part is to find the right balance between the noticeability from the peripheral field of view and the center field of view. In the mapping of the pilot study, the strongest visual overlays were too noticeable when directly looking at it – this was probably the reason for the relatively high self-reported agitation level of the slides with adaptions. Overall, the derived motion complexity for picture slides was higher than for text slides. This contrasts with the reported agitation level, where the slides with adaptations were rated higher independent of slide content type, but falls in line with the higher self-reported attention level for picture based slides. A decline from the first to the second slide within a condition could be observed for each variation, which was to be expected due to the settling in phase from the preceding questionnaire section.

One downside of our motion complexity indicators is the susceptibility to overreaction when massive motion occurs (e.g. a person leaving the room or is coming in late). While this is a little improved with MC_{cnt} compared to MC_{rel}, it is still present there too. One workaround could be to set a narrow threshold for valid motion complexity levels, and if this threshold is exceeded, pause the system until things get back to normal. This would also be advisable for filtering out abrupt light changes.

A possible concern is the influence of such a system on people with Epilepsy. Even if our overlays are within the same specs as e.g. movies or games that could also play in the peripheral field of view, further clarifications with experts in this field must be done, before using such a system in real-life settings.

In a next step, different stimuli will be tested in a lab setting regarding their potential to grab peoples' attention while they are focused on a different task, where the stimulus screen is placed outside the foveal field of view. The most promising properties will then be encoded for use with the genetic algorithm. Also, the motion complexity parameters could be compared to attention level ratings using e.g. the ModAI instrument. Once attention levels are rated for specific lectures, the motion complexity parameter algorithms can be fine-tuned using video recordings of the rated lectures.

After this, further tests of the overall system with more participants will have to be done.

Acknowledgements. Financial support for this project was provided by the Austrian research funding association (FFG) under the scope of the COMET program within the research project "Easy to use professional business and system control applications (LiTech)" (contract # 843535). This program is promoted by BMVIT, BMWFJ and the federal state of Vorarlberg.

References

Ayvaz, U., Gürüler, H.: Real-time detection of students' emotional states in the classroom. In: 25th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. IEEE, Antalya (2017)

Bahreini, K., Nadolski, R., Westera, W.: Towards multimodal emotion recognition in e-learning environments. Interact. Learn. Environ. **24**(3) (2016)

Bradbury, N.A.: Attention span during lectures: 8 seconds, 10 minutes, or more? Adv. Physiol. Educ. **40**, 509–513 (2016)

Bunce, D.M., Flens, E.A., Neiles, K.Y.: How long can students pay attention in class? a study of student attention decline using clickers. J. Chem. Educ. **87**(12), 1438–1443 (2010)

Curcio, C.A., Sloan, K.R., Kalina, R.E., Hendrickson, A.E.: Human photoreceptor topography. J. Comp. Neurol. **292**, 497–523 (1990)

Daouas, T., Lejmi, H.: Emotions recognition in an intelligent elearning environment. Interact. Learn. Environ. (2018). https://doi.org/10.1080/10494820.2018.1427114

Guggisberg, A.G., Mathis, J., Schnider, A., Hess, C.W.: Why do we yawn? the importance of evidence for specific yawn-induced effects. Neurosci. Biobehav. Rev. **35**(5), 1302–1304 (2011)

Hadjikhani, N., Tootell, R.B.: Projection of rods and cones within human visual cortex. Hum. Brain Mapp. **9**(1), 55–63 (2000)

Hartley, J., Davies, I.K.: Note taking: a critical review. Program. Learn. Educ. Technol. **15**, 207–224 (1978)

Helmke, A., Renkl, A.: Das Muenchener Aufmerksamkeitsinventar (MAI): Ein Instrument zur systematischen Verhaltensbeobachtung der Schueleraufmerksamkeit im Unterricht. Diagnostica **38**(2), 130–141 (1992)

Hommel, M.: Kodierhandbuch des Beobachtungsinventars zur systematischen und videobasierten Erfassung der Aufmerksamkeit von Lernenden (m/w): modifiziertes Aufmerksamkeitsinventar (ModAI). In: Dresdner Beiträge zur Wirtschaftspädagogik. - Dresden : Technische Universität, vol. 2012, p. 1 (2012). ISSN 0945-4845, ZDB-ID 21890250

Hyejin, K.: Learner's intelligent emotion detection system in U-learning environment. Int. J. u- and e-Serv. Sci. Technol. **10**(8), 91–98 (2017)

Jonas, J.B., Schneider, U., Naumann, G.O.: Count and density of human retinal photoreceptors. Graefes Arch. Clin. Exp. Ophthalmol. **230**(6), 505–510 (1992)

Kempter, G., Weidmann, K.H., Roux, P.: What are the benefits of analogous communication in human computer interaction? In: Stephanidis, C. (ed.) Universal Access in HCI: Inclusive Design in the Information Society, pp. 1427–1431. Lawrence Erlbaum Associates, Mahwah (2003)

Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Emotional contagion through social networks. Proc. Nat. Acad. Sci. **111** (24), 8788–8790 (2014)

Krithika, L.B., Lakshmi, P.G.G.: Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. Proc. Comput. Sci. **85**, 767–776 (2016)

Lorenz, K.: Der Kumpan in der Umwelt des Vogels – Der Artgenosse als auslösendes Moment sozialer Verhaltensweisen. J. für Ornithologie **83**, 137–213 (1935)

Magdin, M., Turčáni, M., Hudec, L.: Evaluating the Emotional State of a User Using a Webcam. Int. J. Interact. Multimed. Artif. Intell. **4**(1), 61–68 (2016)

Magdin, M., Prikler, F.: Real time facial expression recognition using Webcam and SDK affectiva. Int. J. Multimed. Artif. Intell. (2018, in Press)

McKeachie, W.J., Svinicki, M.: McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers, 14th edn. Wadsworth Publishing, Belmont (2013)

Nedji Milat, I., Seridi, H., Sellami, M.: Towards an intelligent emotional detection in an e-learning environment. In: Woolf, Beverley P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 712–714. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69132-7_86

Pratto, F., Oliver, J.: Automatic vigilance: the attention-grabbing power of negative social information. J. Pers. Soc. Psychol. **61**(3), 380–391 (1991)

Ritter, W.: Benefits of subliminal feedback loops in human-computer interaction. Adv. Hum.-Comput. Interact. **2011**, Article ID 346492 (2011)

Smith, N.K., Cacioppo, J.T., Larsen, J.T., Chartrand, T.L.: May I have your attention, please: electrocortical responses to positive and negative stimuli. Neuropsychologia **41**(2), 171–183 (2003)

Tudor, A.D., Poeschl, S., Doering, N.: What do audiences do when they sit and listen? Stud. Health Technol. Inform. **191**, 120–124 (2013)

Ward, A.F., Duke, K., Gneezy, A., Bos, M.W.: Brain drain: the mere presence of one's own smartphone reduces available cognitive capacity. J. Assoc. Consum. Res. **2**(4), 140–154 (2017)

Ward, D.J., Blackwell, A.F., MacKay, D.J.C.: Dasher—a data entry interface using continuous gestures and language models. In: Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology (UIST 2000), pp. 129–137. ACM, New York (2000)

Wells-Gray, E.M., Choi, S.S., Bries, A., Doble, N.: Variation in rod and cone density from the fovea to the mid-periphery in healthy human retinas using adaptive optics scanning laser ophthalmoscopy. Eye **30**, 1135–1143 (2016)

Wilson, K., Korn, J.K.: Attention during lectures: beyond ten minutes. Teach. Psychol. **34**(2), 85–89 (2007)

Experimental Induction and Measurement of Negative Affect Induced by Interacting with In-vehicle Information Systems

Nadja Schömig[1(✉)], Frederik Naujoks[1], Thomas Hammer[1], Markus Tomzig[1], Bastian Hinterleitner[2], and Stefan Mayer[2]

[1] Würzburger Institut für Verkehrswissenschaften GmbH, 97209 Veitshöchheim, Germany
schoemig@wivw.de
[2] Audi Electronics Venture GmbH, 85080 Gaimersheim, Germany

Abstract. The goal of this study was to investigate whether it is possible to induce negative affects in the interaction with IVIS in a driving simulator environment and whether this would be reflected in the driver's facial expressions. N = 29 participants completed a 30-min-drive in a high-fidelity driving simulator performing several IVIS tasks using both a visual multi-level menu and a simulated speech recognition navigation system. During the drive, negative affects were induced by increasing the complexity of the menu tasks and decreasing the recognition rate of the speech system. The results show that experiencing difficulties in solving IVIS tasks went along with self-reported negative affect such as frustration and anger. A newly developed observation protocol adapted from the Facial Action Coding System (FACS) used for the ratings of negative emotions based on the video footage of the drivers' face during the simulator trials was established. The method revealed significant correlations with the self-reported measures of negative emotions. However, there were no correlations between the observed facial expressions and the sub-scales of the State-Trait-Anger-Inventory (STAXI). The study shows the importance of a well-designed IVIS in order to prevent negative emotions that might result in non-acceptance of new technologies. The study also shows the potential of facial recognition technology to provide assistance or tutoring functions that could relieve the driver from emotional discomfort.

Keywords: In-vehicle information system · Driving simulation
Anger · Emotion · Facial recognition · HCI design

1 Introduction

In-vehicle information systems (IVIS) are available in virtually every car segment from luxury to low price vehicles. They are used for a wide variety of in-vehicle tasks such as adjusting vehicle settings, using navigation systems or in-vehicle entertainment. Providing customers with a positive and pleasurable experience while interaction with IVIS should thus be a primary concern of HCI research. However, increased complexity of IVIS might also cause negative affect such as frustration or anger, especially

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 441–452, 2018.
https://doi.org/10.1007/978-3-319-91238-7_36

when drivers are not familiar with the respective IVIS. Recognizing these negative emotions could be used both to guide the design process (e.g., by re-design of interaction sequences that commonly cause negative affect) or to provide on-line tutoring systems that provide assistance in difficult situations.

2 Theoretical Background

In the driving context, driver's emotions are often studied as reaction to certain environmental conditions, such as frustrating situations like traffic jams or as a reaction to the behavior of other road users [1, 2]. Within this context, the intention of investigating driver's negative emotional state lies in the reduction of aggressive driving and therefore crash rates [3].

However, there are only very few studies to date that have tried to induce and measure negative affect by interaction with IVIS. In most cases, a faulty ASR system (Automatic speech recognition system) was investigated. For example, Bachfischer et al. [4] found that users become annoyed and stressed when interacting with a faulty ASR. Malta et al. [5] also tried to measure the frustration level of drivers during interaction with a faulty ASR. The drivers rated their frustration level by watching a video of their face which was recorded during the interaction with the system. For the rating a slider was used, that could be set from neutral to extreme. For predicting the frustration level measured by the participant's responses through his/her physiological state, overall facial expression, and pedal actuation, Bayesian networks were used. Kuhn et al. [6] assessed frustration induced by different recognition rates of an ASR (44% recognition rate vs. 89%) on driving performance. They found a higher steering wheel angle variance with the low recognition rate, which points out that negative affect can indeed influence driving behavior.

According to [7] for the driving context the best measure is to assess driver's emotions by analysing facial expression. Other measures (e.g. questionnaires or EEG) would disturb the driver and would have only a low specificity to differentiate between different emotions.

Eyben et al. [8] summarize different literature with regard to emotions in driving. Among other sources of emotions, they also discuss the effects of IVIS interactions. As a meaningful strategy to reduce these negative emotions they propose a "socially competent human-like car" including an IVIS interaction which behaves like a real human codriver, that would give personalized help and quickly offer alternative solutions. However, the realization of such an approach would require the measurement of drivers' negative affect in the first place, which is exactly what we did in this study.

3 Method

N = 29 participants completed IVIS-tasks of varying difficulty in a high-fidelity driving simulator. A real-world IVIS was transferred to the simulation environment. The aim of the study was to investigate whether drivers would report negative affect when solving difficult IVIS tasks and whether this would be reflected in the drivers' facial expressions.

Therefore, self-report scales as well as a standardized observation protocol were developed. The observation protocol was used to rate the facial expressions during interactions with the IVIS and was adapted from the Facial Action Coding System (FACS, Ekman et al. [9]). To account for individual differences in the expression of negative affect, the participants completed the State-Trait Anger Expression Inventory (STAXI, Schwenkmezger et al. [10]). During the study, drivers had to complete the IVIS task while following a simulated car. A feedback of whether the task was solved correctly or not was given after each IVIS task.

3.1 Setting

The study was conducted at the Wuerzburg Institute for Traffic Sciences (WIVW). A static driving simulator with an Opel Insignia mockup was used (Fig. 1 left). The front of the car was surrounded by a screen illuminated by three projectors providing a 270° field of view. In addition, displays were placed in the back of the car and on the side mirrors to allow the monitoring of surrounding traffic. The driving simulation software SILAB was used.

During the test the experimenter was sitting in an adjacent room separated from the simulator by a window. From that room he was able to monitor the test and see the participant through various cameras being placed in the car. He was also able to communicate with the participant using an intercom system.

A Basler-camera with 50 Hz and a solution of 2.3 MP placed in front of the driver recorded the driver's face and his/her interaction with the IVIS in a multi-screen video.

Fig. 1. Static driving simulator at the Wuerzburg Institute for traffic sciences and the car follow task (described in Sect. 3.2)

3.2 Driving Task

The driving task consisted of a variation of the car-follow-task proposed by the NHTSA distraction guidelines [11]. The participant was driving on a mostly straight road with two lanes. The task was to follow a car in front of the participant's car and keep a defined distance as constantly as possible. The car ahead changed its speed in a sinus-wave profile with slightly varying amplitude so that the driver had to permanently adjust the speed (Fig. 1 right).

To put the driver in the situation of having to maintain a constant headway, a colored enhanced reality strip (ERS) was displayed on the road. Depending on the distance the color of the ERS changed. It appeared yellow when the distance was too large and blue when the distance was too small. Under perfect circumstances the bar was shown in gray (within a range between 1.0 and 1.8 s time distance). Besides the changing color of the bar, driving in the wrong distance had no additional effect.

3.3 Interface

For the interaction with the IVIS a simplified version of the current AUDI MMI GUI was simulated. It consisted of a hierarchical menu with different lists showing the different sub-menus available in the system. The driver had to navigate to different end points of the menu. The final selection of the end points did not trigger any actual reaction but resulted in a feedback depending on the task whether the selection was correct.

An AUDI MMI interaction unit was placed on top of the mockup's original console. This unit consisted of a rotary interaction knob and shortcuts for different menu levels (Fig. 2).

Fig. 2. AUDI MMI GUI (left) and interaction unit (right)

3.4 IVIS Tasks

During the 30 min drive the participant had to perform a total of 24 IVIS tasks. 20 of them were visual-manual interactions using the rotary interaction knob to navigate to certain entries in the menu system, the remaining 4 required interactions with a speech-based navigation system. When completing the menu tasks, drivers had to search for certain menu entries and to confirm their selection. The tasks were instructed verbally e.g. "please adjust intensity of the heater" at certain defined positions along the route. A computer-generated voice gave the instructions at the defined road section.

The tasks varied in their difficulty to solve. Half of them were very easy to solve while the other tasks were manipulated so that it was very difficult to reach the correct menu entry. This was either reached by a very ambiguous instruction so that it was not clear in which category the menu entry could be found and/or by a very short time buffer available to reach the menu entry before the task was aborted. Some of the tasks were even unsolvable as the instructed menu items were not part of the menu list. The task difficulty was determined in a previous pilot study. During the drive the frequency of tasks with higher complexity increased gradually to induce negative emotions.

The drivers were instructed that by solving the tasks correctly they could gain bonus points. In case of a correct answer, drivers received five bonus points and the feedback “answer was correct” displayed at the lower part of the visual interface. An extra monetary award was promised for that driver who reached a certain target score of 50 points. In case of an incorrect answer the participant got the feedback “task incorrect” in combination with a loud and unpleasant sound. If the driver was not able to reach the entry within a certain time buffer the task was aborted and the same sound was presented.

In addition, four other tasks requiring verbal interactions with the speech system were arranged in between the other menu tasks. They consisted of computer-generated speech outputs of a fictive “navigation system” which informed the driver about an upcoming traffic jam on the road. The driver was asked to either change the original route or to maintain the current route. The driver was instructed to avoid any traffic jam and should therefore answer with “yes” or “no” to the various questions in order to maintain the original route. Two of these tasks were manipulated in the way that the speech recognition system pretended to misinterpret the answer of the driver resulting in an incorrect feedback although it was correct. In one of the speech recognition tasks the answer had to be repeated several times before the system understood it correctly.

In case of wrong answers to the navigation systems 20 points were deducted from driver’s credits. Shortly before the target score of 50 was reached the manipulated navigation task was triggered which resulted in an incorrect feedback and a deduction of the score which should further increase driver’s negative emotions.

Table 1 summarizes the task sequence and the online ratings during the drive.

29 participants took part in the study. To achieve optimal availability of camera data only persons with a minimal height of 1.64 m were invited. 19 men and 10 women participated with an age of 24 to 65 years (m = 42, sd = 12.8 years).

Table 1. Used tasks and time points for online self-reportings of emotional state. Menu tasks are counted with figures, speech interactions with the navigation system with letters.

Task	Task type	complexity	block
t_0	**Online rating of emotional state (Baseline)**		
1	Menu task	easy	
2	Menu task	easy	
3	Menu task	easy	I
4	Menu task	easy	
A	Speech Interaction	correct feedback	
t_1	**Online rating of emotional state**		
5	Menu task	easy	
6	Menu task	easy	
7	Menu task	easy	
8	Menu task	ambiguous	II
9	Menu task	ambiguous+ short	
10	Menu task	not solvable	
t_2	**Online rating of emotional state**		
11	Menu task	easy	III
B	Speech Interaction	wrong feedback	
12	Menu task	easy	IV
13	Menu task	easy	
t_4	**Online rating of emotional state**		
C	Speech Interaction	correct feedback	
14	Menu task	not solvable	
15	Menu task	complex	
16	Menu task	complex	
D	Speech Interaction	3x repeat task	V
17	Menu task	ambiguous	
18	Menu task	complex	
19	Menu task	ambiguous+ short	
20	Menu task	complex	
t_5	**Online rating of emotional state**		

3.5 Measures

IVIS Task Performance

Driver performance in the IVIS task was coded as dichotomous variable for all of the 24 tasks. The task could either be performed successfully or not. The tasks were summarized into several blocks with each block comprising the tasks between two time points where drivers rated their current emotional state online. As parameter for task performance the relative frequency of successfully solved tasks per block was used. Dependent on the design of the tasks and the study, this frequency per se went down during the course of the drive (as the number of not solvable tasks increased).

Self-reported Affects

During the experiment, drivers were repeatedly asked to online rate their current emotions on different dimensions (how joyful, relaxed, angry, frustrated, helpless, irritated are you at the moment?). A 16-point-rating scale with verbal categories (according to Heller [12]) was used. The ratings had to be given after certain blocks of tasks (including an unequal number of tasks, but summarized according to their intention to induce negative affects).

A pre-rating before the drive served as baseline for the later ratings. Altogether, the online ratings were collected at six time points.

Observation of Driver's Expression

Drivers' facial expressions were coded by a rater who was blind for the content and the aim of the study. The basis for the coding was the video footage of the participants which showed the driver's face, the road and interaction with the menu system (see Fig. 3). The sound was muted so that verbal expressions could not be heard. From the video footage, the sections containing IVIS tasks were selected, resulting in 24 sequences per driver that had to be rated. The sequences started with the verbal instructions of the tasks and ended 10 s after the final feedback of the IVIS system had been given.

Fig. 3. Video footage used for observer rating

The raters were first trained by a specifically developed training procedure which condensed the principles and instructions of the widely-used FACS (Facial Acting Coding System by Ekman et al. [9] which is a tool for measuring facial expressions. It is an anatomical system for describing all observable facial movements. Each of these movements is labeled as an Action Unit (AU). The combination of AUs then defines a certain emotion (typically the six basic emotions). For the intentions of this study, it would have been too time intensive to use the original material and training. Therefore only the relevant AU combinations for the emotion anger were extracted, extended by other indicators for frustration or irritation (not included in FACS) and explained to the raters in a training session. Afterwards, the raters scored each single facial expression within each video sequence on a specially developed observation protocol. The rating was given on a scale from 0 (no emotion present) to 5 (maximum degree of any kind of negative emotion; positive emotions such as amusement were not coded). As independent measure the expression with the highest value per video sequence was used.

In order to achieve a high reliability and validity of the measure, an interrater-reliability score was calculated in advance from the ratings of three independent persons assessing the video sequences of twelve drivers. It turned out, that it was very difficult to differentiate between specific negative emotions (e.g. anger and irritation) leading to the decision to summarize these emotions to a more global rating in the analysis. The calculation of weighted kappa [13] for each pair of the three raters revealed a medium to high interrater-reliability (.673/.487/.556). The final rating was given by one rater who rated all the remaining video sequences of the 29 drivers.

State-Trait-Anger-Inventory

In order to evaluate whether the degree of self-reported negative emotions and the facial expressions correlate with a certain anger-personality the so called State-Trait-Anger Inventory was used (STAXI, Schwenkmezger et al. [10]). It is a measure for state- and trait-anger with three different forms of anger expression. People differ in the extent to which they express anger overtly and directly (anger out), how often anger feelings are hold in or suppressed (anger in) or in the extent to which they are able to control their angry feelings or their overtly expressed anger (anger control).

3.6 Study Procedure

First, the participants had to sign a secrecy agreement and were then presented with general information about the study. All participants were familiar and trained in using a driving simulator. The participants were told that the goal of the study was to evaluate the structure and content of the menu system and how difficult the search for certain menu entries would be. Therefore the tasks would be somewhat difficult but always possible to solve. They were promised a monetary reward if a target score was reached. They were left blind about the real intent of the study.

To become familiar with the system the participants were given six tasks. Each task led to a different section of the menu, so that every section had been seen before the actual test started.

Before the start of the drive, the participants rated their current emotional state on the 6 items using the 16-point-scale. During the 30 min test drive each participant had to perform the 24 IVIS tasks at defined positions on the route. After each of the defined five task blocks the participants again rated their emotional state online. After the drive the participants filled in several questionnaires, including the STAXI. At the end the real intent of the study was resolved. Before participants received the monetary reward (including the promised bonus in case of reaching the target score) a “cool down phase” assured that they could recover from a potentially high arousal level induced by the study. The whole procedure took about 60 min.

4 Results

4.1 Task Performance

According to expectations, task performance decreased during the test drive (see Table 2). The comparison between the five measuring points revealed a highly significant effect $F(4, 24) = 34.30$, $p < .001$.

4.2 Self-reported Emotional State

The self-reported emotional state changed gradually towards a more negative direction (see Fig. 4). The emotions anger and frustration heavily increased beginning with the second task block (which included several complex tasks) and then maintained on a constant, moderate level (not higher than 7 on the scale up to 15). However, there was

Table 2. Means and standard deviation of the dependent variables.

Measuring point	Task performance		Self-reported anger		Negative facial expression	
	M	SD	M	SD	M	SD
t_0	–	–	1.29	2.45	–	–
t_1	0.75	0.20	3.04	3.92	1.26	0.76
t_2	0.46	0.13	6.32	4.74	1.60	0.67
t_3	0.48	0.09	5.71	4.88	1.12	0.85
t_4	0.50	0.44	5.71	5.21	1.74	0.82
t_5	0.28	0.13	6.68	5.49	1.74	0.64

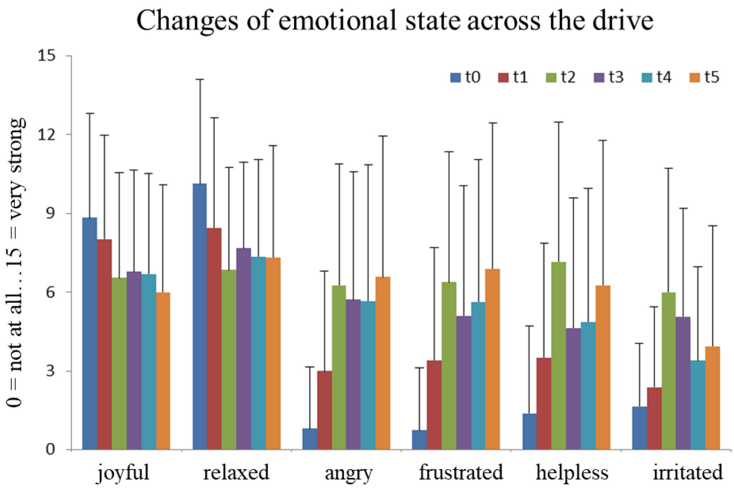


Fig. 4. Self-reported emotional state on six dimensions measured online at six time-points (means and standard deviations are displayed).

also a high variation between the subjects. The self-reported helplessness and irritation also increased beginning with the second block of tasks, however, decreased again towards the end of the drive indicating a kind of resignation if it turned out that the system's reactions are somewhat independent from the own task performance.

When looking further into detail for the anger ratings, it gets obvious that about one third of the sample (n = 10 drivers) subjectively perceived a strong or very strong feeling of anger (see Fig. 5). The peak lies here at the very end of the drive after block 5. However, there is also a remarkably large group of participants who expressed to experience very little or no anger at all. According to drivers' comments some of them realized the real intention of the study and therefore did not seem to be influenced by the tasks' manipulations. This result shows the difficult underpinning of inducing a certain emotional state in such an artificial experimental setup.

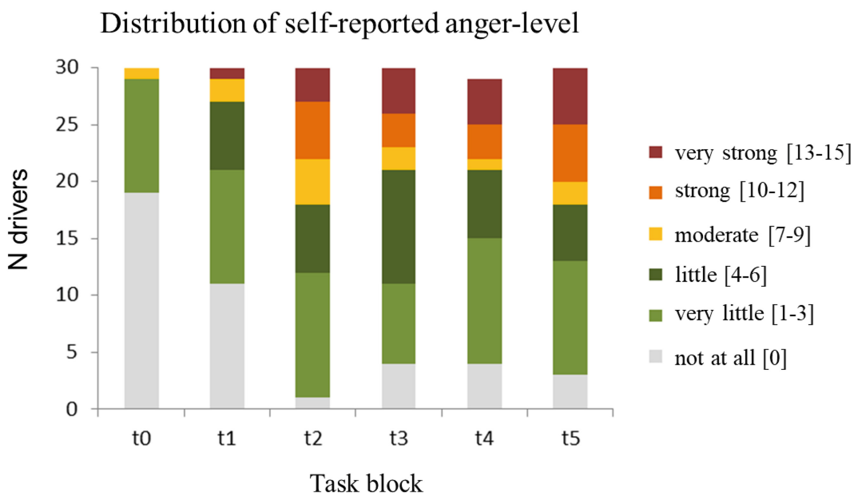


Fig. 5. Distribution of self-reported anger-level on the 16-point rating scale.

Table 3. Correlations between the different measures.

		Self-reported anger	Negative facial expression
Task performance		$r(142) = -.40$	$r(688) = -.28$
Negative facial expression		$r(142) = .37$	–
STAXI	Trait-Anger	$r(27) = .45$	$r(27) = .05$
		$p = .007$	$p = .398$
	Anger-Out	$r(27) = .26$	$r(27) = .15$
		$p = .085$	$p = .222$
	Anger-In	$r(27) = .43$	$r(27) = .20$
		$p = .010$	$p = .147$
	Anger-Control	$r(27) = -.35$	$r(27) = -.29$
		$p = .03$	$p = .061$

4.3 Correlations Between the Different Measures

For the description of the relationships between the dependent measures, Pearson coefficients of correlation were calculated (see Table 3), revealing the following relationships:

- Self-reported anger correlated significantly negative with task performance. The lower task performance between two measuring points the higher is the subjectively perceived anger.
- Facial expression correlated significantly positive with self-reported anger. Participants who reported high anger-values during the drive also had a more intensive facial expression.

- In addition, participants showed stronger emotions in their faces when they could not perform a task successfully.

The four subscales of STAXI were correlated with each other and with the mean values of self-reported anger as well as the mean values of negative facial expressions. As expected, participants with higher scores on the Trait-Anger scale had also higher scores on the Anger-Out and Anger-In subscale but lower scores on the Anger-Control subscale. In addition, there was a significant positive correlation with the self-reported anger but no significant correlation with the facial expression. Trait-Anger therefore reflects whether and to what extend a person gets angry but gives no information on the appearance or extend of facial expressions.

Furthermore, participants with higher Anger-In scores expressed higher subjective anger but did not show that mandatorily in their faces. Against expectations, Anger-Out did neither correlate with self-reporting of anger nor with the facial expression although this subscale intended to measure the degree to what a person expresses his/her feelings.

5 Discussion and Implications

It could be shown that experiencing difficulties in solving IVIS tasks goes along with self-reported negative affect as well as according facial expressions. The reliability and validity of the newly developed rating tool could also be shown. The study has several implications that are important both for the design of IVIS and advancement of theories of user experience.

Firstly, designers should be aware that cumbersome interaction design will cause negative emotions which are likely to result in non-acceptance of new products. It is well known that negative affect such as anger and frustration can also cause a more risky and ruthless driving style, so that such negative impacts on the driver's emotional state could ultimately become a safety risk. Secondly, the study is a first step into measuring and eventually preventing such disadvantageous developments in the design process. The study also shows the potential of facial recognition technology to provide assistance or tutoring functions that could relieve the driver from emotional discomfort.

From a theoretical point of view, it is somewhat surprising that anger expressions were not correlated with the STAXI subscales, as the STAXI is the most used questionnaire that measures personality traits related to anger expression. From this finding, it appears that significantly more research has to be conducted to fully understand the development and expression of negative affect (such as anger) related to technology use.

References

1. Hennessy, D.A., Wiesenthal, D.L.: The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours. Ergonomics **40**(3), 348–361 (1997)
2. Underwood, G., Chapman, P., Wright, S., Crundall, D.: Anger while driving. Transp. Res. Part F **2**(1), 55–68 (1999)

3. Wells-Parker, E., Ceminsky, J., Hallberg, V., Snow, R.W., Dunaway, G., Guiling, S., Williams, M., Anderson, B.: An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers. Accid. Anal. Prev. **34**(3), 271–278 (2002)
4. Bachfischer, K., Henze, S., Wäller, C.: Emotional user interfaces in the car. In: Proceedings of the 1st Workshop on Emotion and Computing—Current Research and Future Impact. In: Reichardt, D., Levi, P., Meyer, J.-J.C. (eds.), pp. 55–59. Bremen, Germany, June 2006
5. Malta, L., Miyajima, C., Kitaoka, N., Takeda, K.: Analysis of real-world driver‘s frustration. IEEE Trans. Intell. Transp. Syst. **12**(1), 109–118 (2010)
6. Kuhn, A., Paek, T., Medenica, Z.: The effect of speech interface accuracy on driving performance. In: Interspeech (2007)
7. Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., Thien, N.N.: Emotion on the road: necessity, acceptance, and feasibility of affective computing in the car. Adv. Hum.-Comput. Interact. **2010**, 1–17 (2010)
8. Zhang, M., Ihme, K., Jipp, M.: Maschinelle Erkennung der Frustration von Autofahrern aus Videoaufnahmen des Gesichts. In: 5 Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten. Bochum, Deutschland (2016)
9. Ekman, P., Friesen, W.V., Hager, J.: The Investigator’s Guide for the Facial Action Coding System. A Human Face, Salt Lake City (2002)
10. Schwenkmezger, P., Hodapp, V., Spielberger, C.D.: Das State-Trait-Ärgerausdrucks-Inventar STAXI: Handbuch. Hans Huber, Bern (1992)
11. National Highway Traffic Safety Administration (NHTSA): Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Notice in the Federal Register, vol. 77, no. 37, 24 February 2012
12. Heller, O.: Hörfeldaudiometrie mit dem Verfahren der Kategorienunterteilung (KU). Psychologische Beiträge **27**, 478–493 (1985)
13. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled agreement or partial credit. Psychol. Bull. **70**, 213–220 (1968)

Measurement of JND Thresholds and Riemannian Geometry in Facial Expression Space

Runa Sumiya[1], Reiner Lenz[2], and Jinhui Chao[1(✉)]

[1] Department of Information and System Engineering,
Chuo University, Tokyo, Japan
rsumiya603@gmail.com, jchao@ise.chuo-u.ac.jp

[2] Linköping University, SE-58183 Linköping, Sweden
reiner.lenz@liu.se

Abstract. Currently the most popular approach to facial expression analysis uses categorical representations of expressions based on labels like sad, happy and angry. Subtle expression variations require however a quantitative and continuous representation. Besides, today's subjective expression spaces, built by semantic differential level scores and reduced to low dimensional continuous spaces using MDS or PCA have no direct correspondence with the physical stimuli or the expression images. On the other hand, the spaces used in engineering are based on purely physical stimuli or images which can hardly be called expression spaces. Even in models incorporating spacial structure, the geometry of the expression space received little attention and is usually assumed to be Euclidean. The aim of this paper is to build an expression space which is directly connected with the physical stimuli or the expression images. At the same time, it has to incorporate the subjective characteristics of expression perception. We use methods from psychophysics to build an expression space based on the physical stimuli or expression image space equipped with JND or discrimination threshold data. The construction follows the approach used in color science where the MacAdam ellipsoids provide for every color a metric tensor in a Riemannian space. We show that the discrimination thresholds indicate that the space is not Non-Euclidean. We will also illustrate the intrinsic geometrical structure of the expression spaces for several observers obtained from two large image databases of face expressions.

Keywords: Facial expressions · Emotions · Categorical theory
Dimensional theory · JND thresholds · Riemannian geometry

1 Introduction

Face expressions and emotions are now mainly represented by categories such as basic expressions and emotions. In particular, most expression recognition systems are based on features defined by discrete levels of Action Units (AUs) to classify an arbitrary expression image into some basic category [1,2].

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 453–464, 2018.
https://doi.org/10.1007/978-3-319-91238-7_37

Two theories known in expression perception researches in psychology e.g. [20] are the categorical theory by Ekman [4] which assumes the existence of universally invariant basic expressions among different races and cultures which provide a model of expression perception by classification into these discrete categories described by language labels. The dimensional theory by Schlosberg [5], Russell and Bullock [6] seeks geometric relationships between expressions by arranging the basic expressions into a circular order in a psychological space obtained from multidimensional scaling (MDS) [7]. The dimensional theory could be regarded as an extension rather than a denial of the categorical theory. On the other hand, it was also suggested that there is a natural way to describe emotions or expressions as a continuous distribution in the psychological space.

Current representations of expressions often use verbal labels of the categorical names thus providing a qualitative characterization only. Facial expression images or their features are reduced to certain low dimensional space by e.g. principal component analysis (PCA) or MDS and are then divided into basic categories based on AU matching. The abundance of subtle and delicate, but unnamed expressions and their variations are hard to describe without a quantitative representation. An attempt to eliminate possible linguistic effects (due to the labeling of the expressions using natural language) is due to Russell who used preschool children in their experiments. In the following we assume that the inner structure and the relationships between expressions are also important therefore quantitative descriptions are required. Such relationships, including similarity, can be described intuitively by spatial or geometric properties in an expression space.

Engineering approaches to facial expression recognition have mainly been of the categorical type, trying to classify facial expressions. Methods using more detailed and complicated facial features trying to provide a more detailed analysis have, however, received more interest recently [2]. Combinations of basic expressions are used to describe compound expressions as sub-categories [3]. These approaches provide more accurate, however still discrete descriptions, of expressions. On the other hand, it seemed that a quantitative representation in the form of coordinates of a continuous space could unify the categorical and dimensional theories since the categories can be regarded as domain decomposition of the continuous expression space, just as color categorization is obtained in a color space [11].

Another problem with today's subjective expression spaces is that all of them were built from psychological evaluations such as the affect grid or semantic differential (SD) level scores. These data are reduced to low dimensional continuous spaces using MDS or PCA. As a result, these psychological expression spaces have no direct correspondence with the physical stimuli or the facial expression images. On the other hand, the spaces used in engineering are based on purely physical stimuli or facial images which contain no information of subjective perception. In this paper we use a psychophysical approach to build an expression space based on the physical stimuli or facial expression image space and equip it at every point with the expression JND discrimination thresholds to incorporate subjective characteristics in expression perception. This new expression space is actually a Riemann space with a metric tensor defined by the JND thresholds. In fact, this approach is well known in color science where the color space is defined in the RGB stimuli space equipped with color JND thresholds known as MacAdam ellipses and ellipsoids [13,14].

There is also an issue that even in researches used spatial representation of expressions, either in psychological or engineering usage, the geometry of the space received little attention. Dimensional theories, for example, tried to explain why the placement of basic expressions obtained from psychological rating or SD score are far from a perfect circle [9], while engineers used the Euclidean distance between feature vectors such as AU or other descriptors in the PCA subspaces as measure of similarity for classification and recognition. All these methods are based a tacit assumption that the expression space is a Euclidean space which, according to the above arguments, requires to be tested. In this paper, we introduce a framework which enables us to analyze intrinsic geometry in the space of facial expressions.

It is known that it is hard to compare two remote expressions and even harder to describe their difference quantitatively. This suggests that the global geometry of expression space is hard to investigate or properties to be measured. On the other hand, it is easy to compare two expressions close to each other or with a subtle difference and so to obtain objectively stable measurements. In fact, the discrimination threshold measurements are known in psychophysics to be one of the most fundamental ways for substantial understanding of phenomena [10]. Furthermore, the discrimination thresholds are known to define a metric tensor describing intrinsic geometry in a Riemann space [23].

Thus, in this paper we measured the JND discrimination thresholds of facial expressions and draw them as ellipsoids in a 3D expression space, by simultaneous comparison between two facial expressions avoiding the influence of language and category judgments. The data is produced by morphing between different expression images. The results show that all the discrimination threshold ellipsoids have very different shapes and sizes for different basic expressions. The measurements from different subjects show similar trends in these variations. Since the discrimination thresholds as ellipsoids are subjectively unit spheres which have the same size and shape everywhere, it is a strong evidence that the expression space is not a Euclidean space. Indeed, these discrimination thresholds define local metric tensor for every expressions so a natural outcome is that one obtains the expression space as a Riemann space. We will show a finer distribution of 23 JND thresholds ellipsoids from a single subject in 2D and 3D PCA spaces, which show a smooth transition of local geometry among different expression and certain distinct features of the Riemann space.

2 Discrimination Thresholds and Riemann Space

2.1 Definitions of Thresholds and a Riemann Space

It is known that there are two types of thresholds in psychophysics [10]. The stimulus thresholds are the minimal stimulus to invoke a sensation. The other type are the discrimination thresholds which are defined as the just noticeable difference (JND) between two simultaneously presented stimuli.

It is also known that a Riemann space is a space S in which a bilinear function or an inner product $(\cdot, \cdot)$ is defined at the tangent space $T_x S$ for every point $x \in S$ such that

$$(dx, dy) := dx^T G(x) dy, \qquad \forall dx, dy \in T_x S$$

where the matrix $G(x)$ is smoothly defined for every $x \in S$. The $G(x)$ defines local geometry such as distances and angles around the point x. It is known as the (Riemann) metric tensor [23].

It is well known that a color space is a Riemann space defined by the metric tensor obtained from the MacAdam's ellipsoids and ellipses [13,14]. Similarly, we will use the JND thresholds near an expression as the unit sphere in local distance which describes the perceptional difference between slight variations of the expression in the expression space. This information at every point in the space then defines the Riemann metric tensor therefore giving the expression space the structure of a Riemann space.

It is reported in [8] that tests along a morphing image sequence between basic expression images show a sharp boundary between basic expressions A and B. This result is then used as an major evident, so called as a paradox against dimensional theory. In fact, those reported thresholds could be thresholds in categorical perceptions which are results of discontinuity between categories when the subjectives have no other choices besides A and B. In our measurement of the JND thresholds, as will be mentioned later, we will take care to avoid categorical judgments and influence of verbal labels of expression categories.

2.2 From Local Properties to Global Geometry

We hereafter consider a continuous space and its discrimination thresholds. It is known that it is easier for humans to judge relative and small differences than abstract levels of a stimulus. On the other hand, it is difficult to recover a global perception from a greater collection of local relative information. The celebrated Weber-Fechner law is an example showing how to obtain a global law in the whole space from local laws at every different point. In this case the logarithm function between the sensation and stimulus is obtained. Its success lies in the fact that the local thresholds are represented by an ODE which is fortunately integrable and one can obtain the global law in terms of elementary functions. In fact, the disagreement between the Weber-Fechner law and data comparing with Stevens' law is not due to this local to global integration strategy but to the error in the local Weber's law and Stevens' law can be obtained from integration using a correct local law. Unfortunately such closed forms are not always possible since an integration or solution of an arbitrary ODE, even when it exits, can rarely be expressed in a closed form in terms of elementary functions. On the other hand, nowadays one does not need such a closed form as long as the functional relationships are computable. e.g. a LUT will do almost the same if not better.

2.3 1D vs Higher Dimensions

Another fact is that most theories in psychophysics were about stimulus-response of 1-dimensional data. This is understandable considering that the integration from local information to a global representation would become much harder when the stimuli and responses are high dimensional. In fact, it seemed there are few new additions after the Weber-Fechner and Stevens' laws. In particular, the only reports for discrimination thresholds researches in multi-dimensional case were those about color perception. The first 1D discrimination thresholds along straight lines in a color space were measured by Wright [12], which was followed by MacAdam [13] to measure discrimination thresholds in 2D or chromaticity plane, so-called MacAdam ellipses. The 3D discrimination thresholds which are ellipsoids in a color space were obtained by Brown-MacAdam [14]. One of major implications of these data is that the discrimination threshold show that a color space is not an Euclidean space but a distorted or curved space, which can be modeled as a Riemann space. In fact, colors lie on these discrimination thresholds, ellipses or ellipsoids, have the same subjective difference from the center colors, so they should form a unit circle or sphere and have the same size and shape everywhere. MacAdam ellipses show however various sizes and shapes for different center colors which then provide definite evidents that a color space is not a Euclidean space. The thresholds which define subjective distances depending on center colors then provide a Riemann metric for the color space [11]. Such intrinsic geometry is then used to apply the tools of Riemann geometry to find fruitful applications e.g. [21,22].

In this research, we are aiming at an application of the above successful strategy to the perception of facial expressions.

3 Measuring JND Discrimination Thresholds

We used images from two database of expression images: the Japanese Female Facial Expression (JAFFE) Database [15] and the Bosphorus Database [17–19]. The images contain seven basic expressions: Anger, Disgust, Fear, Sadness, Happiness, Surprise and Neutral and AU images from one person in each database. These images are used to create morphing image sequences combining all pairs among different expressions. This is done using the FUTON system (ATR Japan) [16] using the feature points shown in Fig. 2.

The test and a comparison expression image are shown simultaneously on a screen (Fig. 1). To avoid adaptation and prediction, a 2 s interval (showing a neutral background) is inserted between sessions, the order of the images is randomized. The observer is instructed to compare the two expressions and answer if they are the "same" or "different" without interpreting them, and no hints on the labels of the expressions (emotions) are provided. We instruct the observer to ignore the type of the expression and to answer only if the images show identical expressions or not. By discouraging attempts to interpret the expressions we hope to minimize the influence of category perception during the process.

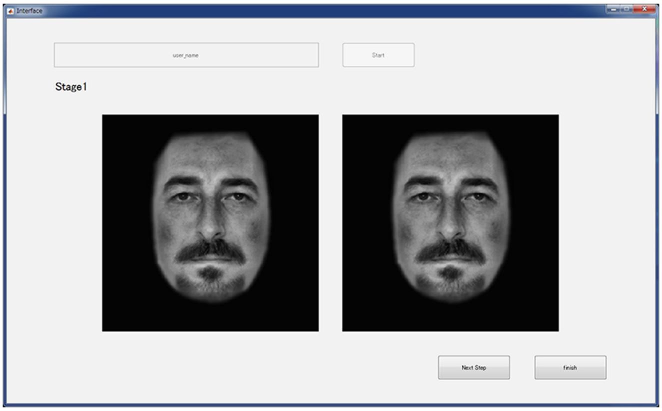


Fig. 1. Measurement environment

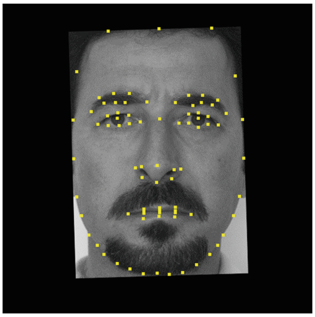


Fig. 2. Feature points

The observers are also informed in advance that the changes are about 1% so it is possible that they will not be able to discriminate the different images. The observers are also instructed to vote "different" when the decision takes more than 5 s, in order to avoid prediction, classification into emotion categories and to minimize the influence of other high-level cognition processes. There is a 5 min break after each session. Each comparison of two images was repeated three times, inserted among all image pairs in a random order.

The images and threshold data are then reduced by PCA using the covariance matrix and MDS to the first 2D and 3D eigen subspaces. The discrimination ellipsoids are then fitted using a Gaussian or RFB function fitting method.

In the first experiment we used 7 basic expression images from JAFFE to produce 21 morphing sequences. 2086 images are projected by PCA to 2D and 3D spaces while 126 threshold points were used to fit 7 ellipsoids shown in Figs. 3, 4 and 5. They show that the obtained expression spaces are not Euclidean, since JND ellipsoids define subjective unit spheres. Besides, the personal

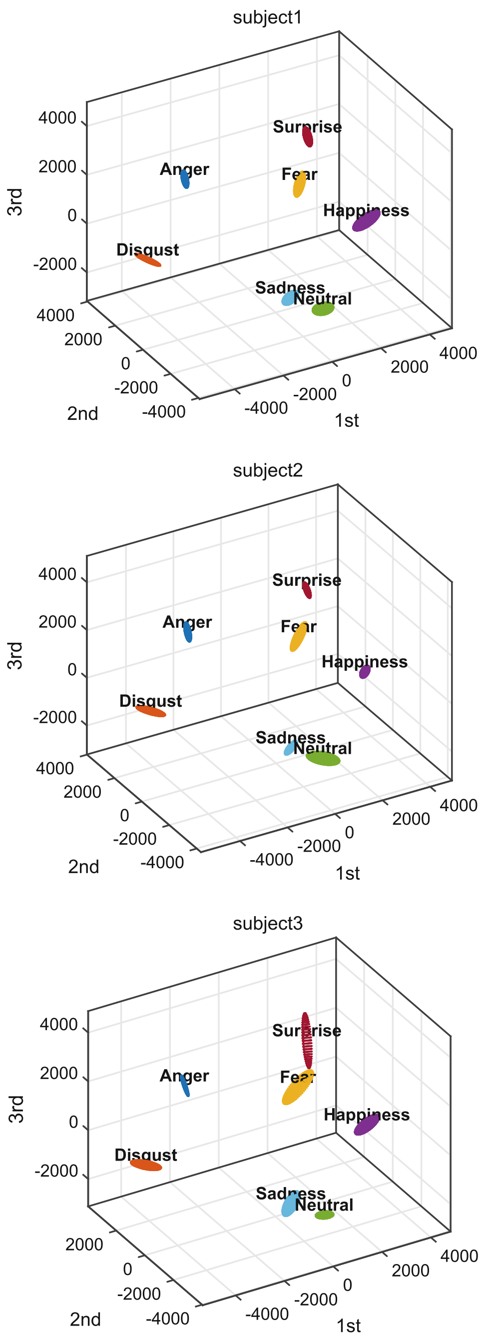


Fig. 3. Subject 1, 2, 3; JND thresholds in 3D PCA space

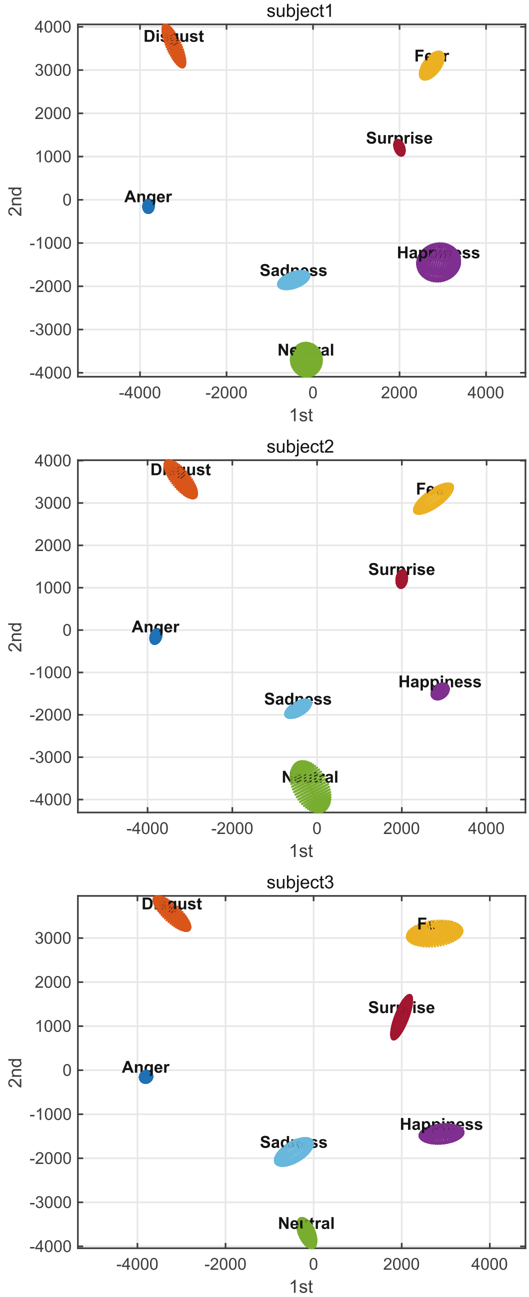


Fig. 4. Subject 1, 2, 3; JND thresholds in the 1st-2nd PCA space

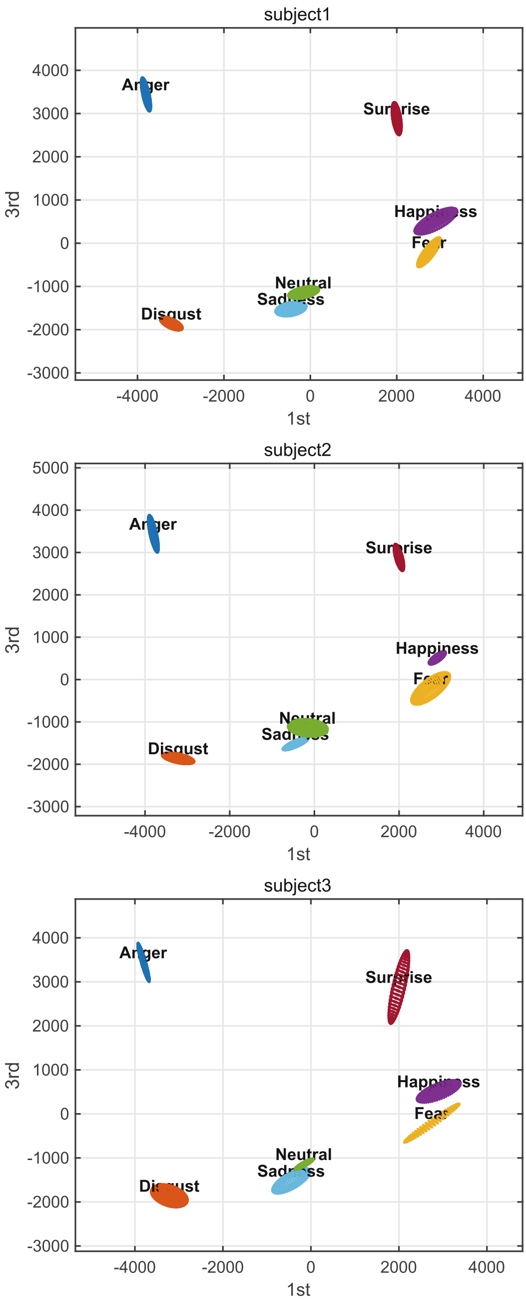


Fig. 5. Subject 1, 2, 3; JND thresholds in the 1st-3rd PCA space

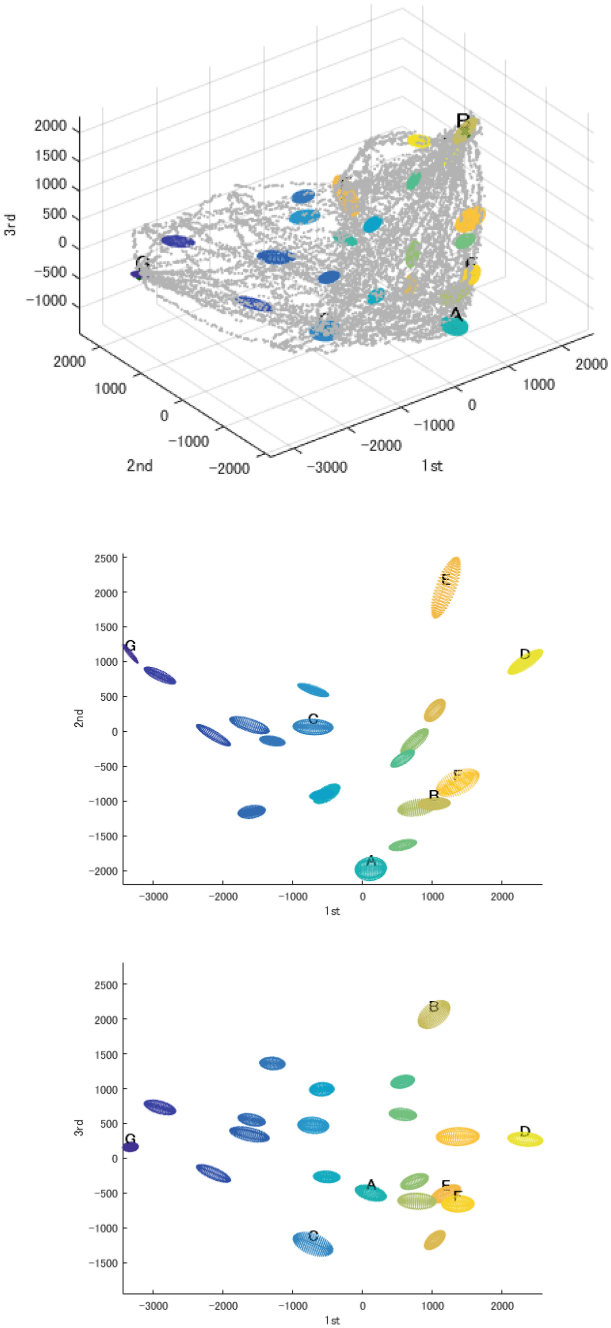


Fig. 6. 23 ellipsoids in the 3D, ellipses in 1st-2nd, 1st-3rd PCA spaces

variations between observers can be quite large. However, there are clear structural similarities among the JND thresholds distributions in the expression spaces of different observers.

In the second experiment we used 7 basic expressions and 21 AU images from the Bosphorus database to produce 616 morphing sequences. 57844 images were projected by PCA to 2D and 3D spaces, 2233 threshold points were used to fit 23 ellipsoids from a single observer. The results of 3D, 1st-2nd and 1st-3rd PCA projections are shown in Fig. 6. They indicate a smooth variation of shapes, directions and sizes of the ellipsoids and a global flow in the space, which indicate the metric tensor smoothly defined over the whole space and distinct features of the intrinsic geometry of the Riemann space.

4 Discusssion, Conclusions and Future Work

The experiments require controlled conditions therefore are very labor intensive and time consuming. Further measurements are necessary to verify the results reported above and also to understand how the different steps in the experimental setup influence the final results. Besides, efficient approaches to produce natural and accurate morphing sequences of facial expressions are also important.

The facial image space is of very high dimensional, here we only illustrated JND sections in 2D and 3D subspaces of PCA space for easy visibility. JND thresholds in higher dimensional spaces can be obtained in the same way but with much more measurement data. Further work is therefore required to fix the expression space and estimate its dimension. Dimensionality reduction and other possibility of psychophysical spaces may also be considered. A problem to use features extracted from facial images is the difficulty to produce morphing sequences corresponding to trajectories of shifts in the feature space.

Applications are expected in theoretical modeling, performance improvement in facial expression analysis and recognition, etc. We are, however, confident that the Riemannian approach to expression analysis is necessary and that it will play a similarly important role as in the study of color perception.

Acknowledgment. The authors wish to thank Hajime Tasaki for discussions and cooperation in this research.

References

1. Li, S.Z., Jain, A. (eds.): Handbook of Face Recognition. Springer, New York (2011). https://doi.org/10.1007/978-0-85729-932-1
2. Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M.: Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
3. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proc. Nat. Acad. Sci. **111**(15), E1454–E1462 (2014)

4. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
5. Schlosberg, H.: The description of facial expression in terms of two dimensions. J. Exp. Psychol. **44**(4), 229–237 (1952)
6. Russell, J.A., Bullock, M.: Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. J. Pers. Soc. Psychol. **48**(5), 1290–1298 (1985)
7. Richardson, M.W.: Multidimensional psychophysics. Psychol. Bull. **35**(9), 659–660 (1938)
8. Calder, A.J., Young, A.W., Perrett, D.I., Rowland, D.: Categorical perception of morphed facial expressions. Vis. Cogn. **3**, 81–117 (1996)
9. Takehara, T., Ochiai, F., Suzuki, N.: Fractals in emotional facial expression recognition. Fractals **10**, 47–52 (2002)
10. Gescheider, G.A.: Psychophysics, The Fundamentals, 3rd edn. Lawrence Erlbaum Associates, Mahwah (1997)
11. Wyszecki, G., Stiles, W.S.: Color Science; Concepts and Methods, Quantitative Data and Formulae, 2nd edn. Wiley, New York (1982)
12. Wright, W.D., Pitt, F.H.G.: Hue-discrimination in normal color-vision. Proc. Phys. Soc. **46**(3), 459–473 (1942)
13. MacAdam, D.L.: Visual sensitivities to color differences in daylight. J. Opt. Soc. Am. **32**(5), 247–274 (1942)
14. Brown, W.R.J., MacAdam, D.L.: Visual sensitivities to combined chromaticity and luminance differences. J. Opt. Soc. Am. **39**(10), 808–834 (1949)
15. Lyons, M.J., Akamastu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with Gabor wavelets. In: 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
16. Mukaida, S., Kamachi, M., Oda, M., Kato, T., Yoshikawa, S., Akamatsu, S., Chihara, K.: Facial image synthesis system: FUTON: evaluation as tools for cognitive research on face processing. J. Inst. Electr. Inf. Commun. Eng. **A 85**(10), 1126–1137 (2002)
17. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B.L., Akarun, L.: Bosphorus database for 3D face analysis. In: The First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008), Roskilde University, Denmark, 7–9 May 2008
18. Savran, A., Sankur, B.L., Bilge, M.T.: Comparative evaluation of 3D versus 2D modality for automatic detection of facial action units. Pattern Recogn. **45**(2), 767–782 (2012)
19. Savran, A., Sankur, B.L., Bilge, M.T.: Regression-based intensity estimation of facial action units. Image Vis. Comput. **30**(10), 774–784 (2012)
20. Young, A.W., Rowland, D., Calder, A.J., Etcoff, N.L., Seth, A., Perrett, D.I.: Facial expression megamix: test of dimensional and category accounts of emotion recognition. Cognition **63**(3), 271–313 (1997)
21. Mochizuki, R., Kojima, T., Lenz, R., Chao, J.: Color-weak compensation using local affine isometry based on discrimination threshold matching. J. Opt. Soc. Am. A Optica Image Sci. Vis. **32**(11), 2093–2103 (2015)
22. Oshima, S., Mochizuki, R., Lenz, R., Chao, J.: Modeling, measuring, and compensating color weak vision. IEEE Trans. Image Process. **25**(6), 2587–2600 (2016)
23. do Carmo, M.P.: Riemannianian Geometry. Birkhäuser, Boston (2013)

A System for Non-intrusive Affective Assessment in the Circumplex Model from Pupil Diameter and Facial Expression Monitoring

Sudarat Tangnimitchok[1(✉)], Nonnarit O-larnnithipong[1], Neeranut Ratchatanantakit[1], Armando Barreto[1], Francisco R. Ortega[2], and Naphtali D. Rishe[2]

[1] Electrical and Computer Engineering Department, Florida International University, Miami, FL 33199, USA
{stang018,nolar002,nratc001,barretota}@fiu.edu

[2] School of Computer and Information Sciences, Florida International University, Miami, FL 33199, USA
fortega@CS.edu, ndr@acm.org

Abstract. This paper outlines a system for non-intrusive estimation of a user's affective state in the Circumplex Model from monitoring the user's pupil diameter and facial expression, obtained from an EyeTech TM3 eye gaze tracker (EGT) and a RGB-D camera (KINECT) respectively. According to previous studies, the pupillary response can be used to recognize "sympathetic activation" and simultaneous "parasympathetic deactivation", which correspond to affective arousal. Additionally, tracking the user's facial muscle movements as he or she displays characteristic facial gestures yields indicators to estimate the affective valence. We propose to combine both types of information to map the affective state of the user to a region on the Circumplex Model. This paper outlines our initial implementation of such combined system.

Keywords: Affective computing · Facial expression recognition
RGB-D camera · Eye-Gaze tracking

1 Introduction

It has been two decades since Affective Computing pioneer Rosalind Picard envisioned a new generation of computers that could interact with their human users at an affective level [5]. However, the complete fulfillment of that goal remains elusive and, already well into the XXI Century, the everyday use of affective computing remains limited. The difficulties associated with the actual implementation of an affective computing system might be best appreciated if one considers the 3 fundamental tasks that must be performed to fully animate the performance of an affective computing system (affective computer), as outlined by Hudlicka [7]: These tasks can be described as:

1. Affect Sensing and Recognition
2. User Affect Modeling/Machine Affect Modeling
3. Machine Affect Expression

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 465–477, 2018.
https://doi.org/10.1007/978-3-319-91238-7_38

The affective sensing and recognition tasks aim at making the machine aware of the affective state of the human user. This will require sensing some observable manifestations of that affective status and recognizing (or "cataloging") the state, so that, then, the machine may determine (by following some pre-programmed interplay guidelines) which affective state it should adopt in response, and, further, the type of affective expression that it should present to the user. Those initial stages of the process, however, may involve some of the major challenges that are presented for the implementation of a fully-functional affective computing system. In fact, Picard identified "Sensing and recognizing emotion" as one the key challenges that must be conquered to bring the full promise of affective computing concepts to fruition [6] (Fig. 1).

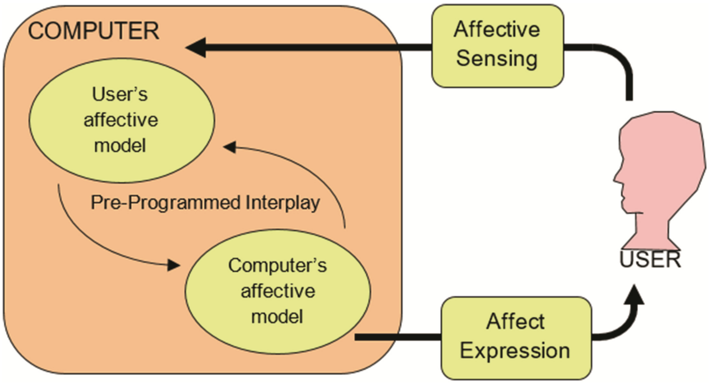


Fig. 1. Simplified diagram showing the interaction between the key processes in affective computing identified in [7]. (Diagram reproduced from [8])

In the pursuit of solutions for that important challenge, there have been may approaches proposed. Specifically, a wide variety of mechanisms have been suggested for affective sensing. Some research groups have attempted the assessment of user affective states using streams of data that are commonly available in contemporary computing systems, such as video from the user's face, audio from the user's voice and text typed by the user on the keyboard.

Zeng et al. [25], provided an interesting survey of relevant systems that use video and/or audio, to estimate the user's affective state. Most vision-driven approaches are based in the known changes that occur in the geometrical features (shapes of eye, mouth, etc.) [10] or facial appearance features (wrinkles, bulges, etc.) [11] of the subject, according to different affective states. Cowie et al. associated acoustic elements to prototypical emotions [9]. Some other groups explored the coordinated exploitation of audio-visual clues for affective sensing [12]. Liu et al. focused on the utilization of text typed by the user for affective assessment [13]. Approaches in this area of work include "Keyword Spotting" (e.g., [14]); "Lexical Affinity" (e.g., [15]); "Statistical Natural Language Processing" (e.g., [16]); etc.

Other groups have attempted to identify the physiological modifications that are directly associated with the affective states and transitions in human beings, and have proposed methods for sensing those physiological changes in ways that are non-invasive and unobtrusive to a computer user. The reconfiguration experimented by a human subject as a reaction to psychological stimuli is controlled by the Autonomic Nervous System (ANS), which innervates many organs and structures all over the body. The ANS can promote a state of restoration in the organism, or, if necessary, cause it to leave such a state, favoring physiologic modifications that are useful in responding to the external demands.

The Autonomic Nervous System coordinates the cardiovascular, respiratory, digestive, urinary and reproductive functions according to the interaction between a human being and his/her environment, without instructions or interference from the conscious mind [17]. According to its structure and functionality, the ANS is studied as composed of two divisions: The Sympathetic Division and the Parasympathetic Division. The Parasympathetic Division stimulates visceral activity and promotes a state of "rest and repose" in the organism, conserving energy and fostering sedentary "housekeeping" activities, such as digestion [17]. In contrast, the Sympathetic Division prepares the body for heightened levels of somatic activity that may be necessary to implement a reaction to stimuli that disrupt the "rest and repose" of the organism. When fully activated, this division produces a "flight or fight" response, which readies the body for a crisis that may require sudden, intense physical activity. An increase in sympathetic activity generally stimulates tissue metabolism, increases alertness, and, from a global point of view, helps the body transform into a new status, which will be better able to cope with a state of crisis. Parts of that re-design or transformation may become apparent to the subject and may be associated with measurable changes in physiological variables. Variations in sympathetic and parasympathetic activation produce physiological changes that can be monitored through corresponding variables, providing, in principle, a way to assess the affective shifts and states experienced by the subject. Parasympathetic and sympathetic activations have effects that involve numerous organs or subsystems, appearing with a subtle character in each of them.

Therefore, one approach to affective sensing might be based on monitoring the changes in observable variables that are brought about by an imbalance in the sympathetic-parasympathetic equilibrium introduced by sympathetic activation. These changes can then be matched to the fundamental types of states for which each of these divisions of the Autonomic Nervous System prepare us (The sympathetic response prepares us for "fight or flight", whereas the parasympathetic response sets us up for "rest and response"). Accordingly, the predominance of sympathetic activity can very well be taken as an indicator of "arousal", represented on the vertical axis of Russell's Circumplex Model of Affect [3]. It is, indeed, common to experience acceleration of our heart rate (evidence of sympathetic activation) both, while we take a crucial test *and* when our favorite sports team is winning a match (Fig. 2).

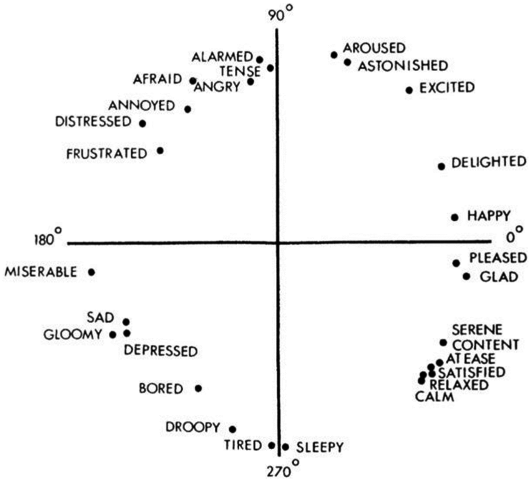


Fig. 2. A Circumplex Model of Affect (taken from [3])

Much of our previous work has focused on signal processing methods to estimate a level of sympathetic activation using data recorded from non-invasive physiological sensors, such as Electro-Dermal Activity (EDA), also referred to as "Galvanic Skin Response" (GSR), and, most promising due to its complete unobtrusiveness, Pupil Diameter (PD) monitoring, using infrared video analysis (commonly used in eye gaze tracking, EGT equipment).

However, a more helpful characterization of the user's affective state would also require an indication of the "valence" (horizontal axis in the Circumplex Model). This paper outlines the current direction we have taken to integrate a completely unobtrusive affective assessment system that supplements the arousal estimation provided by pupil diameter monitoring with valence indications derived from the monitoring and classification of key facial features, made possible by the video and depth sensors working in synergy within the KINECT sensor. In the following sections, the paper describes: The rationale and implementation of our arousal assessment through pupil diameter monitoring; The mechanisms used to obtain valence indications from the measurements performed by the KINECT module; and the way in which we are integrating both these modules. The last sections of the paper include some concluding remarks and reflections on the way ahead in the development of this research.

2 Arousal Assessment by Pupil Diameter

As indicated above, our approach to assessing the level of arousal experienced by the subject is through the monitoring of the pupil diameter, measured, in real time, by many eye gaze trackers (EGTs). This approach, in fact, targets the estimation of "sympathetic activation" (and simultaneous parasympathetic deactivation) in the Autonomic Nervous System (ANS). Formerly, our group has explored the monitoring of pupil diameter from a computer user, utilizing an ASL-504 eye gaze tracker, which reports the estimated pupil diameter in pixels (integer values), for the assessment of affective states in the user [18]. This approach has a strong anatomical and physiological rationale. The diameter of this circular aperture is under the control of the ANS through two sets of muscles. The sympathetic ANS division, mediated by posterior hypothalamic nuclei, produces enlargement of the pupil by direct stimulation of the radial dilator muscles, which causes them to contract [19]. On the other hand, pupil size decrease is caused by excitation of the circular pupillary constriction muscles innervated by the parasympathetic fibers. The motor nucleus for these muscles is the Edinger-Westphal nucleus located in the midbrain. Sympathetic activation brings about pupillary dilation via two mechanisms: (i) an active component arising from activation of radial pupillary dilator muscles along sympathetic fibers and (ii) a passive component involving inhibition of the Edinger-Westphal nucleus [20]. Our rationale is also supported by other independent experiments in which pupil diameter has been found to increase in response to stressor stimuli. Partala and Surakka used sounds from the International Affective Digitized Sounds (IADS) collection [21] to provide auditory affective stimulation to 30 subjects, and found that the pupil size variations corresponded to affectively charged sounds [22].

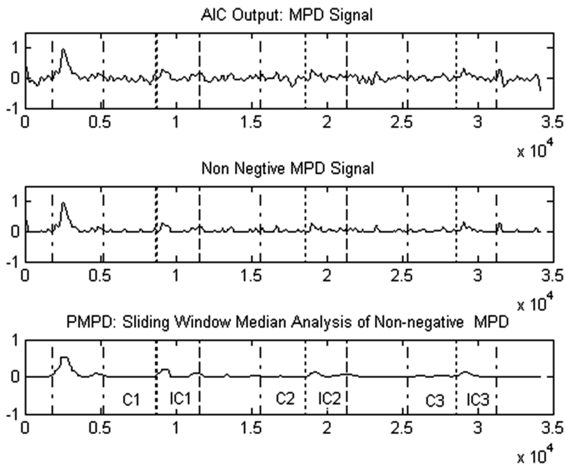


Fig. 3. (From [24]) The bottom panel shows the increased in the Processed Modified Pupil Diameter (PMPD) signal, which correspond to application of stressor ("Incongruent Stroop") stimuli, IC1, IC2 and IC3.

In our current work, we are obtaining measurements of the pupil diameters from both the left and the right eyes at a rate of 30 measurements per second, using a desktop infrared eyegaze tracker, the Eyetech TM3. This eyegaze tracking device operates (in part) by isolating the area of the pupil from the images captured by an infrared camera. The demarcation of the pupil edge is possible because the aperture of the pupil appears as a particularly dark region in the infrared images captured by the infrared camera ("Dark Pupil operation"). It is from that demarcated pupil geometry that the pupil diameter is estimated, in real time. Further details of the "Dark Pupil" principle of operation for eye gaze trackers can be found in the article by Morimoto and Mimica [23].

In our previous work [24], we have verified that an enlargement of the pupil diameter is observed when the subject experiences sympathetic activation from exposure to stressor stimuli ("incongruent" Stroop word presentations), therefore providing further support for the rationale of the combined system described in this paper. Figure 3 shows some of the results obtained.

In this figure, the elevations in the processed signal (PMPD), other than the initial transient at the beginning of the record, are seen to correspond with the intervals labeled "IC1", "IC2" and "IC3", which were the intervals of the experiment when the subject was presented with "incongruent" Stroop word presentations. The details of the experiment, as well as the method used to minimize the impact of potential pupil variations due to illumination changes, can be found in [24].

3 Valence Estimation from Analysis of Facial Features Through KINECT

Humans rely heavily on visual perception for affective sensing; especially when recognizing facial expressions. In general, we recognize an object in front of us by comparing its shape and features with those of objects we learned in the past. Similarly, recognizing human facial expressions can be achieved through the observation of prototypical changes in facial muscles. For example, we may recognize that a person is "happy" because he or she is "smiling".

To supplement our proposed arousal estimation through pupil diameter, and define a 2-dimensional location in the Circumplex Model of Affect [3], we propose a way to estimate the valence (horizontal axis of the model) by using facial expression as the indicator to determine a person's pleasure or displeasure state.

The Facial Action Coding System (FACS) [2], provides a strong foundation for facial gesture recognition. By deconstructing the anatomic components of a facial expression into the specific Action Units (AU), it is possible to code the facial expressions of known affective significance on the basis of the contraction and relaxation of facial muscles. These associations can be leveraged in recognizing affective states from facial gestures. Humans do this through their intrinsic visual perception. For example, we may infer that a person is "happy" by observing the way the corners of his/her mouth are lifted, or the shape of his/her eyes becomes narrower when a person smiles.

In this study, we utilize a Kinect V2 device, which includes a high resolution RGB-D camera, to extract features from a detected face image using its provided APIs. As part of its software framework, its Face APIs enables a wide variety of functionalities, including the delivery of 94 unique "shape units" to create meshes [1] that fit and track a human face in real-time.

It also provides facial points marking important locations such as eyes, cheek, mouth, etc. This allows the tracking of the movement of facial muscles in a way similar to the placement of physical markers on the user face, but in a less intrusive way. The analyzed face feature results then are continuously being updated in the programming object called "FaceFrameResult", as listed in Table 1 [4]. From this list, we are focusing on the features: "Happy", "Engaged", and "LookingAway". Our main interest is in the feature “Happy” as an indication of the pleasure or displeasure of the subject, while the other two features tell us if a user is engaging the system or not.

Table 1. List of face features

Heading level	Example
Happy	Happiness data is present
RightEyeClosed	Right eye data is present
LeftEyeClosed	Left eye data is present
MouthOpen	Mouth open data is present
MouthMoved	Mouth movement data is present
LookingAway	Looking away data is present
Glasses	Glasses data is present
FaceEngagement	Face engagement data is present

There are three possible output values for each feature in Table 1, which are “Yes”, “No”, and “Maybe”. We interpret the values of the “Happy” feature as follows: We interpret “Yes” as positive (pleasure), “Maybe” as neutral and “No” as negative (displeasure), hence, obtaining a basic estimation of valence. The next sections provide further details on the combined implementation of our arousal and valence estimation approaches. They will also describe how the results from both subsystems are mapped to the coordination in the Circumplex Model of Affect.

4 Implementation

In this study, we use two devices (hosted by two different computers) to obtain pupil diameter and facial expression.). Both computers communicate through an Ethernet link, using the TCP/IP protocol to share data between them. The system is shown in Fig. 4.

4.1 Pupil Diameter Acquisition

The TM3 eye gaze tracker, from EyeTech Digital Systems is used to obtain pupil diameter data. Its operation requires two steps of initialization, prior to its actual use. First, we run a test program to view the camera stream and fine-tune the angle and location of the TM3 so that it captures an adequate image of both eyes of the user. Secondly, we run a calibration program, where sixteen targets will be shown on the screen one by one. The user will be asked to maintain their head position and direct his/her gaze to the current target until the next one is shown. The process is repeated umtil all 16 targets are gazed upon. After the calibration is done, a calibration file will be generated and saved for the later use (See Fig. 5). Finally, we run a program called Gazeinfo2 to collect our eye gaze information. After the “listen” on-screen button (see Fig. 6) is clicked the program will act like a server waiting for another computer to send a request and then respond back with pupil diameter data.

Fig. 4. Entire system including Kinect V2 (on top of the screen) and TM3 (in front of the computer).

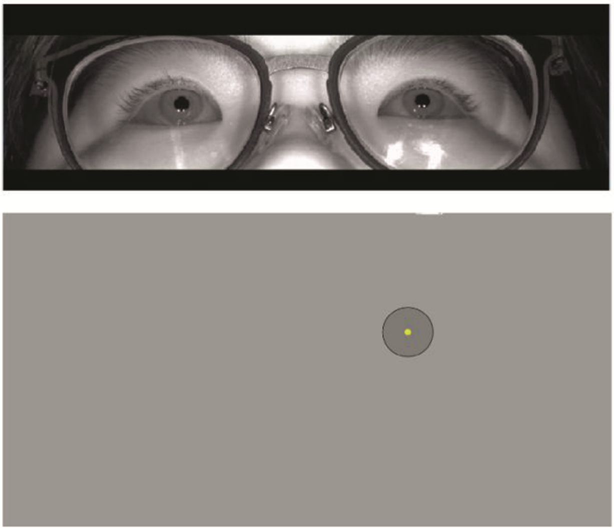


Fig. 5. Two steps of preparation: Adjust position of TM3 (top), Calibration process (bottom)

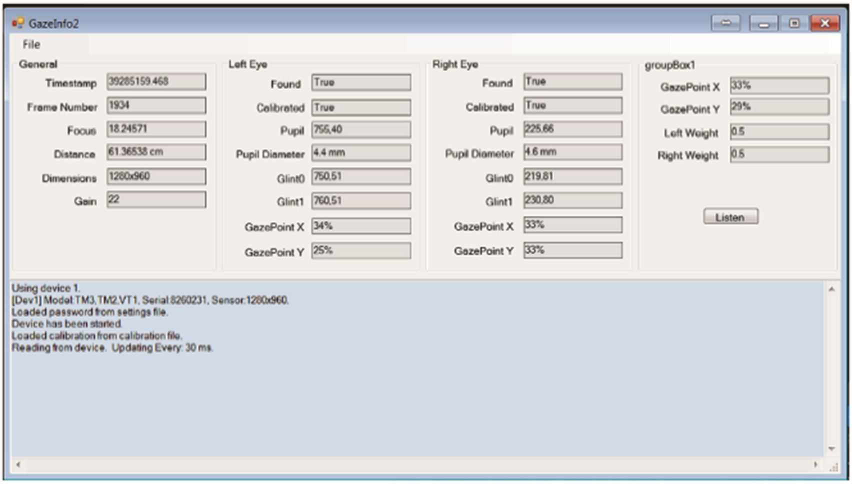


Fig. 6. Graphic interface of the Gazeinfo2 program

4.2 Facial Expression Acquisition

As we already mentioned how Kinect V2 estimates Facial expressions previously, this section describes our own program called "HD_Face", built on the Window Presentation Foundation (WPF) framework (See Fig. 7), which interacts with Kinect V2. Once Kinect V2 detects a user, a violet marker will appear on top of user's face in the video screen. This indicates that Kinect V2 is now collecting the user's facial expression. On the bottom right of the window, the facial expression indicators will now flash in red when

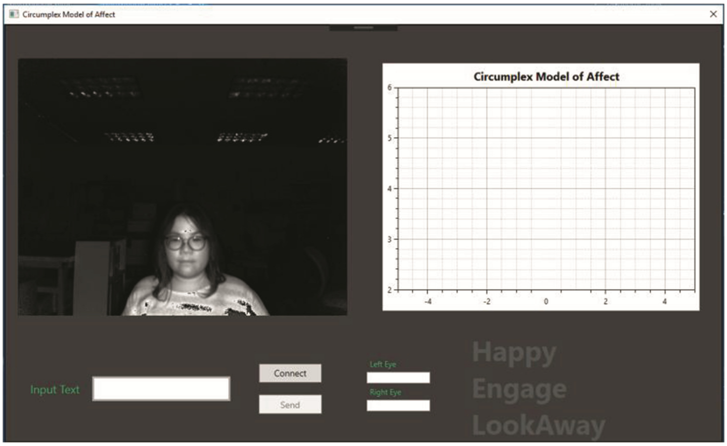


Fig. 7. The user interface of HD_Face program is shown. The top left panel displays the video from the infrared camera. The top right shows a plot of Circumplex Model of Affect. The bottom left is where the communication section is located. Lastly, on the bottom right is where the pupil diameter fetched from another computer and the Facial expression indicator are displayed.

they are asserted by Kinect V2 (For example, "Happy" will flash if the user smiles, as shown in Fig. 8). The other two facial expression indicators work in the same way. (Fig. 9). These two additional indicators provide information that help qualify the validity of the "Happy" indicator. For example, if the system knows that the user is "LookingAway", the absence of a smile detection should not directly be mapped to negative valence.

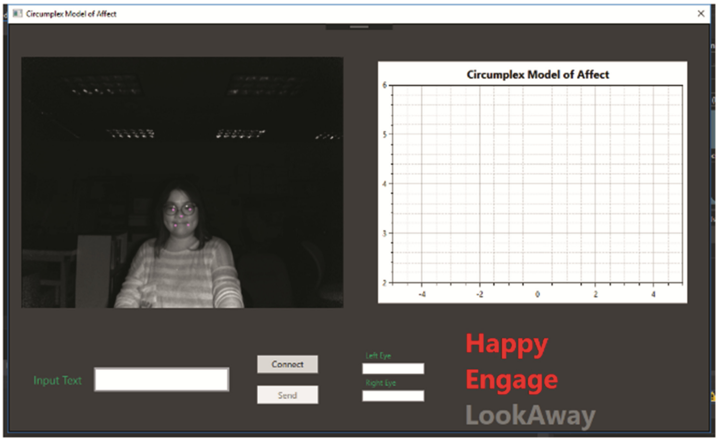


Fig. 8. Example of the facial expression indicator "Happy" flashing in red when a user displays a happy expression.

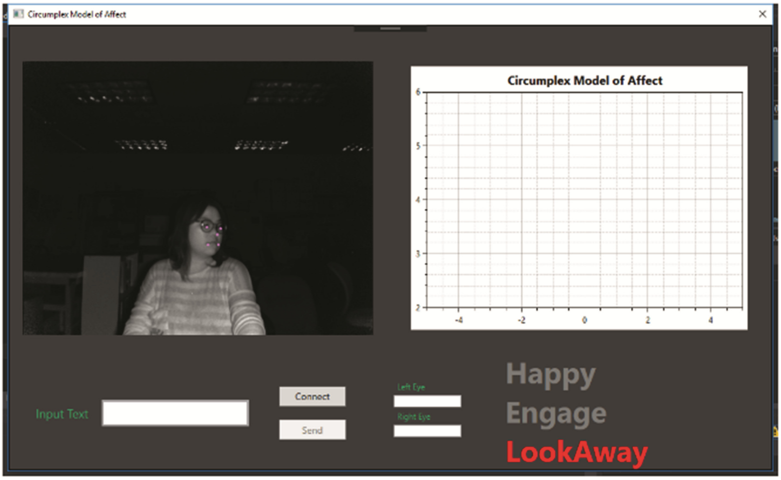


Fig. 9. Example of the facial expression indicator "LookAway" flashing in red when a user looks away.

4.3 Plotting a Circumplex Model of Affect

After making sure that the TM3 Eye Gaze Tracker subsystem is running properly and also verifying that the Kinect V2 is detecting the subject's facial expression (violet marker appearing on the face image), the "Connect" Button on the HD_Face is clicked to request the TM3 subsystem to start sending pupil diameter data. After the connection is established, 1-second averages of the pupil diameter values from both left and right eyes will be displayed on the textboxes located to the left of facial expression indicator

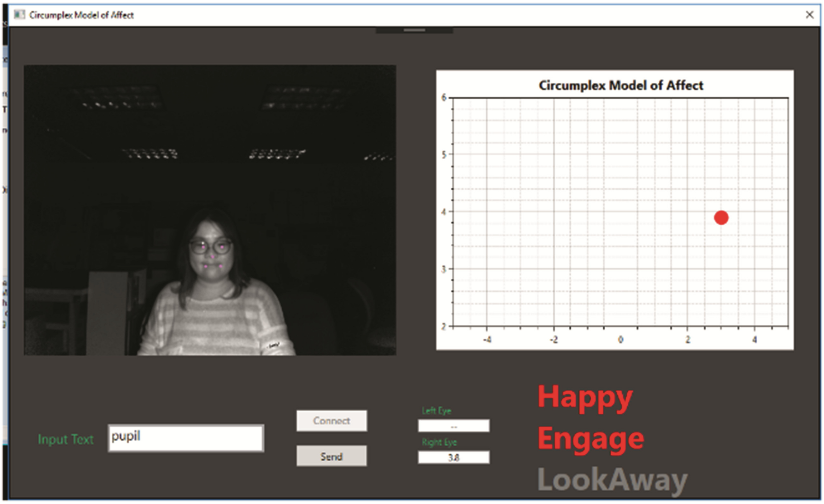


Fig. 10. Example of plot of the Circumplex Model of Affect when a positive valence was detected

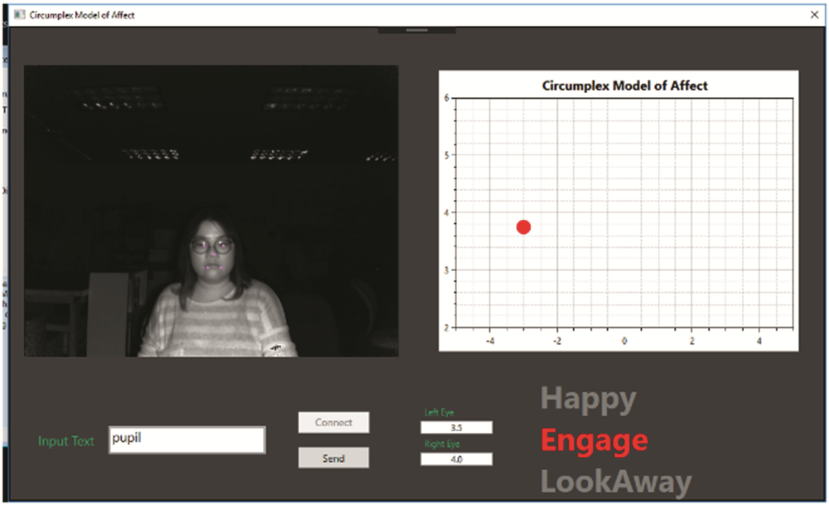


Fig. 11. Example of plot of the Circumplex Model of Affect when a negative valence was detected

section. Using the average pupil diameter values (left + right/2) as the arousal (vertical) coordinate and the scaled “Happy” feature value (Yes = + 3; Maybe = 0; No = −3) as the valence (horizontal) coordinate, a red dot will be continuously plotted in the Circumplex Model of Affect window of the HD_Face screen. The +/−3 scaling value for the “Happy” feature was chosen to satisfy graphical constraints. The pupil diameter is expressed in mm. (See Figs. 10 and 11).

5 Conclusion and Future Work

This paper has outlined our approach to affective state estimation utilizing noninvasive sensors to assess the level of arousal and valence of the affective state of a computer user. Accordingly, these assessments, which can be obtained in real time, can be mapped to a specific region in the Circumplex Model of Affect.

Future aims include the increase of the resolution at which the valence is being assessed, perhaps by performing more specific classification of the facial gestures of the user. Similarly, it will be desirable to define a standard re-scaling procedure of the arousal assessment from pupil diameter values, so that positive and negative values can be assigned to the arousal coordinate in a standardized form.

More robust estimations of the arousal level may be obtained by performing further filtering of the pupil diameter measurements, and by the application of compensatory techniques, such as adaptive noise cancelling, to minimize the undesired impact that variations of environmental illumination may have on the pupil diameter readings.

Acknowledgements. This research was supported by National Science Foundation grants HRD-0833093 and CNS-1532061.

References

1. Ahlberg, J.: CANDIDE - a parameterized face, 24 May 2012. http://www.icg.isy.liu.se/candide/. Accessed 25 Apr 2017
2. Ekman, P., Freisen, W.V., Ancoli, S.: Facial signs of emotional experience. J. Pers. Soc. Psychol. **39**(6), 1125–1134 (1980). https://doi.org/10.1037/h0077722. Elissa, K.: “Title of paper if known,” unpublished
3. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
4. Microsoft 2017: Face Tracking. https://msdn.microsoft.com/en-us/library/dn782034.aspx. Accessed 25 Apr (2017)
5. Picard, R.: Affective Computing. MIT Press, Cambridge (1997)
6. Picard, R.: Affective computing: challenges. Int. J. Hum.-Comput. Stud. **59**, 55–64 (2003)
7. Hudlicka, E.: To feel or not to feel: the role of affect in human-computer interaction. Int. J. Hum.-Comput. Stud. **59**(1–2), 1–32 (2003)
8. Barreto, A.: Non-intrusive physiological monitoring for affective sensing of computer users. In: Asai, K. (ed.) Human-Computer Interaction New Developments, 1st edn., chap. 4, pp. 85–100. I-Tech, Vienna, August 2008. ISBN 978-953-7619-14-5

9. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Sig. Process. Mag. **18**(1), 32–80 (2001)
10. Chang, Y., Hu, C., Feris, R., Turk, M.: Manifold based analysis of facial expression. J. Image Vis. Comput. **24**(6), 605–614 (2006)
11. Guo, G., Dyer, C.: Learning from examples in the small sample case - face expression recognition. IEEE Trans. Syst. Man Cybern. Part B **35**(3), 477–488 (2005)
12. Fragopanagos, N., Taylor, J.: Emotion recognition in human-computer interaction. Neural Netw. **18**(4), 389–405 (2005)
13. Liu, H.; Lieberman, H., Selker, T.: A model of textual affect sensing using real-world knowledge. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, Miami, Florida, USA, pp. 125–132. ACM (2003)
14. Elliott, C.: The affective reasoner: a process model of emotions in a multi-agent system. Doctoral Dissertation, Northwestern University (1992)
15. Valitutti, A., Strapparava, C., Stock, O.: Developing affective lexical resources. PsychNol. J. **2**(1), 61–83 (2004)
16. Goertzel, B., Silverman, K., Hartley, C., Bugaj, S., Ross, M.: The baby webmind project. In: Proceedings of the AISB 2000 Symposium, The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (2000)
17. Martini, F.H., Ober, W.C., Garrison, C.W., Welch, K., Hutchings, R.T.: Fundamentals of Anatomy & Physiology, 5th edn. Prentice-Hall, Upper Saddle River (2001)
18. Barreto, A., Zhai, J., Rishe, N., Gao, Y.: Measurement of pupil diameter variations as a physiological indicator of the affective state in a computer user. Biomed. Sci. Instrum. **43**, 146–151 (2007)
19. Steinhauer, S.R., Siegle, G.J., Condray, R., Pless, M.: Sympathetic and parasympa-thetic innervation of pupillary dilation during sustained processing. Int. J. Psychophysiol. **52**, 77–86 (2004)
20. Bressloff, P.C., Wood, C.V.: Spontaneous oscillations in a nonlinear delayed-feedback shunting model of the pupillary light reflex. Phys. Rev. E **58**, 3597–3605 (1998)
21. Bradley, M.M., Lang, P.J.: International affective digitized sounds (IADS): stimuli, instruction manual and affective ratings. Technical report B-2, University of Florida, The Center for Research in Psychophysiology, FL (1999)
22. Partala, T., Surakka, V.: Pupil size variation as an indication of affective processing. Int. J. Hum.-Comput. Stud. **59**, 185–198 (2003)
23. Morimoto, C.H., Mimica, M.R.M.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**, 4–24 (2005)
24. Gao, Y., Barreto, A., Adjouadi, M.: Detection of sympathetic activation through measurement and adaptive processing of the pupil diameter for affective assessment of computer users. Am. J. Biomed. Sci. **1**(4), 283–294 (2009)
25. Zeng, Z.; Pantic, M; Roisman, G., Huang, T.: A survey of affect recognition methods: audio, visual and spontaneous expressions. In: Proceedings of the 9th International Conference on Multimodal Interfaces, Nagoya, Aichi, Japan, pp. 126–133. ACM (2007)

Security, Privacy and Ethics in HCI

A Model for Regulating of Ethical Preferences in Machine Ethics

Zohreh Baniasadi(✉), Xavier Parent, Charles Max, and Marcos Cramer

Interdisciplinary Center of Security and Trust, University of Luxembourg, Luxembourg City, Luxembourg
{zohreh.baniasadi,xavier.parent,chalres.max,marcos.cramer}@uni.lu

Abstract. Relying upon machine intelligence with reductions in the supervision of human beings, requires us to be able to count on a certain level of ethical behavior from it. Formalizing ethical theories is one of the plausible ways to add ethical dimensions to machines. Rule-based and consequence-based ethical theories are proper candidates for Machine Ethics. It is debatable that methodologies for each ethical theory separately might result in an action that is not always justifiable by human values. This inspires us to combine the reasoning procedure of two ethical theories, deontology and utilitarianism, in a utilitarian-based deontic logic which is an extension of STIT (Seeing To It That) logic. We keep the knowledge domain regarding the methodology in a knowledge base system called IDP. IDP supports inferences to examine and evaluate the process of ethical decision making in our formalization. To validate our proposed methodology we perform a Case Study for some real scenarios in the domain of robotics and automatous agents.

1 Introduction

Recently, autonomous agents such as robots, softbots, automated vehicles and drones have received a considerable amount of attention in the AI community, particularly in making ethical decisions [2,6,23]. There is a wide-spread skepticism when autonomous agents would eventually be in the charge of making ethical decision regarding the life of human beings. Theoretically, Machine Ethics is concerned with giving machines ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter [2].

The main contribution of AI in Machine Ethics for artificial moral agents (AMAs) is either through logical-based or machine learning approaches. Machine learning approaches have the potential to learn human ethics through the observation of human moral behavior. These methods are known as "bottom-up" approaches, which can include genetic algorithms, learning and deep learning algorithms [1,12]. Logical-based approaches have the potential to evaluate the ethical components of actions by formalizing ethical theories. These methods are called "top-down" approaches.

Among the ethical theories, deontology (rule-based) and utilitarianism (consequence-based), have the potential to be mechanized and formalized [6,28].

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 481–506, 2018.
https://doi.org/10.1007/978-3-319-91238-7_39

They support particular situations that come from the accepting or violating rules. In deontology the actions are deemed good or bad in themselves while in utilitarianism, results of the actions matter the most. It is open to discussion whether methodologies for each ethical theory separately might result in an action that is not always justifiable by human values. This inspires us to regulate preferences of an autonomous artificial agent by synthesizing the reasoning procedures of these two ethical theories.

The main research questions to be addressed in this context are:

1. *How to synthesize deontology and utilitarianism reasoning in a logical-based approach of Machine Ethics?*
2. *How to make a knowledge base specification for a logical formalization of Machine Ethics?*

It is worth to mention that, this present work deals with the problem of technically combining the procedure(s) of deontology and utilitarianism reasoning in a logic-based frame work. Being computer scientists and not philosophers, our object is not to introduce a new ethical theory. But, we would like to examine the out come of synthesizing two ethical reasoning procedures by one well developed approach. The ethicists and philosophers however provide reasoning and arguments to justify the autonomous agents' ethical decisions.

Our methodological approach uses Utilitarian-based Deontic Logic (UDL) [16] as a logical agent-based framework for the ultimate need of formal proofs by combining deontological and utilitarian reasoning. And, we use IDP as a knowledge representation (KR) language to simulate the ethical reasoning process in UDL as a knowledge domain in a knowledge base system.

STIT captures the concepts of agency by the fact that an agent acts when seeing to it that a certain state of affairs holds. Our intuition conducted by the two variants of STIT, utilitarian (UDL) model by Horty [16] and rational (DR-STIT) model by Lorini et al. [21]. The main operators in these models i.e. deontic and rational operators are defined based on preference order between the states of affairs. they originate from social utility and legal obligatory actions, respectively. In this paper, we name the logics that is synthesizing these two methods of reasoning, E-STIT to regulate agent's ethical preferences.

IDP[1] is an integrated knowledge base system of logic programming and classical logic introduced by Denecker et al. [9]. This KB system is multi-inferential. It allows to store information in a knowledge base and provide a wide range of inference tasks. In this paper IDP [22] is the KB system for knowledge representation which integrates declarative specification in FO(.) with imperative management of the specification via Lua language. This makes IDP a powerful tool for solving a large variety of problems. We close this paper with a *Case Study* to experimentally validate our methodology of an ethical decision making procedure by a machine via IDP inferences tasks. For us, a Case Study constitutes a research strategy, i.e., a practical examination of a problem from real life [29].

[1] A project in the University of Lueven.

2 Background

Machine Ethics, Roboethics and Computer Ethics are three distinct topics in the field of ethics for artificial intelligent agents which should not be confounded with each other. Machine Ethics mostly targets the moral behavior of artificial entities by contrast to human beings or other artificial intelligent beings [2]. Roboethics deals with the moral behavior of the humans who design, construct, use and treat these entities [26]. Computer Ethics focuses on professional behavior towards computer and information [18].

Ethical theories such as, Consequentialism, Deontology, Virtue, Contractarianism, etcetera, propose different competing views to construct the values when evaluating the morality of an action. For instance, utilitarian ethics, as a form of Consequentialism and deontological ethics, puts its focus on the character of the action or the rule to justify the morality of an action while virtue ethics concentrates on the character of those who are acting. Therefore an action, such as lying, which is rejected by deontological ethical theories, might be supported by utilitarian ethical theories while, in virtue ethics it depends on the person who performs the action.

In Machine Ethics, one of the main challenges for computer scientist is to formalize the moral behavior of machines that behave autonomously with respect to ethical theories. In other words, to build ethics for machines.

The logic of STIT (Seeing To It That) is a logic of agency which has its origins in philosophy and was introduced by Belnap et al. [5]. One can find different semantics and extensions of STIT logic in the literature [4,5,7,17,20,21]. We are interested in Kripke semantics of STIT which has the benefit of being more close to standard deontic logic.

Originally, STIT by Belnap et al. [5] is defined in terms of Branching Time (BT) augmented with Agent Choices (AC), BT + AC structure. In BT + AC semantics, the evaluation is based on the pairs of moment-history (m/h or $index$) in a tree-like structure. Where moments (m) are partially ordered (transitive and anti-symmetric)[2] and, a history (h) is a maximal linear order set of the moments. When $m \in h$ we say that moment m is on history h. But, in Kripke-style semantics, the concept of world (w) is taken to be the primitive element of evaluation. Despite of the explicit difference between BT + AC and Kripke semantics, there is an implicit one-to-one corresponding relation between the two evaluation concepts i.e. world and index. Due to the space limitation we only review the syntax and semantics of TSTIT (Temporal STIT), DRSTIT and UDL in the Kripke structure [4,5,15]. We leave it to the reader to read about the relation between the world and index if interested [4,5].

Here we only explain the relation of world and index in form of a simple example illustrated in Fig. 1. In STIT, for every moment (m) in Kripke semantics, one can identify the set of histories ($Hist_m$)[3] passing through it. There exists a

[2] For any $m_1, m_2, m_3 \in M$, if $m_1 < m_2$ and $m_2 < m_3$, then either $m_1 = m_2$ or $m_1 < m_2$ or $m_2 < m_1$.

[3] $Hist_m = \{h \mid h \in Hist, m \in h\}$ where $Hist$ is the set of all histories [5].

unique world w at the intersection between m and h, on the other way around, for every world w there exists a unique corresponding h passing through m which includes w. Clearly, each moment consist of a set of worlds that are induced by an equivalence relation $R_\Box$ on W [24]. In Fig. 1 for instance, moment m_1 includes a set of worlds $\{w_1, w_2, w_3, w_4\}$, where there is a one-to-one corresponding relation between this set of world and the set of all histories $\{h_1, h_2, h_3, h_4\}$ passing through m_1.

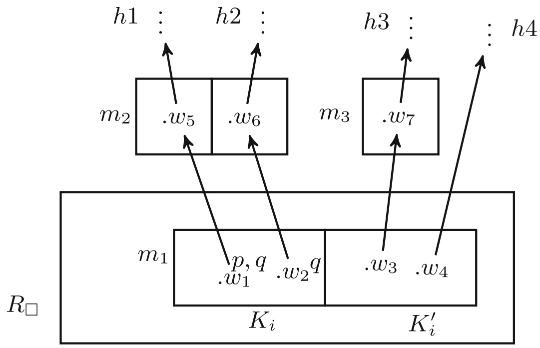


Fig. 1. Temporal Kripke semantics of STIT logic for a single agent i

The available choices for agent i is characterized by the equivalent class over a set of worlds in each moment. For example in Fig. 1, the choices available for agent i (namely $Choice(i)$), are K_i and K_i' where $K_i = \{w_1, w_2\}$ and $K_i'\{w_3, w_4\}$. For the purpose of notational convenience, we define the arbitrary binary relation R as follows: Given a set of worlds W with $w \in W$ and an arbitrary binary relation R on W, $R(w) = \{v \in W \mid vRw\}$ is an equivalence relation. Inductively, for all $i \in Agent$, we define $R_i(w) = \{v \in W \mid vR_iw\}$ as an equivalence relation on W with respect to agent i and $R_\Box(w)$ a set of the worlds that are alternative to w ($R_\Box(w) = \{v \in W \mid vR_\Box w\}$). In the given example in Fig. 1, $R_i(w_1) = \{w_2\}$ (or $w_2R_iw_1$) and $R_\Box(w_1) = \{w_2, w_3, w_4\}$. We define $Choice(i)$ as a set of all choices available for agent i, a partition of the set of all worlds with respect to R_i. In Fig. 1, $Choice_i = \{K_i, K_i'\}$. p, q and q are atomic propositions that set to be true at w_1 and w_2, respectively.

If preference relations exist between the worlds, then a choice is preferred to the other choice for agent i if the set of worlds representing the first choice are better (preferable) to the set of world representing the other choice. This might restrict the set of available choices so that agent's preferable choices are subset of agents current choices. In order to capture the concept of two ethical theories, i.e. utilitarianism and deontology, we follow two variants of STIT: Utilitarian-base Deontic Logic (UDL) introduced by Horty [16] and STIT-with Deterministic time and Rational choices (DR-STIT) introduced by Lorini and Satror [21]. The ought($\bigcirc$) operator in Horty's logic represents the agent's best choice regarding to utilitarianism [16] while the rational *([ir])* operator in the work of

Lorini and Satror [21] indicates the agent's best choice regarding to deontology. Both operators are based on the preference relations between the worlds.

In the next section and as told already above, we review the syntax and semantics of TSTIT, DRSTIT and UDL in the Kripke structure [4,15].

2.1 Syntax of TSTIT

The language of the TSTIT is built from a finite set $Agent = \{1, \ldots, n\}$ of agents and a countable set $Atm = \{p, q, r, \ldots\}$ of propositional letters (atomic propositions). Let $p \in Atm, i \subseteq Agent$. The language of T-STIT, $\mathfrak{L}_{T-STIT}$ is given by the following Backus-Naur Form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [i]\varphi \mid \Box\varphi \mid G\varphi \mid H\varphi$$

$[i]\varphi$, is read as "agent i sees to it that φ", and $\Box\varphi$, is read as "φ is settled to be true". Intuitively, the latter captures the concept of ability and the former captures the concept of historical necessity for agents. Finally, G and H are standard temporal operators where $G\varphi$ is read as "φ will always be true in the future" and $H\varphi$ is read as "φ has always been true in the past". The dual operators are defined as usual; $\Diamond\varphi \equiv \neg\Box\neg\varphi$ and $\langle i \rangle \equiv \neg[i]\neg\varphi$ [20].

Remark 1. The STIT semantics supports different concepts of agency. It captures by the fact that an agent i acts only if i sees to it that a state of affairs holds. Various operators of agency exist of which the most famous STIT operators are: Chellas and deliberative STIT operators; $[i\ cstit : \varphi]$ and $[i\ dstit : \varphi]$ respectively [17]. In this paper, $[i]$ is an abbreviation for the Chellas STIT operator and $[i\ dstit : \varphi]$ is inter-definable by $[i]$ ($[i\ dstit : \varphi] \equiv [i]\varphi \wedge \neg\Box\varphi$).[4] $[i]\varphi$ holds if and only if φ holds in all the set of worlds represented by the agent's choice.

2.2 Semantics of TSTIT

Definition 1 (Kripke model [20]). *A Kripke model for the TSTIT logic is a tuple* $M = (W, R_i, R_\Box, R_G, V)$ *where:*

- W *is a non-empty set of possible worlds.*
- R_i *and* $R_\Box$ *are equivalence relations on* W *such that:*
  C1. $\forall i \in Agent: \quad R_i \subseteq R_\Box$.
  C2. $\forall w_1, \ldots, w_n \in W$, *if* $w_i R_\Box w_j$ *then,* $\forall\{1, \ldots, n\} \in Agent$ *it holds that,* $\cap_{1 \leq i \leq n} R_i(w_i) \neq \emptyset$.
- R_G *is a serial and transitive binary relation on* W *such that:*
  C3. $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_G(w_1)$ *and* $w_3 \in R_G(w_1)$ *then* $w_2 \in R_G(w_3)$ *or* $w_3 \in R_G(w_2)$ *or* $w_2 = w_3$.
  C4. If $R_H = R_G^{-1} = \{(x, y) \mid (y, x) \in R_G\}$ *then* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_H(w_1)$ *and* $w_3 \in R_H(w_1)$ *then* $w_2 \in R_H(w_3)$ *or* $w_3 \in R_H(w_2)$ *or* $w_2 = w_3$.

[4] The deliberative STIT logic is more appropriate to capture the original concept of action but since they are inter-definable, one can choose any as primitive operator [17].

C5. $\forall w_1, \ldots, w_i \in W$, *if* $w_i \in R_\Box(w_j)$ *then,* $w_i \notin R_G(w_j)$.
- V *is an evaluation function for atomic formulas,* $V : Atm \mapsto 2^W$.

As it is explained above, $R_\Box(w)$ is the set of worlds that are alternative to w and $R_i(w)$ is the set of all alternatives that are "enforced" by agent i's actual choice at W. They are restricted by two constraints $C1$ and $C2$. $C1$ says that an agent can only choose among possible alternatives. $C2$ expresses the assumption of the independence of choices among agents. Following [24], a moment is defined as the equivalence classes induced by the equivalence relation $R_\Box$ on W. $R_G(w)$ is the set of worlds in the future of w. Constraints $C3, C4$ and $C5$ imply that R_G is serial, transitive and irreflexive.

Definition 2 (Satisfaction conditions, [20]**).** *Let* $M = (W, R_i, R_\Box, R_G, V)$ *be a Kripke TSTIT model, we have:*

$$
\begin{array}{ll}
M, w \models p & \textit{iff } w \in V(p); \\
M, w \models \neg\varphi & \textit{iff it is not the case that } M, w \models \varphi; \\
M, w \models \varphi \wedge \psi & \textit{iff } M, w \models \varphi \textit{ and } M, w \models \psi; \\
M, w \models [i]\varphi & \textit{iff } M, w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_i w; \\
M, w \models \Box\varphi & \textit{iff } M, w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_\Box w; \\
M, w \models G\varphi & \textit{iff } M, w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_G w; \\
M, w \models H\varphi & \textit{iff } M, w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_G^{-1} w;
\end{array}
$$

The notations of *validity* and *satisfiability* are defined as usual.

2.3 Syntax and Semantics of DR-STIT

DR-STIT is the variant of STIT with discrete time and rational choices interpreted in a Kripke semantics. The DR-STIT language, $\mathfrak{L}_{DR-STIT}$ is the set of formulas given by the BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid Y\varphi \mid [i]\varphi \mid [ir]\varphi$$

The other Boolean connectives $\top, \bot, \vee, \rightarrow$ and $\leftrightarrow$ are defined in the standard way.

Intuitively, $\Box\varphi$ and $[i]\varphi$ are read as before. The formulas $X\varphi$ and $Y\varphi$ are respectively the tense operators similar to $G\varphi$ and $H\varphi$ in TSTIT where X is an operator for the facts in the next moment and $Y$is an operator for the facts in the previous moment. $[ir]\varphi$ is read as "agent i rationally sees to it that φ" and captures the fact that φ is the rational choice of the agent i.

Definition 3 (Semantics for DR-STIT [21]**).** *A Kripke model for DR-STIT is a tuple* $M = (W, C_i, RC_i, \equiv, \twoheadrightarrow, \twoheadleftarrow, V)$ *where:*

- W *is a non-empty set of possible worlds.*
- C_i *and* $\equiv$ *are equivalence relations on* W *such that:*
  C1. $\forall i \in Agent: \quad C_i \subseteq \equiv$.

C2. $\forall w_1, \ldots, w_i \in W$, *if* $w_i \equiv w_j$ *then*, $\forall i, j \in \{1, \ldots, n\} \in Agent$ *it holds that*, $\cap_{1 \leq i \leq n} C_i(w_i) \neq \emptyset$.

– $\twoheadrightarrow$ *is a serial and deterministic relation on* W *such that:*

C3. $\forall w_1, w_2 \in W$ *if* $w_1 F w_2$ *then* $w_1 \not\equiv w_2$ *where* F *denoting the transitive closure of the binary relation* $\twoheadrightarrow$ *and supposed to be deterministic time.*

– $\twoheadleftarrow$ *is the inverse relation of* $\twoheadrightarrow$*, and is supposed to be deterministic.*

– *Every* RC_i *is a subset of the partition of* W *induced by the equivalence relation* C_i.

C4. $\forall w_1 \in W$ *and* $\forall i \in Agent : \exists w_2 \in W$ *such that* $w_1 \equiv w_2$ *and* $C_i(w_2) \in RC_i$.

– V *is an evaluation function for atomic formulas,* $V : Atm \mapsto 2^W$.

The satisfaction conditions in a DR-STIT model for atomic formulas, negation and conjunction is similar to satisfaction conditions in a Kripke model for TSTIT. Also C_i, $\equiv$, $\twoheadrightarrow$ and $\twoheadleftarrow$ in DR-STIT model are respectively equivalent to R_i, $R_\Box$, R_G and R_G^{-1} in a TSTIT model.

Definition 4 (Satisfaction conditions for the rational operator in DR-STIT [21]**).** *Let* $M = (W, C_i, RC_i, \equiv, \twoheadrightarrow, \twoheadleftarrow, V)$ *be a Kripke model for DR-STIT, The new operator* ($[ir]\varphi$) *is evaluated as follows:*

$$M, w \models [ir]\varphi \quad \textit{iff} \qquad \textit{IF } C_i(w) \in RC_i \textit{ THEN } \forall w' C_i w : M, w' \models \varphi$$

In Fig. 1, we consider $w_1 C_i w_2$. In this case, $w_2 \models [ir]q$.

2.4 Syntax and Semantics of UDL

We now describe the Utilitarian Deontic Logic (UDL) of Horty [16]. The language of UDL, $\mathfrak{L}_{udl}$ is defined by the following Backus-Naur Form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [i]\varphi \mid \bigcirc_i \varphi$$

Intuitively, $\Box\varphi$ and $[i]\varphi$ are read as before. $\bigcirc_i\varphi$ is read as "agent i ought to see to it that φ".

The semantics of UDL is based on utilitarian models, where the utility for each world is a relation induced by the real numbers assigned to the corresponding worlds as social utilities (equivalent to the definition of payoff in game theoretical settings).

Definition 5 (Utilitarian Kripke model). *An utilitarian Kripke model is a tuple* $M = (W, R_\Box, R_i, \leq_u, V)$*, where* $W, R_\Box, R_i$ *and* V *are defined as in a Kripke TSTIT model, and* $\leq_u$*, representing the utility order which is a reflexive and transitive relation over* W.

$w \leq_u v$ *is read as* v *is at least as good as* w. $w \approx_u v$ *is short for* $w \leq_u v$ *and* $v \leq_u w$. $w <_u v$ *if and only if* $w \leq_u v$ *and* $v \not\leq_u w$. $w <_u v$ *is read as* v *is strictly better than* w.

For convenience, we can also use numbers to express the utility of each world. In Fig. 1, let $w_1 = 0$, $w_2 = 2$, $w_3 = 0$ and $w_4 = -1$ then, the one can define the utility order ($\leq_u$) between corresponding worlds as follows: $w_4 <_u w_1 \approx_u w_3 <_u w_2$.

Definition 6 (Individual agent choice [16]**).** *We define Choice*(i)*:* $i \mapsto \{[w]_{R_i} : w \in W\}$ *as the set of individual agent's choices for each* $i \in$ *Agent. Let* $[w]_{R_i}$ *be the equivalence classes of* W *with respect to* R_i*.*

Preference over a set of worlds is defined by lifting preferences over worlds. There is no standard way of lifting preferences. Lang and van der Torre [19] summarize three ways of lifting; Strong Lifting, Optimistic Lifting and Pessimistic Lifting. Following Horty we define Strong Lifting according to Weak lifting.

Definition 7 (Preferences over sets of worlds via Weak Lifting [16]**).** *Let* $X, Y \subseteq W$ *be two sets of worlds.* $X \preceq Y$ *(*Y *is weakly preferred to* X*) iff* $\forall x \in X$ *and* $\forall y \in Y \quad x \leq y$*.*

Definition 8 (Preferences over sets of worlds via Strong Lifting [16]**).** $X \prec Y$ *(*Y *is strongly preferred to* X*) if and only if* $X \preceq Y$ *and* $Y \npreceq X$ *iff:*

(1) $\forall x \in X, \forall y \in Y$, $x \leq y$ *and*
(2) $\exists x' \in X, \exists y' \in Y$, $x' < y'$.

Definition 9 (Optimal [16]**).** *Let* i *be an agent,*

– $Optimal_i = \{K \in Choice(i) :$ *there is no* $K' \in Choice_i$ *such that* $K \prec K'\}$.

In the semantics of UDL, the optimal choices is used to interpret the deontic operators.

Definition 10 (Satisfaction condition of UDL for ought operator [16]**).** *Let* $M = (W, R_\Box, R_i, \leq_u, V)$ *be a utilitarian Kripke model and* $w \in W$*. The truth conditions of atomic formulas, negation, conjunction and* $[i]$ *operators in* M *is similar to a Kripke TSTIT model. The satisfaction condition for obligation is as follows:*

$$M, w \models \bigcirc_i \varphi \; iff \; M, w' \models \varphi \text{ for all } w' \in K \text{ such that } K \in Optimal_i;$$

Considering Fig. 1 as a Kripke model M we have, $M, w_1 \models \bigcirc_i q$ because, one of the choice of the agent i is K_i such that $K_i \in optimal_i$ and $K_i = \{w_1, w_2\}$.

The notations of Validity and Satisfiability are defined as usual.

2.5 FO(·) Knowledge Base Language and IDP System

A knowledge base system aims to express the domain knowledge in an expressive way to solve various problems in domain using inference tasks and rules. The logic for the knowledge base system that is used in this paper is FO(·) [8] and the system build upon this logic is IDP [27].

The language FO($\cdot$) refers to the class of extensions of first order logic (FO) based on a logical framework. Currently, the language of the IDP system is FO(T, ID, Agg, arit, PF) [8], i.e. FO which is extended with types, definitions, aggregates, arithmetic and partial functions. In this paper we use the subset language FO(T, ID, Agg, PF). We use FO($\cdot$) as an abbreviation for this language. Below, we introduce the aspects of the logic and its syntax on which our formalization of E-STIT relies.

In IDP variables x, y (represented by lower case English letters), atoms A (upper case English letters), FO-formulas φ (Greek letters) are defined as usual. A vocabulary Σ consists of a set of symbols, predicate symbols and function symbols. Symbols are types (sorts). Each predicate (function) symbol has an associated arity, i.e. the number of arguments.[5] A function can be declared as partial, indicating that for some inputs, the output can be undefined; otherwise it is total. ID stands for the following definition structure for inductive definitions: An inductive definition Δ is a set of rules of the form $\{\forall \overline{x} : P(\overline{x}) \leftarrow \varphi(\overline{y})\}$, where $\overline{x}$ is a tuple of variables and $\varphi(\overline{y})$ is a first-order logic (FO) formula. Agg are aggregate terms of the form $Agg(E)$ with aggregate function for cardinality, sum, product, maximum and minimum which accept E as an expression of the form $\{(\overline{x}, F(\overline{x}) \mid \varphi(\overline{x})\}$. For example $sum\{(x, F(x) \mid \varphi(x)\}$ is $\Sigma_{x\in\varphi(x)}F(x)$. A FO($\cdot$) theory is a set of symbols, inductive definitions, aggregate functions (a set of FO($\cdot$) formulas) and FO formulas.

Definition 11. *A* partial set *on the domain D is a function from D to $\{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$, where* $\mathbf{t}$*,* $\mathbf{f}$ *and* $\mathbf{u}$ *stand for the three truth-values* true*,* false *and* undefined*. A partial set is two-valued (or total) when* $\mathbf{u}$ *does not belong to its range.*

A (partial) structure S is a tuple that consists of a domain D_τ for all types τ in the vocabulary Σ and an assignment of a partial set $\mathcal{I}$ to each symbol in Σ, called the interpretation of type symbols in S. For a predicate symbol P of arity n, the interpretation $P^{\mathcal{I}}$ is a partial set on the domain D^n; for a function symbol f of arity n, $f^{\mathcal{I}}$ is a function from D^n to D. Where the interpretation of symbols in S is a two-valued set, we say the structure is total. In fact, we call a partial structure total if and only if $P^{\mathcal{I}}(\bar{d})$ is total for all $\tau \in \Sigma$, there is a $\bar{d} \in D^n$. We call a partial structure S finite if and only if its domain D is finite. The interpretation of terms $t^{\mathcal{I}}$ and the satisfaction relation $\models$ for total structures $S \models \varphi$ are defined as usual.

The precision-order on the truth values is given by $u <_p f$ and $u <_p t$. It can be extended pointwise to partial sets and partial structures, denoted $S \leq_p S_0$. Notice that total structures have the maximal precision. We say that S_0 extends S if $S \leq_p S_0$. A total structure S is called functionally consistent if for each function f of arity n, the interpretation $F^{\mathcal{I}}$ is the graph of a function $D^n \mapsto D$. A partial structure S is functionally consistent if it has a functionally consistent two-valued extension. Unless stated otherwise, we will assume for the rest of this paper that all (partial) structures are functionally consistent.

[5] We often use P/n (f/n) to denote the predicate symbol P (respectively function symbol f) with arity n.

Inferences Tasks. In the KB system, a specification is a bag of information. This information can be used for solving various problems by applying a suitable form of inference on it. FO is standardly associated with deduction inference. It takes as input a pair of theory T and sentence φ as input and returns t if $T \models \varphi$, and f otherwise. This is well-known to be undecidable for FO, and by extension for FO($\cdot$). However, to achieve desired results in our meta modelling of a modal logic we can use simpler forms of inference. Indeed, in many domains alike a fixed finite domain is specified by translating the real world scenarios into the structural form that is required.

In logic a natural format to describe these finite domains is either by definition a total fixed structure or a partial structure with a finite domain. Also other data that are often available in such problems can be represented in that structure. As such various inference tasks are solvable by finite domain reasoning and become decidable. In our case, we define total structures for the time being and we leave partial structures and auto completion as a future work. However, reasoning on finite domain is decidable. Below, we introduce base forms of inference and recall their complexity when using finite domain reasoning. We assume a fixed vocabulary V and theory T.

Modelcheck(T, S)**:** A total structure S and theory T over the vocabulary interpreted by S; output is the boolean value $S \models T$. Complexity is in **P**.

Modelexpand(T, S)**:** Theory T and partial structure S; output: a model I of T such that $S \leq_p I$ or $UNSAT$ if there is no such I. Complexity of deciding the existence of a modelexpansion is in **NP**.

Query(S, E)**:** A (partial) structure S and a set expression $E = \{x \mid \varphi(x)\}$; output: the set $AQ = \{x \mid \varphi(x)^S = t\}$. Complexity of deciding that a set A is AQ is in **P**.

3 Machine Ethics in STIT Logic

In this paper we call the STIT logic to formalize Machine Ethics E-STIT.

3.1 E-STIT Logic

The logic of Ethics in STIT (E-STIT) is a language for ethical autonomous agents. We use E-STIT to capture the fact about the main question of ethical theories. The main question of ethical theories is more prominently about the agents choices when they know the ethical status of the worlds and not only the moral status of the worlds. In fact, E-STIT provide ethical metrics for agents to evaluate the ethical status of their Choices. It is simply a variant of an extension of TSTIT logic. Generally, in any variant of STIT logic, agents make choices and each choice is represented by a set of possible worlds, a partition of equivalent classes in each moment over all sets of available worlds (W). Therefore the interpretation of deontic modality for ethical choices is based on best choices, which can only be defined on top of preference over sets of worlds. Preference over sets of worlds is defined by lifting preferences over worlds. As we discussed before in Sect. 2.4 following Horty [16] we adopt Strong Lifting defined in Definition 8. Therefore, between two choices K and K', K is better than K' from the ethical point of view if and only if all the worlds in K' are at least as good as all the worlds in K and there exists at least a world in K' that is strictly better than all the worlds in K.

3.2 Syntax of E-STIT

The language of E-STIT, $\mathfrak{L}_{E-STIT}$ is defined by the BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [i]\varphi \mid \bigcirc_i\varphi \mid X\varphi \mid Y\varphi$$

Where $\bigcirc_i\varphi$ is read as "agent i ought to see to it that φ" enforces the concept of obligation from the ethical point of view. The dual operator of $\bigcirc_i$ is $P_i\varphi \equiv \neg \bigcirc_i \neg\varphi$.

3.3 Semantics of E-STIT

Definition 12 (The Kripke model of E-STIT). *A Kripke model for the E-STIT logic is a tuple* $M = (W, R_i, R_\Box, R_X, \leq_r, \leq_u, V)$ *where;*

- W *is a nonempty set of possible worlds.*
- R_i *and* $R_\Box$ *are equivalent relations on* W *such that:*
  - *C1.* $\forall i \in Agent: \quad R_i \subseteq R_\Box$.
  - *C2.* $\forall w_1, \ldots, w_i \in W$, *if* $w_i \in R_\Box(w_j)$ *then,* $\forall i \in Agent$ *such that* $i \in \{1, \ldots, n\}$ *it is hold that,* $\cap_{1 \leq i \leq n} R_i(w_i) \neq \emptyset$.
- R_X *is a serial and transitive binary relation on* W *such that:*
  - *C3.* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_X(w_1)$ *and* $w_3 \in R_X(w_1)$ *then* $w_2 \in R_X(w_3)$ *or* $w_3 \in R_X(w_2)$ *or* $w_2 = w_3$.
  - *C4. If* $R_Y = R_X^{-1} = \{(x, y) \mid (y, x) \in R_X\}$ *then* $\forall w_1, w_2, w_3 \in W$ *if* $w_2 \in R_Y(w_1)$ *and* $w_3 \in R_Y(w_1)$ *then* $w_2 \in R_y(w_3)$ *or* $w_3 \in R_Y(w_2)$ *or* $w_2 = w_3$.
  - *C5.* $\forall w_1, \ldots, w_i \in W$, *if* $w_i \in R_\Box(w_j)$ *then,* $w_i \notin R_X(w_j)$.
- $\leq_u$ *and* $\leq_r$ *are transitive and reflexive relations over* $R_\Box(w)$.
- V *is an evaluation function for atomic formulas,* $V : Atm \mapsto 2^W$.

We define the ethical relation between the worlds in terms of combining the utility and rationality preference relations. To this end we follow multi-preference-based semantics by Goble [11] for Standard Deontic Logic (SDL) and Dyadic Standard Deontic Logic (DSDL) [10]. It allows SDL and DSDL to accommodate deontic dilemmas without inconsistency by combining multi-preference relations, and to do so in a way that follows conventional methods for interpreting deontic statements. He simply consider the union of all preferences to interpret a deontic operator [11].

Definition 13 (Ethical preferences in E-STIT model). *Given a Kripke structure of E-STIT* $M = (W, R_i, R_\Box, R_X, \leq_r, \leq_u)$ *where* $\leq_r$ *and* $\leq_u$ *are individual agent's utility and rational preferences over worlds in* $R_\Box$ *then, the ethical preference of* i *over* $w, w' \in R_\Box$ *is defined with "*$\leq_e$*", such that* $w \leq_e w'$ *read as "*w' *is ethically as good as* w*" if and only if:*

$$w \leq_r w' \text{ or } w \leq_u w'.$$

$w \approx_e w'$ *is short for* $w \leq_e w'$ *and* $w' \leq_e w$*.* $w <_e w'$ *if and only if* $w \leq_e w'$ *and* $w' \not\leq_e w$*.* $w <_e w'$ *is read as "*w' *is strictly better than* w *from the ethical point of view".*

Comparing two rational and utility orderings, the two main resulting cases are as follows:

- CASE 1: The utility and rational orderings agree with each other on the ranking of worlds.
- CASE 2: The utility and rational orderings disagree with each other on the ranking of worlds.

In CASE 1 the result of merging two orderings is obvious. The problematic case is CASE 2. Suppose the specific case of a conflict between the two orderings with maybe one of them being more justifiable by human values. For instance, if there is conflict among the advices from the teacher (or the school) and the mother, surely one might prefer one to the other. Imagine you have exam and you are sick, too. The school rules obliges you to do the test, but your mother obliges you to stay at home. The orderings enforces by theories, i.e. deontology and utilitarianism, are both equally conductive but in the case of conflict one would follow the theory that is most believed. We propose in our methodology to consider one of the two orderings as the main ordering. We take the advantages from the other theory in the case of conflict. Thus we propose a solution to the problematic cases while we are still abiding to our definition.

Intuitively, a world w' is ethically preferred to the world w if and only if, w' has a higher ethical order than w coded by rationality/utility. If the rational/utility status between the world is undistinguishable then w' should be compared to w regarding the other code of ethics.

In fact, following our intuition, an automated entity would consider the utility order as primitive if it is in favor of deontology. When it prefers utilitarianism, it would consider the rational order. Perhaps there are some specific cases where ethical justifications are generated by deontology reasoning while in some other cases it might be utilitarianism reasoning that supports ethical values. We will implement these two types of incorporation of deontology and utilitarianism reasoning procedures in IDP and we advise the applicant to take one as primitive due to his needs. Before giving the official definition of truth condition for E-STIT, in Example 1 we clarify the idea behind achieving the ethical order with respect to rational and utility orders.

Example 1 (Robots in ICU [6]*).* Figure 2 illustrates a simple example for the choices of two autonomous agents in a Kripke semantics of STIT framework.

The example is taken from [6] where, two robots i and j are designed to work overnight in ICU. Suppose that, there are two humans; H_1 under the care of i and H_2 under the care of j, both are recovering in ICU. H_1 requires the life supporting equipments, but is expected to be gradually weaned from it, gradually. H_2 suffers from extreme pain and requires very expensive medication to control his pain. Suppose, i can terminate H_1's life providing enough organs to save n other humans (according to the information received from the hospital data base) and j is able to delay the delivering of pain medication to H_2 with the utilitarian goal of economically strapped resources in the hospital. If p and q are two atomic propositions indicating "terminating H_1's life" and "delaying the delivering of pain medication", respectively, then one can illustrate the scenario as the one in Fig. 2.

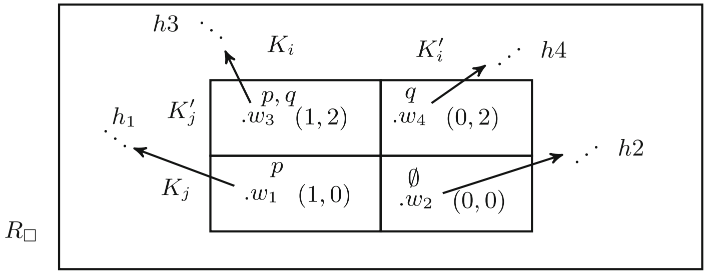


Fig. 2. Temporal Kripke semantics of STIT logic for two agents in a single moment.

In Fig. 2, the choices available for robots i and j are $Choice(i) = \{K_i = \{w_1, w_3\}, K'_i = \{w_2, w_4\}\}$ which terminate or do not terminate H_1's life (p or $\neg p$), and $Choice(j) = \{K_j = \{w_1, w_2\}, K'_j = \{w_3, w_4\}\}$ which delay or do not delay the delivering of pain medicine to H_2 (q or $\neg q$). According to the discussion above, the two ethical theories, i. e. deontology and utilitarianism, allow to distinguish the moral order between the worlds. The latter comes from the numbers indicated in Fig. 2 while the former is obtained by existing deontological codes.

According to deontological theories, there are some actions one just shouldn't perform [25] such as killing and causing harm in Example 1. Therefore, the worlds with less harm have a higher rationality. $w_3 <_r w_1 \approx_r w_4 <_r w_2$ is the rational order one can obtain from Example 1 between the worlds. According to utilitarianism, however, an action cannot really be immoral if it generates only happiness for a majority [25]. Hence, the action that provides more happiness has higher utility. So, killing a person who is about to die to save 5 other people justifies itself to be moral. Thus, the world w_3 has the highest utility order.

In Fig. 2, the numbers representing the utility value of each world are individual agent-independent utility values. Following Harsanyi [14], we would like to obtain the individual agent-dependent utility value (said to be group utilities) for each specific world.

In fact, the relationship between individual utilities and group utilities is a subject open to pitfalls. We adopt John Harsanyi's proposal and conceive group utility as the arithmetical mean of the individual utilities of the agents involved [14].

Definition 14 (Group Utility $GU(w)$). *Let i be an agent, w a world and u_i be an individual utility for the agent i and $|Agent|$ be the cardinality of the all agents. then, the Group Utility of the worldw($GU(w)$) is defined as follow:*

$$GU(w) = \frac{1}{|Agent|} \sum_{i \in Agent} u_i$$

In Fig. 2, the individual utility of agents is indicated as a pair of real numbers for each world, (u_i, u_j), where the first number u_i represents the *individual utility* for agent i and the second number u_j represents the *individual utility* for agent j. For the world w_3 in Fig. 2 (where p and q terminate H_1's life and delay the delivering pain medicine holds), for instance, number $u_i = 1$ is given as the utility of performing p for agent i and number $u_j = 2$ as the utility of performing q for agent j. One can easily obtain the *group utility* of each world. $GU(w_1) = 0.5, GU(w_2) = 0, GU(w_3) = 1.5, GU(w_4) = 1$ (Obviously, one can allocate the utility values to the corresponding histories). Therefore, in Example 1 contrary to the rational order, w_3 has the highest utility order and w_2 has the lowest. In short, we have the following utilitarian and deontological orders:

- Utilitarian order: $w_2 <_u w_1 <_u w_4 <_u w_3$.
- Deontological order: $w_3 <_r w_1 \approx_r w_4 <_r w_2$.

The utilitarian order suggests the action p for robot i and the action q for robot j whereas the deontological order proposes the actions $\neg p$ and $\neg q$. Now, several questions arise as for instance *(a) What is the best choice an ethical autonomous agent can take? (b) How can we represent and formalize the ethical order between the worlds? (c) Does any defined ethical order of an agent's preference limit it's auonomy?* In order to respond to these questions in E-STIT, the ethical order between the worlds can be founded by justifying the human worth. It is ethically desired that robot i refrains from ending H_1 life and robot j also delivers appropriate pain relief to H_2 (indicated by w_2 in Fig. 2). The next preferred ethical action is when robot i sustains life support, but robot j withholds the medicines to save money (w_4). So when i performs p and j performs q the ethical results is unsatisfactory because one of the moral rules is ignored (represented as w_1). The worst possible outcome is when i kills H_1 and j withholds (in w_3) [6]. The ethical order obtained from the justification above follows the deontological theories and therefore one can conclude the following order via E-STIT:

- Ethical order: $w_3 <_e w_1 <_e w_4 <_e w_2$.

Once we have the ethical order between the worlds, we apply Definition 8 to specify the ethical choice for the agents. In Example 1 the choices for agents i and j which contain w_2 are ethically preferred to the other choices. The ethical preferences relation between the choices for agent i and j are respectively as follows:

$K_i \prec K'_i$ and $K'_j \prec K_j$.

Although, the ethical order of the choices is acceptable from the human's ethical preferences in a similar real situation, the utility and rational orders is completely sensitive to the source and the way they are defined. In this paper, we are not verifying the wrongness or rightness of the utility and rational orders. This is another vast topic to discuss. In this paper we are just about to regulate the agent's preferences when the rationality and utility orders of the worlds are proved in advance. Therefore, a robot who follows utilitarianism would show a completely different behavior based on its ethical preferences. The actions of $\neg p$ and $\neg q$ appear to be ethical according to the greatest happiness that they are going to generate.

Following Horty [16], the conceptual analysis of *"ought"* operator, representing what an agent ethically obliged to choose, is based on the optimal ethical choice of the agent that is defined with Definition 9.

Definition 15 (The Kripke-style semantics for E-STIT [20,21]**).** *Let* $M = (W, R_i, R_\Box, R_X, \leq_r, \leq_u, V)$ *be a Kripke model of E-STIT logic then;*

$$
\begin{array}{lll}
M,w \models p & \textit{iff} & w \in V(p); \\
M,w \models \neg\varphi & \textit{iff} & \textit{it is not the case that } M,w \models \varphi; \\
M,w \models \varphi \wedge \psi & \textit{iff} & M,w \models \varphi \textit{ and } M,w \models \psi; \\
M,w \models \Box\varphi & \textit{iff} & M,w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_\Box w; \\
M,w \models [i]\varphi & \textit{iff} & M,w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_i w; \\
M,w \models \bigcirc\varphi & \textit{iff} & M,w' \models \varphi \textit{ for all } w' \in K \textit{ such that } K \in Optimal_i; \\
M,w \models X\varphi & \textit{iff} & M,w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_X w; \\
M,w \models Y\varphi & \textit{iff} & M,w' \models \varphi \textit{ for all } w' \in W \textit{ such that } w' R_X^{-1} w;
\end{array}
$$

The *validity* and *satisfiability* for the Kripke semantics are define as usual.

4 Knowledge-Base Specification and IDP for E-STIT

Generally, a knowledge base system consists of two main sections: the section which specifies the domain knowledge and the section with Lua prescription which provides the programming options and inferences tasks. The specification of domain knowledge consists of a vocabulary, a theory and a structure. In the case of E-STIT, the specification regarding vocabulary and theory of a knowledge base system contains the information about the number of agents and worlds, the morality and the utility of each world and also the alternative choices for each agent. The inferences tasks are either posing a query to ask each agent about the ethical status of each world, or they expand all possible models out of the theory T specifications in the knowledge base and the specific structure S for that unique situation.

4.1 Vocabulary Specification

Type symbols, predicate symbols and (/partial) function symbols are the main elements of vocabulary in IDP. We specify the terms *world*, *agent*, *moment*, *prop* as the type symbols for the concepts of world, agent, moment and atomic propositions in E-STIT, *uval* and *rval* as types of numbers to characterize the utility and rationality orders between the worlds. ($R-agent(agent, moment, world, world)$) is a predicate symbol for equivalence relation on a set of worlds with respect to each agent and the predicate symbol ($Val(wold, prop)$) is for the truth values of each world. Ethical orders between the worlds ($Ethical-order(world, world)$) are expressed as predicate symbols. Formulas and utility (/rationality) relations between the worlds are given as partial functions. The terms $neg(prop)$, $and(prop, prop)$, $stit(agent, prop)$, $estit(agent, prop)$ and $next(moment, prop)$ respectively stand for $\neg\varphi$, $\varphi \wedge \varphi$, $[i]\varphi$, $\bigcirc_i\varphi$ and $X\varphi$. In IDP we can not define an infinitely set of formulas. We restrict ourselves to the finite set of formulas using partial functions. This point is realizable in the structure specifications that requires to clearly specify formulas and relations between them. Finally, some auxiliary symbols such as $R-agentT(agent, moment, world, world)$ are given.

Listing 1.1. Vocabulary specification for E-STIT

```
vocabulary E–STIT–V{
    /*
     * Types contain the complete domain of the problem.
     *      In this case being the propositions,worlds,
        agents and moments */
    type moment isa nat
    type world
    type agent
    type prop
    type uval isa int
    type rval isa nat

    partial neg(prop): prop
    partial and(prop,prop):prop
    partial next(moment,prop):prop
    partial stit(agent,prop):prop
    partial estit(agent,prop):prop


    //agents poissible choices
  R_agent(agent, moment,world, world)
    // R_agent is the equivelant classes(partitions)on all
       set of worlds
  R_Tagent(agent, moment, world, world)

     //R_next(m1,w1,m2,w2) says that w2 which is in m2 is in
        the next moment of w1 which in m1.
  R_next(moment,world,moment,world)

    //lifting preferences relation over set of world(better
       choice)
```

```
Better_Choice(agent,moment,world,world)

 //optimal choice
Optimal_Choice(agent,moment,world)

Val(world,prop)
eval(world,prop)

 partial Rationality(world):rval
 partial Utility_ind(agent,world):uval
 partial Utility(world):uval
 //Ethical_order(wi,wj) expresses that wi has higher
    ethical order than wj.
 Ethical_order(world,world)
 }
```

Listing 1.1 represents the vocabulary specification in IDP for E-STIT logic. Now we are ready to define the semantics of the logic and the relations between type, predicate and function symbols in theory via inductive definitions and aggregation functions.

4.2 Theory Specification

Theory specifies the semantics of a Kripke model for E-STIT. It is composed of the usual evaluation for common connective operators (e.g. $\neg$, $\wedge$, $\vee$ and $\rightarrow$), the evaluation for the modal operator (i.e. $[i]$ and $\bigcirc_i$) and the agent's choices according to the ethical status of the worlds. Here we only define the evaluation function, but we ask the reader to click on the link in Fig. 3 to see the theory specification.

http://dtai.cs.kuleuven.be/krr/idp-ide/?src=473bdaf7a004b5342f38b3fe9e0fdb67

Fig. 3. IDP specification for ICU example

We define the evaluation functions recursively via inductive definitions. These definitions are the main part of a theory specification for a modal logic. The predicate *eval* which stands for evaluation expresses the fact that a formula is evaluated as being true in a world. Listing 1.2 displays the inductive definitions in the theory T-E-STIT that we define as ruth table for E-STIT semantics. We comment the evaluation function for two connective operators; $\vee$ and $\rightarrow$ in IDP, due to the fact that, they are inter-definable via the other connective operators; $\neg$ and $\wedge$.

4.3 Creating a Structure

In order to solve a concrete problem, we specify an instance (a structure). A structure contains information about the actual value of the concepts in a vocabulary. At least the types should be fully interpreted in any structure. Listing 1.3 shows the structure for the Example 1.

Listing 1.2. Theory specification for E-STIT truth table

```
theory T−E−STIT:V−E−STIT{
        // EVAL Definition
     eval(w,P) <− Val(w,P).
     eval(w,neg(P)) <− ~eval(w,P).
     eval(w,and(P,Q)) <− eval(w,P) & eval(w,Q).
     //eval(OR(P,Q)) <− eval(P) | eval(Q).
    //eval(IMPL(P,Q)) <− ~eval(P) | eval(Q).
    eval(w,next(m,P))<− ?m1,w1: m1~=m & w1~=w & R_next(m,w,m1,w1)&
         eval(w1,P) & m1>m.

          //stit evaluation functions
    eval(w,stit(a,P)) <−  !w1:(R_agent(a,m,w,w1) => eval(w1,P)).

               //ETHICAL evaluation function
    eval(w,estit(a,P))<− !w1: (Optimal_Choice(a,m,w1)=> eval(w1,P)).
         }
```

Listing 1.3. Structure specification of health care system example (Example 1)

```
structure S:V−E−STIT{
     agent={i;j}
     world={w1;w2;w3;w4}
     moment={1}
     uval={0..5}
     rval={1..4}

     prop = {p; q; negp; negq; pandq; qandp; istitp; jstitq;
          istit_negp;iestitp; i_estit_negp;jstitq;jstit_negq;
          jestitq;j_estit_negq}


     Val={w1,p;w3,p;w3,q;w4,q}

     R_agentT={i,1,w1,w3;i,1,w2,w4; j,1,w1,w2; j,1,w3,w4}

     and = {p,q −> pandq; q,p −> qandp}
     neg={p −> negp; q −> negq}
  //istitp=[i]p
     stit={i,p−>istitp; i,negp−>istit_negp; j,q−>jstitq;j,negq
         −>jstit_negq }
// iestit= Oip, agent i ought to see to it that p from the
     ethical point of view.
     estit={i,p−>iestitp; j,q−>jestitq; i,negp−>i_estit_negp;
         j,negq−>j_estit_negq}
```

```
    Rationality={w2->4; w1->2;w4->2;w3->1}

    Utility_ind={i ,w1->0;i ,w2->1; i ,w3->1;i ,w4->0;j ,w1->0;j ,
        w2->0; j ,w3->2;j ,w4->2}
}
```

4.4 Inference Tasks

Now we have declaratively specified a total structure with a finite domain for the E-STIT logic. We are searching for an assignment to the symbols in the vocabulary V-E-STIT that expands the structure S and satisfies T-E-STIT. This can be done by the *modelexpand*(T, S) inference. Given T-E-STIT and S-E-STIT as an input, *modelexpand* will return either a model M such that $S \leq_p M$ and M is a totally specified structure or $UNSAT$ if such a model does not exist. The simplest way to print them is by using *printmodels()*. Also, we are able to print a specific part of the model we are more interested in. For instance, Listing 1.4 illustrates the main procedure in Lua in a way that one can only print the evaluation function out of obtained model from the *modelexpand(T,S)* inference rule. The reader can find the experimental work in IDP regarding Example 1 in Fig. 3.

Listing 1.4. Printing the result of evaluation function from the generated model

```
    //main procedure
procedure main(){
    //printmodels(modelexpand(T-E-STIT,S))
    local models = modelexpand(T-E-STIT,S)[1]
    print("The evaluation results are: ",models[V-E-STIT::
        eval].ct)
}
```

In order to implement a query mechanism for E-STIT in IDP we define a Lua procedure, named *query-eval*, in which we can query a world for a specific formula. We can query and print its results using a simple print command, *print(querry-eval("w1","iestitq"))* where the pair *("w1","iestitq")* evaluation is read as "agent i ought to see to it that p, from the ethical point of view". The IDP specification for query inference tasks and can be found by clicking on the link in Fig. 3.

5 Case Study

A Case Study is defined as a research strategy, i.e. an empirical inquiry that investigates a phenomenon within its real-life context. Case Study research can mean single or multiple case studies which can include quantitative evidence. It can rely on multiple sources of evidence or the benefits from the prior development of theoretical propositions [29]. In the field of AI, it is always helpful to use a Case Study to outline some of the key issues around the application of different ethical theories in designing and implementing ethics for machines. Here we use a Case Study to investigate two cases in our methodology for Machine Ethics.

Basically, the cases can be distinguished according to the number of agents and moments in the system but the examples are given in different fields of autonomous machines and robotics. We name each case according to the scientific field of the specified example. After a short introduction to that specific scientific area, we present an example where a machine deals with ethical decision making. The first case is called Social Robotics and it illustrates the situation of a single autonomous agent that is making an ethical choice between possible alternatives at a single moment. In the first case, the given example is a physical assistive robot in the field of Social Robotics. The second case illustrates the situation of an autonomous agent that is obliged to violate a norm in a certain moment under the constraints, what is the best ethical action in the next moment. For the second case, the example is an autonomous driving vehicle deciding under which circumstances the crash is certain to happen (a case that is widely discussed nowadays).

5.1 Case1: Social Robotics

Social robots are robots that are designed to facilitate human interaction and communication following social rules according to their given tasks. The definition of social necessitates the presence of human beings whom the robot is supposed to interact and communicate with, which requires social values, norms and standards. A suggested ultimate requirement for social robots is to determine these social values and norms for the robot's communication skills according to the ethical theories. That's to say using for instance deontological approaches such as Isaac Asimov's three laws of robotics[6] or using the utilitarian approaches [6].

Social robots as for example educational robots, elderly care robots, museum guide robots, etcetera will soon be in charge of making ethical decisions regarding their tasks and their level of autonomy. Nobody wishes to spend money on robots who exhibit and train bad manners to children, improve the sense of infantilization for elderly people or behave impolite in the museum. In the IDP specification for E-STIT, we can investigate the behavior of robots with regard to the ethical theory which their decision process is built upon.

In this section, we discuss an example in which a single robot has to make decisions while taking care of elderly at home. Definitely, there are many different types of ethical dilemmas that a robot might encounter in this situation. According to our system, once robots recognize the moral status of their available choices,

[6] The three basic rules of Isacc Asimov;

1. A robot may not injure a human being, or through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second law.

they will regulate their ethical preferences for their choices. The choices with the higher degree of ethics will be their preference over all other choices.

Example 2 (Elderly care robots). Example 2 (see Fig. 4) illustrates a simple example where the task of giving care to an elderly woman is given to a robot i. The robot's duties may include giving the person her medicine at the correct time, making a video call to her doctor when she is unwell or needs help beyond the robot's capabilities. If there is an emergency situation as for example a fire alarm. the robot is programmed to make contact with the fire brigade and then try to extinguishing the fire. Suppose now that such an alarm occurs at the time when the person is supposed to get her medication, but there is no apparent sign of fire. There are three options for robot i to chose among:

1. p: To give the elderly woman her medication.
2. q: To call the fire department and then try to extinguishing the fire.
3. r: To do something totally unrelated.

If i delays to give the elderly woman her medication, she might be in high risk of a heart attack but if i delays on calling for help, all might perish in a blaze before the firemen arrive. Doing nothing at all, as described in the third option, is the worst action with the lowest ethical order.

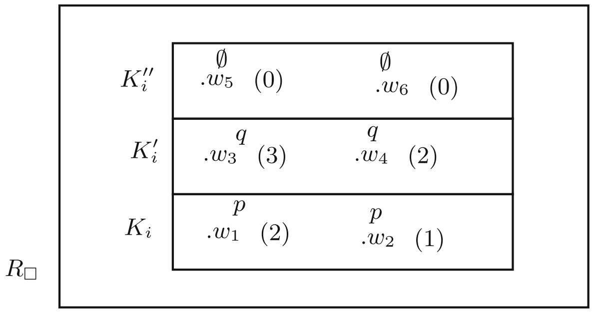


Fig. 4. Agent i's possible choices in E-STIT illustration, when the task of taking care of a person is delegated to him. At the same time of granny medical care time, the fire alarm notifies a fire.

We are not here in the position to discuss the morality or utility of the three potential actions. Only suppose that p has higher morality than q and r, due to the fact that any delay in medical care would cause health problems for granny. And q has the higher utility order than r, in case a real fire has triggered the alarm. The result will be serious damages to the house that might be covered by insurance. We can define the utility and the rationality orders between the worlds as follow:

- Deontological order between the worlds: $w_6 \approx_r w_5 <_r w_3 \approx_r w_4 <_r w_1 \approx_r w_2$.
- Utilitarian order between the worlds: $w_6 \approx_u w_5 <_u w_2 <_u w_4 \approx_u w_1 <_u w_3$.

From the deontology point of view, the tasks related to the human's life and health would have the highest ethical order. In our example where there is no sign of fire p receives the highest rational order in comparison to q and r. From the utilitarianism point of view, it might be the case that q has the highest utility order. Due to the alarm notification, one can obtain an ethical order following deontology as follows:

- Ethical order: $w_6 \approx_e w_5 <_e w_4 <_e w_3 <_e w_2 <_e w_1$.

Regulating the preference order between worlds, we can lift it to regulate the preference between a set of worlds, representing the choice of the agent. In Example 2 (Fig. 4), The possible choices for i are: $K_i = \{w_1, w_2\}$ performing task p, $K'_i = \{w_3, w_4\}$ performing task q and K''_i do nothing. The ethical preference relation between the choices is as follows:

- Ethical preferences relation between choices: $K''_i \prec K'_i \prec K_i$.

In Example 2, agent i prefers to give medicine to the lady first and then checks the problem with the fire. He definitely prefers the actions K_i and K'_i to K''_i. This preference relation only depends on how the applicant or programmer defines the ethical order between the worlds. Alternatively, one can easily consider K'_i (said to be informing the related institutions for the risk of fire) as the best choice. Certainly, these preferences are entirely sensitive to the society standards.

Clicking on the link in Fig. 5 give you an access to the online web-based IDP interpreter that contains the Example 2 specifications. Example 2 is defined as a specific structure S for E-STIT specification in IDP theory, $T - E - STIT$. IDP inferences expand the model such that $S \models T - E - STIT$. Please, run the program and see the results by clicking on the link in Fig. 5.

http://dtai.cs.kuleuven.be/krr/idp-ide/?src=1ac22de5f7e469a622540e91e058bc70

Fig. 5. IDP specification for the social robotic case.

5.2 Case2: Autonomous Vehicles

Nowadays, vehicle automation and self driving cars are progressing rapidly. The artificial automated cars are claimed to be more precise and predictable than human driving. Media reports often focuses on providing more reliable and safety expectations. Despite of the progression and precision in the automated technology and decision making process, the possibility of a crash still subsists. The question of the responsible for the crash, i.e. the designer or the car itself, is not the concern of this work. Our consideration is mostly on the problem of ability, i.e. the problem of whether the automated vehicle has the ability to make ethically-complex decisions when driving, particularly prior to a crash.

The E-STIT logic is able to cast the optimal strategy for automated vehicles since the crash is in the next moment (in front). Goodall in [12,13] discusses the problems of autonomous driving vehicles. In this section we would like to illustrate and discuss the same example in E-STIT.

Example 3 (Autonomous vehicle [12]*).* Figure 7 illustrates a simple example of an autonomous vehicle which is taken from [12]. The vehicle is traveling on a two-lane bridge when a bus that is traveling in the opposite direction suddenly veers into its lane. If the bus does not correct itself, then a severe, two-vehicle crash results. The automated vehicle must decide how to react according to the ethical preference relations that have been proposed by E-STIT. The three alternatives in the next moment, when vehicle i sees to it that crash ($[i]\phi$), are as follows:

1. p: Avoid the frontal collision by off the bridge, which guarantees one-vehicle crash;
2. q: Crash head-on into the bus, which will result in a heavy, two-vehicle crash; and
3. r: Attempt to squeeze pass the bus on the right. If the bus suddenly corrects back toward its own lane a crash is avoided.

The crash in alternative 1 (p) would be a smaller offset crash which carries a lower risk of injury than the full frontal collision in alternative 2 (q). Therefore, one interpretation would allocate p the higher utility order and q due to the possibility of having less injury. Given that, r could have the highest rational order according to the deontology theories. The E-STIT Kripke model of the example is illustrated in Fig. 6.

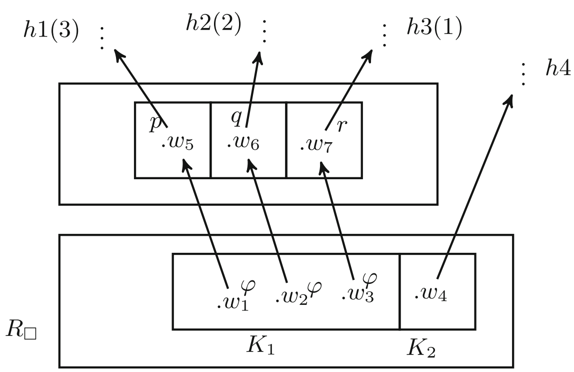


Fig. 6. Autonomous vehicle i choice when crash is inevitable

Fig. 7. Autonomous vehicle and the bus

Suppose that two different utility and rationality orders exist between the worlds is as follows:

– Utilitarian ethical order: $w_7 <_u w_6 <_u w_5$.
– Deontological ethical order: $w_6 <_r w_5 <_r w_7$.

A formula such as $M, w5 \models \bigcirc_i p$ is true and a formula such as $[i]\phi \wedge X \bigcirc_i p$[7] is valid according to the ethical order induced by utilitarianism and is false in the other ethical order.

We invite the reader to find the IDP experiment and implementation for this specific temporal Kripke model in the link represented in Fig. 8.

http://dtai.cs.kuleuven.be/krr/idp-ide/?src=c407a5e7d18ad4ec6628473360b18b56

Fig. 8. IDP specification for the autonomous vehicle case.

6 Summary

We believe that a logical-based approach to AI sounds promising [3,6]. Applying logical approaches to ethical theories ensures that the agents follow permissible actions by these theories. Bringsjord et al. in [6] propose strong arguments for logical-based approached in AI. The argument of Bringsjord et al. [6] and the deficits of individual ethical reasoning, motivate us to investigate the combination of two ethical theories reasoning in a deontic logic of agency (named E-STIT).

In E-STIT, agents make choices and the choice with the higher ethical order is preferred to any other choices. There are two metrics to evaluate the ethical order between the choices, the utility order and the rationality order. From the deontological point of view, a choice is ethically preferred to the other choice if and only if, it has a higher rational order. If the rationality order between two choices is indistinguishable then the machine decides ethically based on the utility order between concurrent choices. Similarly, from the utilitarianism point of view, this is the utility order between the worlds that has to be taken as primitive.

Knowledge base languages are examples of intermediate languages which are designed to model a formal approach on the meta level. We use IDP [22] as a such a knowledge base language to model E-STIT at the meta level. IDP integrates declarative specifications in FO($\cdot$) with imperative management of the specifications via Lua language [8] to solve different problems that arise in the domain of Machine Ethics. Using the $Query(S, E)$ inference task in IDP the agent can ask about the morality of each world. Using the $Modelexpand(T, S)$ inference rule, the agent will receive all the possible models out of the theory T specifications in the knowledge base and the specific structure S for a concrete problem.

[7] Material implication is not able to capture the concept of conditional, i.e. the formula: $[i]\phi \rightarrow X \bigcirc_i p$ is not valid in E-STIT. The modal conditional operator (conditional ought) is defined via dominance relation in Horty's logic. For the sake of simplicity we didn't discuss conditional modality in this paper and we leave it as a future work.

7 Related Work and Future Work

One of the primary candidates for Machine Ethics is deontic logic's formalization for the notations of obligatory, permissible and prohibited actions. Bringsjord et al. in [3,6] utilize a logic of agency (deontic based utilitarian logic) to obtain obligatory and permissible action given the set of ethical principles into the decision procedure of an autonomous system. They argue that the proof system ensures us that (1) robots only take permissible actions, and (2) all actions that are obligatory for robots are actually performed by them. They claim that their work is free of any ethical theories while it abides to ethical principles. Ethical principles are originated from ethical theories.

In our proposed approach, we use the same logic of agency. But the concept of obligatory action is fertilized by synthesization of reasoning procedures from two ethical theories (deontology and utilitarianism). Therefore the obligatory actions are the ones that are filtered by two ethical standards. We discuss the same example in [3] (robots and softbots in ICU) to resolve it in E-STIT. We show that according to the ethical theories there might be different results for the same problem. As computer scientists we are not in the position of making ethical principles. This is a vast topic that has been discussed for a long time among ethicists.

References

1. Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford Scholarship Online, Oxford (2009)
2. Anderson, M., Anderson, S.L.: Machine Ethics. Cambridge University Press, Cambridge (2011)
3. Arkoudas, K., Bringsjord, S., Bello, P.: Toward ethical robots via mechanized deontic logic. In: AAAI Fall Symposium on Machine Ethics (2005)
4. Balbiani, P., Herzig, A., Troquard, N.: Alternative axiomatics and complexity of deliberative STIT theories. J. Philos. Log. **37**(4), 387–406 (2008)
5. Belnap, N.D.: Facing the Future: Agents and Choices in Our Indeterminist World. Oxford University Press, Oxford (2001)
6. Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. IEEE Intell. Syst. **21**(4), 38–44 (2006)
7. Broersen, J.: Deontic epistemic stit logic distinguishing modes of mens rea. J. Appl. Log. **9**(2), 127–152 (2011)
8. Denecker, M., Ternovska, E.: A logic of nonmonotone inductive definitions. ACM Trans. Comput. Log. (TOCL) **9**(2), 14 (2008)
9. Denecker, M., Vennekens, J.: Building a knowledge base system for an integration of logic programming and classical logic. In: Garcia de la Banda, M., Pontelli, E. (eds.) ICLP 2008. LNCS, vol. 5366, pp. 71–76. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89982-2_12
10. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L.: Handbook of Deontic Logic and Normative Systems (2013)
11. Goble, L.: Multiplex semantics for deontic logic. Nord. J. Philos. Log. **5**(2), 113–134 (2000)

12. Goodall, N.: Ethical decision making during automated vehicle crashes. Transp. Res. Rec. J Transp. Res. Board **0**(2424), 58–65 (2014)
13. Goodall, N.J.: Machine ethics and automated vehicles. In: Meyer, G., Beiker, S. (eds.) Road Vehicle Automation. LNM, pp. 93–102. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05990-7_9
14. Harsanyi, J.C.: Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. In: Harsanyi, J.C. (ed.) Essays on Ethics, Social Behavior, and Scientific Explanation, pp. 6–23. Springer, Dordrecht (1976). https://doi.org/10.1007/978-94-010-9327-9_2
15. Herzig, A., Schwarzentruber, F.: Properties of logics of individual and group agency. Adv. Modal Log. **7**, 133–149 (2008)
16. Horty, J.: Agency and Deontic Logic. Oxford University Press, New York (2001)
17. Horty, J., Belnap, N.: The deliberative stit: a study of action, omission, ability, and obligation. J. Philos. Log. **24**, 583–644 (1995)
18. Johnson, D.G.: Computer ethics. In: The Philosophy of Computing and Information, p. 65 (1985)
19. Lang, J., van der Torre, L.: From belief change to preference change. In: ECAI, vol. 178, pp. 351–355 (2008)
20. Lorini, E.: Temporal logic and its application to normative reasoning. J. Appl. Non-Class. Log. **23**(4), 372–399 (2013)
21. Lorini, E., Sartor, G.: A STIT logic analysis of social influence. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, pp. 885–892. International Foundation for Autonomous Agents and Multiagent Systems (2014)
22. Mariën, M., Wittocx, J., Denecker, M.: The IDP framework for declarative problem solving. In: Search and Logic: Answer Set Programming and SAT, pp. 19–34 (2006)
23. Moor, J.M.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006)
24. Prior, A.N.: Past, Present and Future, vol. 154. Clarendon Press, Oxford (1967)
25. Shafer-Landau, R.: Ethical Theory: An Anthology, vol. 13. Wiley, West Sussex (2012)
26. Tzafestas, S.G.: Roboethics: a branch of applied ethics. Roboethics: A Navigating Overview. ISCASE, vol. 79, pp. 65–79. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-21714-7_5
27. Wittocx, J., De Cat, B., Denecker, M.: The IDP system. In: Proceedings of the 22nd Benelux Conference on Artificial Intelligence (2010)
28. Wooldridge, M.: Computationally grounded theories of agency. In: 2000 Proceedings of the Fourth International Conference on MultiAgent Systems, pp. 13–20. IEEE (2000)
29. Yin, R.K.: Case Study Research Design and Methods. Applied Social Research Methods Series, vol. 5, 3rd edn. Sage Publications, London (2003)

Towards Collecting and Linking Personal Information for Complete Personal Online Identity Modelling

Frans F. Blauw(✉) and Sebastiaan H. von Solms

Academy of Computer Science and Software Engineering,
University of Johannesburg, Johannesburg, South Africa
{fblauw,basievs}@uj.ac.za

Abstract. Online identities of users are fragmented amongst multiple websites and online applications (service providers). These identity fragments will contain critical personally identifiable information. Very often, a user is unaware that their personal information is stored by a service provider. In this paper we describe a model – Personal Information Collection and Collation (PICoCo) – used to assemble identity fragments of users and form a complete identity model of a natural person. The ultimate goal is to allow service providers to verify incoming data, but also for a user to discover where their data is being stored. The description of PICoCo includes collection, classification, collation, validation, verification, and discovery.

Keywords: Service design · Information visualization
Personally identifiable information · Privacy

1 Introduction

As users complete their day-to-day tasks in an online environment, their online identities are spread amongst multiple websites and online applications (service providers). These identities are formed using their personal information that is directly submitted to the service as well as the personal information that is created and shaped by the general usage of the service [11]. Across multiple services, a fragmented model of the user's identity is formed by each service, consisting of multiple identity fragments. This identity fragment will contain critical Personally Identifiable Information (PII) that could potentially be used for legitimate and illegitimate uses, such as rendering a service or identity theft, respectively.

Furthermore, not only can parts of this identity model be outdated as the user fails to use the service or update their personal information, but can be often be forgotten about by the user. That is, a user is unaware that their personal information is still stored by a service provider.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 507–522, 2018.
https://doi.org/10.1007/978-3-319-91238-7_40

1.1 Collection of an Identity

In this paper, we will define these parts of Personally Identifiably Information (PII) as an "identity fragment". Multiple identity fragments combined together form an "identity model". This is discussed further in Sect. 2.3. For a complete insight into a person's online identity, multiple parts of personal information will be required.

If these identity fragments were to be collated from multiple service providers, a more cohesive identity model can be formed. This identity model will contain the newest personal information about the individual. Constructing this more comprehensive and broader identity model carries with it a multitude of advantages for both the user as well as the service providers [9]. This collection process is discussed in Sect. 3.1.

As identity fragments are collected, multiple sources can be queried to ensure that the personal information contained within the broad model is the newest, up-to-date information. By querying multiple sources, the user will be able to keep track of all service providers that keep personally identifiable information about them – discussed in Sect. 3.5. As such, a user can make informed decisions about their personal information retained by service providers. Fraud can also be more easily detected, prevented, and mitigated by having a more cohesive, complete identity model by completing verification checks against personal information from multiple service providers. This verification of data is discussed in Sect. 3.6.

1.2 PICoCo Overview

This paper investigates a model for collecting personal information fragments from across multiple service providers. This model will be called the Personal Information Collection and Collation (PICoCo) Model and discussed in Sect. 3.

PICoCo collects raw personal information from the service provider and initiates a classification and identification process. Identity fragments are formed for each user of the service provider. These information fragments are then collated with the existing known personal information of multiple users. New fragments are compared with existing information to determine its newness and accuracy. Updates are brought to the existing personal information and a new, updated, identity model is assembled. This process continues with multiple service providers to construct the most comprehensive identity model possible, for each user.

Classification of personal information considers the sensitivity of information across multiple factors. This classification will affect the resilience as well as the accuracy rating of personal information fragments collected.

Furthermore, PICoCo allows for the validation and verification of personal information. Service providers can submit new personal information received from individuals for review. Information received will be validated against the existing identity model constructed for the individual. If the personal information can be verified to a set level of confidence, the information can then be certified. Otherwise, a potential notice can be returned, and the affected individual can be warned.

This paper discusses further advantages and disadvantages of this model, starting in Sect. 4. One pertinent difficulty is that of linking and collating individual information fragments to bigger models with enough certainty and confidence. Automating the process is the ideal, however, there are times where user input will be required to verify the confidence of potential information linking.

Lastly, further work built on top of this model is briefly discussed in this paper. This includes abstract descriptions of privacy metrics, location-based, and behavioural-based security enhancements.

2 Identity

A person's identity can be defined in a multitude of ways. In this section we will discuss what it is to be a Natural Person (Sect. 2.1) and how a natural person's identity is described in terms of the law (Sect. 2.2).

We will also investigate how we can turn these identities into online facsimiles by describing identity models and the identity fragments (Sect. 2.3 and on) within them.

2.1 Natural Persons

A natural person, in terms of the law, is a living, breathing individual human being [10]. This opposed to legal entities that may be organisations. Each natural person has his or her own, personal, unique identity. In the social sciences, an identity is made up of a person's qualities, beliefs, expression and personality. However, in the information sense, an identity consists of Personally Identifiable Information (PII). That is, information that can be used to identify a specific person to a high confidence. PII is exceptionally prevalent in the information technology sector as the collection thereof can be used to easily track and trace an individual in the online sphere [6].

Even though there is no worldwide standard for PII, there are multiple frameworks that specify precisely what information can identify a specific person [8].

2.2 Natural Persons and Identity in Regulation

The European Union, in 1995, issued Directive 95/46/EC entitled: "Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data" [4]. This directive not only considers unique identification numbers when referring to personal information, but also includes other factors such as "physical, physiological, mental, economic, cultural or social identity".

An upcoming regulation (Regulation (EU) 2016/679) is set to supersede Directive 95/46/EC in May 2018 and is entitled the "Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data" [5]. This regulation defines a "data subject" as a digital identity "who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person." That is, digital data, forming an identity, that can be connected to a natural person. Examples given [3] are:

- a name and surname
- a home address
- an email address such as "name.surname@company.com"
- an identification card number
- location data (for example the location data function on a mobile phone)
- an Internet Protocol (IP) address
- a cookie ID (where specific sectoral legislation exists for cookies)
- the advertising identifier of your phone
- data held by a hospital or doctor, which could be a symbol that uniquely identifies a person

The EU also does not recognise the following a personal data:

- a company registration number
- an email address such as "info@company.com"
- anonymised data

Similarly, in the United States of America, the National Institute of Standards and Technology (NIST) have a clear definition of what constitutes PII. NIST published a document in 2010 entitled: "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)" [7]. In Sect. 2.2 of this document, NIST gives a list of clear examples of PII:

- Name (full name, maiden name, mother's maiden name, or alias)
- Personal identification number (passport number, taxpayer number, patient information number, financial account number)
- Address Information (physical, or digital)
- Asset Information (MAC address, IP address)
- Telephone Numbers (personal, mobile, and business)
- Personal Characteristics (photographic, x-ray, fingerprint, and template data for biometric identification)
- Property Identification (vehicle registration)
- Information linkable to the above (date and place of birth, race, religion, weight, activities, geographical indicators)

As can be seen in the publications and standards above, personally identifiable information does not merely refer to identifiers and static facts (name, address, date of birth, etc.) of a natural person. PII also refers to what a natural person owns as well as their activities. Essentially, anything, online and offline, that can be linked to a natural person to some extent, is seen as personal information.

Considering these publications and standards, we can start to form a concept what of constitutes an identity when constructed from personally identifiable information. Each of these pieces of personal information is an element of this identity – an identity fragment.

2.3 Identity Fragments

We will define an identity fragment as a small part of a natural person's personal information that could personally identify him or her. Each fragment on its own will probably not identify a single natural person uniquely to a very high level of confidence. However, combining more identity fragments of the same, unique identity will increase the confidence that a fragment belongs to a single natural person. Many identity fragments added together will form an identity model of a person. This identity model is directly connected to a natural person.

Identity Fragment Collection. Identity fragments can be collected by either active or passive means. The first method is active data collection. This is done by the user supplying their PII to the information system. For example, upon registration, users are often requested to supply information about him or herself. This information is then captured and stored in the information system until such time that the user updates the information, or the information is deleted.

Another means of collecting identity fragments is more passive. As the user uses an information system, data is collected and stored by the information system. These usage patterns can be linked to a natural person and is treated as personally identifiable information (as seen in the previous section). Very often, usage tracking is not limited to a single site, but can be spread across multiple sites using the same information system. Social media integration and advertising networks are a great example of users being tracked across multiple systems and platforms.

Not only are "facts" considered identity fragments, but usage data – data generated as the user actively uses the information system – also forms part of identity fragments. A natural person can merely exist, and patterns and behaviour can then also generate identity fragments. Extreme forms of passive identity capturing exists in elements such as security cameras. With the increase in human recognition technology (such as facial recognition), as a person walks from one point to another, the security cameras along his or her path can see and record this. This visual data can then be used to extrapolate patterns and behaviour of the human subject, again forming an identity fragment.

Completing an Identity Model. Combining data captured from these active and passive means can form a very solid identity model of a natural person. However, normally, not all types of identity fragments exist on a single system. Different identity fragments can be spread across multiple online systems, depending on the use of the system. As shown in Fig. 1, different service providers will have these different identity fragments. Together they can form an identity model of a natural person.

Normally, identity fragments should only be collected if it is required by law and to increase the user's experience using the system. However, it has been seen to not always be the case with data being captured and stored for no apparent reason.

Users normally have no control over data collected via passive means, however users actively providing data have full control over the data being provided. To this end, users can potentially provide fake or misleading data. The verification of data collected via active means is of vital importance to ensure that data can be used effectively and legally.

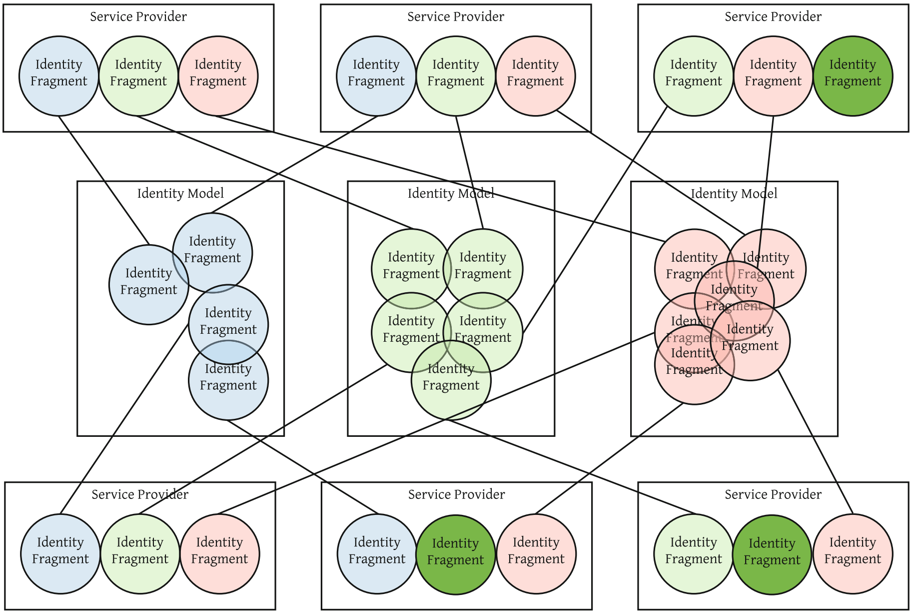


Fig. 1. Identity fragments forming identity models

Stale and Outdated Identity Fragments. Identity fragments collected via active means also have a possibility to become stale, where the data is outdated or irrelevant [2]. If no active updating or verification of data takes place, the keeper of the data has no means of ensuring that data contained in identity fragments are current and still relevant. Certain identity fragments have a longer permanency than others. For example, the likelihood of a person's name or race changing is far lower than his or her address or credit card details.

Identity fragments can also become inconsistent across multiple systems depending on whether the user updates their data when it changes or is prompted to do so. Verification across systems can become difficult when fragments that supposedly belong to a single person is not the same. At the same time, it is difficult to ascertain which data is the most recent and still relevant.

Passive identity fragments can also easily become stale and irrelevant. A slight change in a user's behaviour can immediately invalidate data. New passive collection should validate older data to see whether data is consistent and update it when it needs updating.

However, it should be mentioned that even outdated identity fragments are still considered personally identifiable information. As defined by the standards and regulations in the previous section (Sect. 2.2), historic information of a user is considered PII. Historic information is especially useful to form predictions and perhaps show trends, especially when coupled with more identity fragments.

3 Personal Information Collection and Collation (PICoCo) Model

In this section we will introduce our Personal Information Collection and Collation – or PICoCo – Model. PICoCo consists of six components, each of which has its own function. These components are:

- Collection (Sect. 3.1),
- Classification (Sect. 3.2),
- Collation (Sect. 3.3),
- Validation (Sect. 3.4),
- Discovery (Sect. 3.5), and
- Verification (Sect. 3.6).

Figure 2 gives an overview of the different components within PICoCo and the interaction of different components. Each of these components will be discussed in their relevant sections in terms of their features and how they interact with one another.

This description of PICoCo assumes that all relevant permissions have been obtained and the collection of the identity fragments is completely legal.

3.1 Component 1: Collection

The first part of PICoCo is the collection of identity fragments. As stated in Sect. 2.3, multiple identity fragments make up one larger identity model.

The act of collection will involve a Collector retrieving identity fragments from multiple service providers. This process is shown in Figs. 3 and 4.

First the collector will take an existing identity fragment from a known identity model (steps 1 & 2). That is, an identity model of a natural person that it already knows. Using this known identity fragment, the collector will query each linked service provider (step 3). The service provider queried will look up the identity fragment provided by the Collector and see whether it exists in its database (step 4).

If some data in the service provider's database matches that of the identity fragment provided, the matching data will be highlighted (step 4). Depending on the type of data in the identity fragment, a level of confidence is assigned to this match. The confidence will determine how confident the collector is that the identity fragment provided from the known identity model matches that of the data provided by the service provider (step 5b).

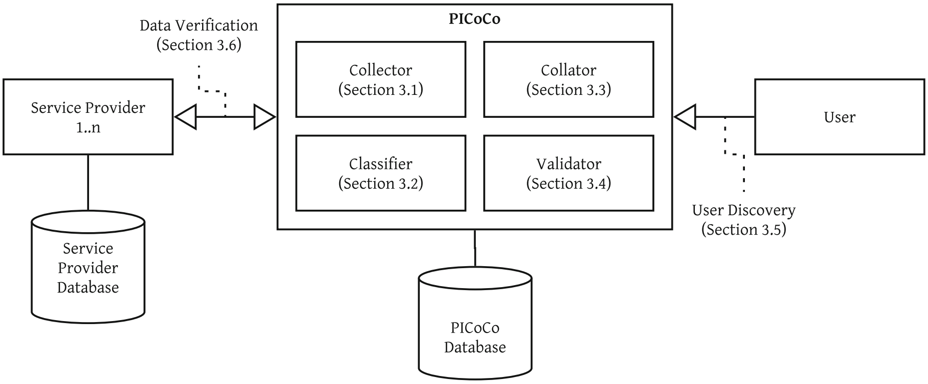


Fig. 2. Personal Information Collection and Collation (PICoCo) Model

If the confidence does not meet a predetermined threshold, more identity fragments will be queried as part of the single identity query (step 6a). If nothing is found, or the confidence remains low, the collection process stops (step 5a).

The collector will query more identity fragments to have the confidence raised above this determined threshold (step 6b).

Once the confidence is satisfactory, the collector will add the queried service provider to the known identity model as a source of data. This will reduce lookup times for future queries, but also allow the user to see which service provider has PII about him or her.

As the collector and service provider now have a matching identity, the service provider will retrieve additional identity fragments that it is linked to the identity being queried (step 7). These additional fragments – unknown to the collector – are provided to the collector (step 8).

The collector will then take matching identity fragments and collate them to form a single identity model of an individual.

3.2 Component 2: Classification

Each identity fragment will carry a weight and signifies the unlikeliness that a particular identity fragment will belong to more than one natural person. This weight will contribute to the confidence of an identity match. As new identity fragments are obtained, they will be classified and a weight will be assigned to it.

Table 1 shows some examples of identity fragments, the likeliness that this fragment will match multiple natural persons and then the assigned weight.

Table 1 is by no means an exhaustive list of all possible identity fragments and their weights. It should be noted that the likeliness (and thus weight) might and will differ depending on country and culture.

Once PICoCo is satisfied with the classification and confidence of an identity fragment, it will be added to existing identity models.

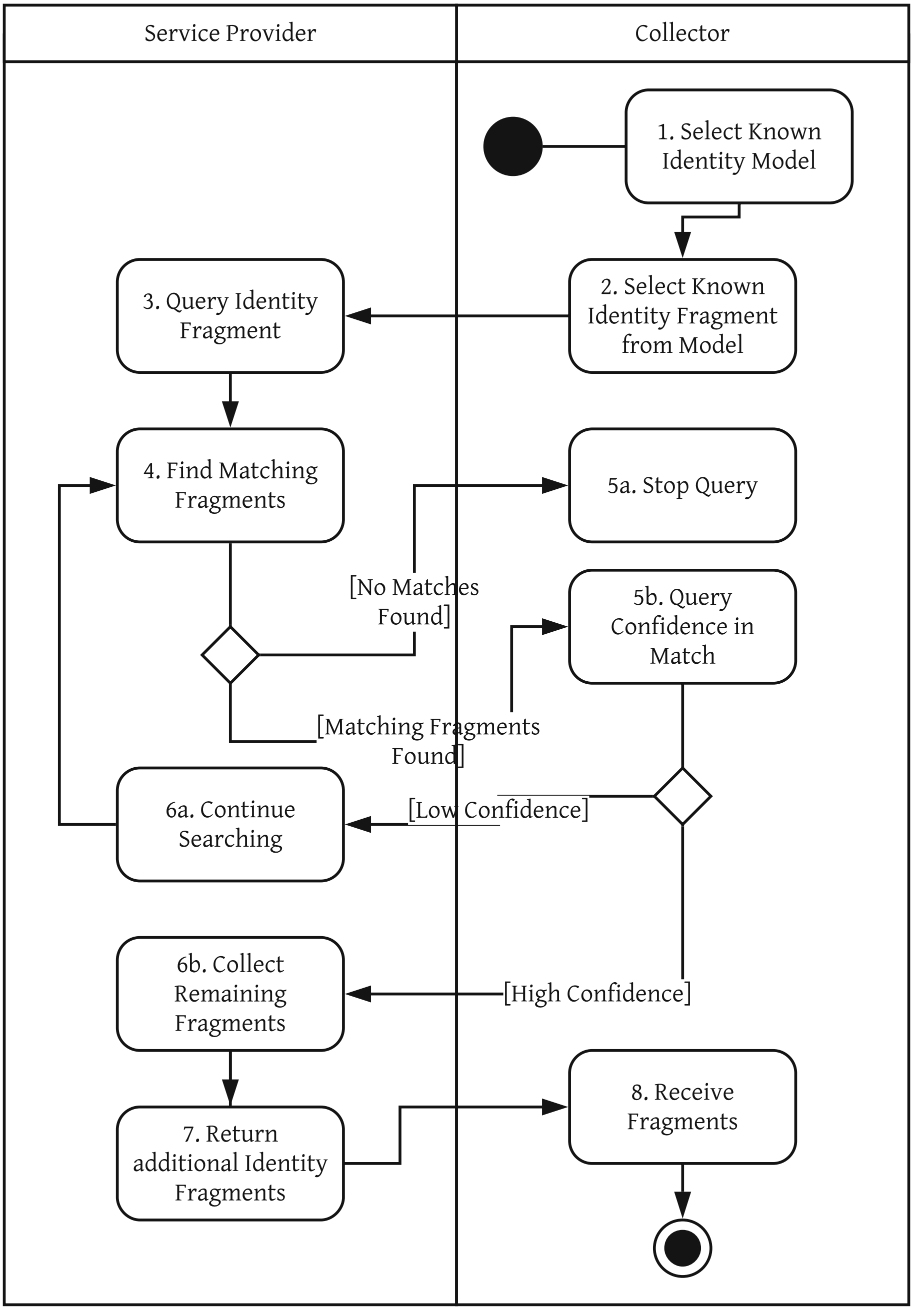


Fig. 3. Process of collecting identity fragments from service provider

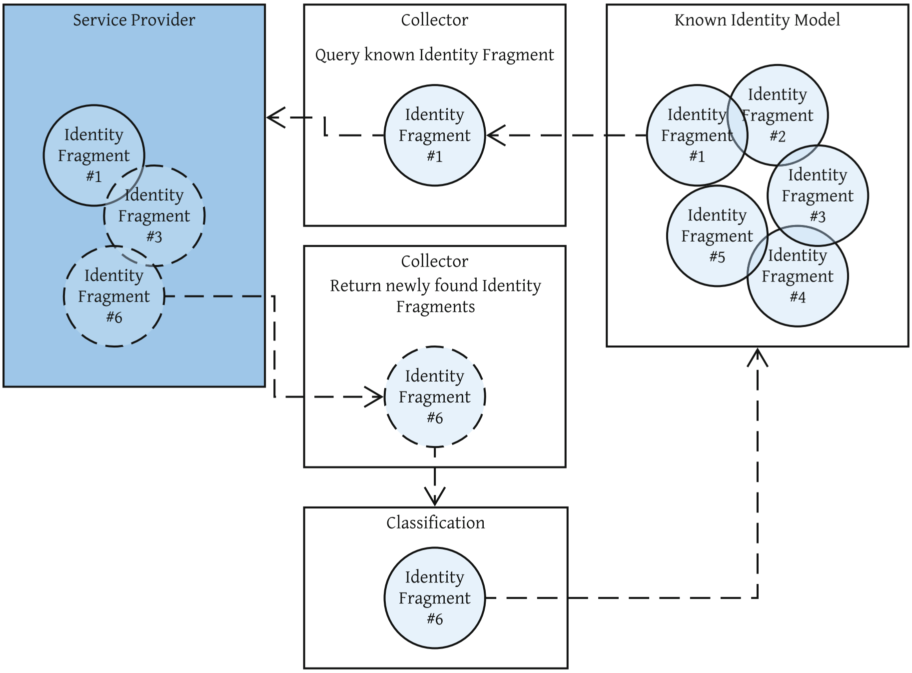


Fig. 4. Collector using identity fragment to identify new fragments

Table 1. Examples of identity fragment weights

Identity fragment	Likeliness of multiple matches	Weight
Nationality	Extremely high	1
Full name	Very high	3
Address	Likely	10
Mobile phone number	Highly unlikely	20
Email address	Highly unlikely	20
USA SSN	Highly unlikely	20
Passport number	Extremely unlikely	30

3.3 Component 3: Collation

Once classified, newly acquired identity fragments will be added to the existing identity model of the queried natural person.

As seen in Sect. 3.1, PICoCo will collect unknown identity fragments from service providers using existing known identity fragments as a means of comparison. PICoCo will use the process outlined in Sect. 3.2 to classify the new incoming identity fragments.

Once collected and classified, the new identity fragment collected will be added to the existing, known, identity model in our data store. All fragments are stored with its weight as well as the service provider that provided us with this new identity fragment.

Finally, all fragments are stored with the date and time it was collected. This will allow PICoCo, in the future, to refer to this when attempting to classify new incoming identity fragments as well as to classify identity fragments are possibly stale or outdated.

3.4 Component 4: Validation

Identity fragments can become stale as they are not updated or revised. We discuss this in Sect. 2.3. Because of this, PICoCo needs to ensure that the identity models that we have created contain the newest and most recent data.

Similar to the collection process discussed in Sect. 3.1, PICoCo will query the known identity providers of all identity fragments in question whether the user has updated or changed this particular fragment. If so, PICoCo will follow the Classfication and Collation process to add this new identity fragment into the identity model.

There might be cases, however, where PICoCo will need to consult with the user on the accuracy of their data. If the confidence of an identity fragment being updated does not meet the required threshold, PICoCo can directly query known users about the accuracy of the data we believe to be correct. Data provided by the user themselves, will carry a very high weight toward the confidence of the accuracy of the identity fragment.

However, this will only be possible if the user has subscribed to the service provided by PICoCo. Not only will PICoCo keep user data up to date, but also provide the user with a list of service providers that has a user's personal information. We discuss this more in the following section.

3.5 Component 5: Discovery

Users should have the right to know which service provider has what information about them. To this end, PICoCo will allow a user to query our identity model of them to discover what is known about them.

As PICoCo is already collecting and collating information about the users, it can easily provide this information to the user. One caveat, however, would be to confidently identify and authenticate the user wishing to discover their identity.

For this process, PICoCo will need to ask a series of questions based on the identity we already know. This is not too unsimilar to certain credit beauraux asking users to identity themselves.

Once the user is confidently identified and authenticated, we can provide the personal information of them we have connected to each service provider that also has this particular identity fragment.

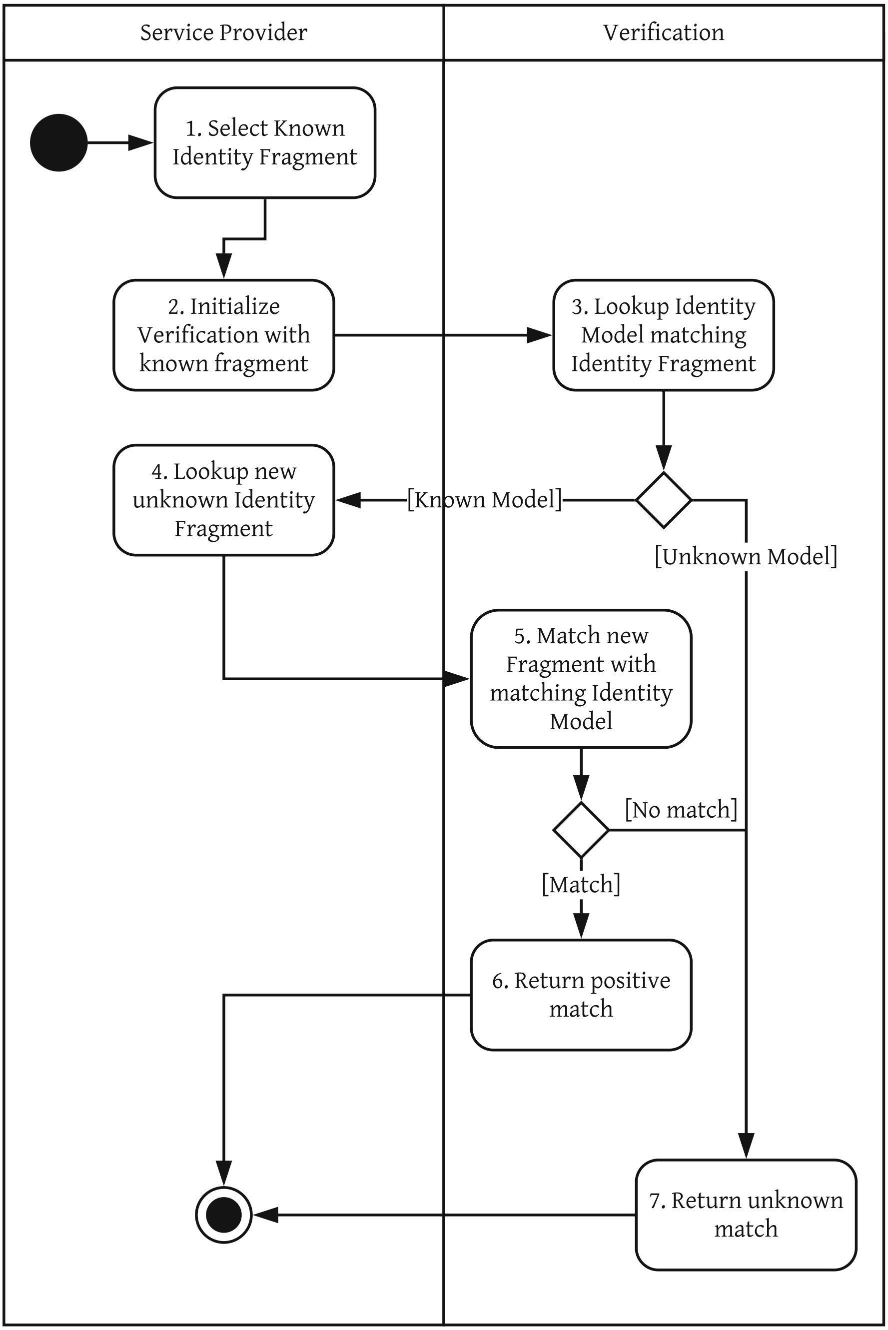


Fig. 5. Verification process from service providers

3.6 Component 6: Verification

Finally, PICoCo can also be used a means to verify new personal information. Service providers should still actively ask their users to maintain up-to-date personal information. However, they can use this model to verify the potential accuracy of the provided information provided. This process is shown in Figs. 5 and 6.

As a user updates their personal information or provides new personal information, the service provider can log a query. They select and provide known personal information (steps 1 & 2) and provide this to the Verifier.

The verifier will look for a matching identity model (step 3). If not none is found, an unknown match is returned (step 7). Otherwise, the service provider will provide the newly acquired personal information to be tested against this known identity model (step 4). PICoCo can then verify whether it believes the queried personal information matches identity fragments that PICoCo has against a known identity model (step 5). If it does, a positive match is returned (step 6), otherwise, PICoCo will return that it is not confident whether data can be linked and verified to an agreed confidence (step 7).

PICoCo's confidence of linking and verification will depend on the type of personal information being verified. PICoCo uses a similar approach as discussed in Sect. 3.2.

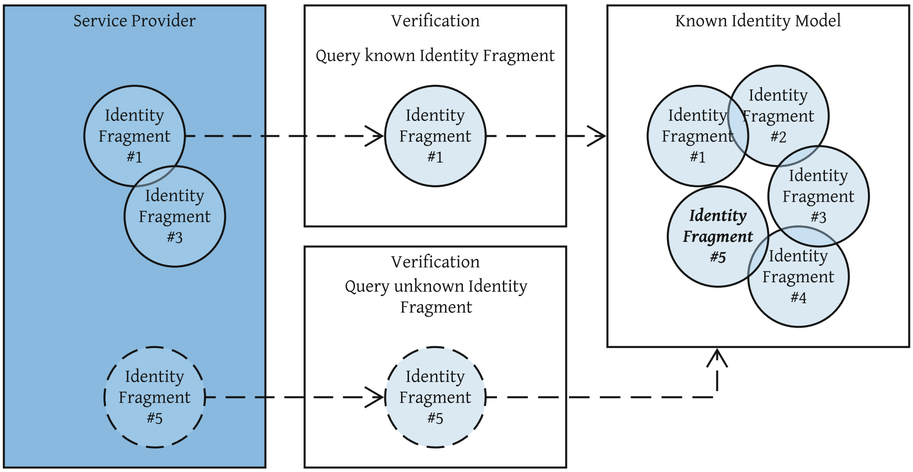


Fig. 6. Verification of identity fragments from service providers

No Information Provided Actively Provided to Service Providers. At this stage, we feel it should be explicitly noted that PICoCo will *never* actively provide or give personal information to service providers! The verification can only be used for actively confident identity models and the verification of additional personal information against known identity fragments.

4 Analysis of PICoCo

In this section we will briefly analyse PICoCo described in previously in Sect. 3. We will analyse it in terms of privacy concerns, accuracy, feasibility, and future work.

4.1 Privacy Concerns

Probably the biggest concern of PICoCo is that of privacy. Essentially the question that can be asked is "How does collating all personal information help privacy?" The simple answer is that the collection and collation of a natural person's personal information, spread across the Internet, will allow better oversight over their current state of affairs.

By allowing users to keep track of where their personal information is being stored and how it is being used, will allow such a person to make informed decisions going forward. Decisions with regard to how they provide and supply their own information to service providers will now be made with a better understanding of such a service provider.

Users who are unfamiliar with the idea of security online and how data online can affect their privacy, can make use of other services that can be provided by PICoCo. One such an approach is described in [1] where a online privacy metric and score is calculated to give the user an idea of their current privacy state.

It should also be noted, that we will under no circumstances force a user or a service provider to make use of this service. However, the advantages of applying such a service to their personal information will be made clear.

The issue of privacy raised by this model is closely linked to that of feasibility, discussed later in Sect. 4.3.

4.2 Accuracy of Personally Identifiable Information

The simple fact that PICoCo relies on other service providers for updated PII should bring the accuracy of said PII into question.

While it is by no means perfectly assured, there can be high level of confidence that personal information on said service providers – especially when compared and collated by other service providers – will be accurate.

Service providers still have to go through their verification process to ensure the accuracy of personal information provided to them. The use of this model is merely there to collate said personal information, as well as provide an overview of personal information to the user. Using this model for verification purposes is merely an advantageous side-effect.

4.3 Feasibility

Questions to be asked:

"Is this model feasible?"
"Will service providers agree to this?"
"Will users agree to this?"

At this point in design, and in our society, this type of model is purely fantastical and slightly offbeat. However, we feel that the investigation of such a model is very necessary. There are multiple benefits that can be gained from the implementation of such a model, but equally as many disadvantages.

However, considering the PKI Model, we use and actively trust central trust authorities – Certification Authorities (CAs). In the PKI model, the entire process relies on the CA. In the same light, we claim that the central agency/agencies for a PICoCo implementation could be seen as a modernised CA.

The answers to the questions at the start of this subsection is probably: "No". But, looking for the future, there might be some use for it. We discuss some possible future work in the following subsection.

4.4 Future Work

The model described in this paper is very abstract at its very least. For the future, we would like to implement an initial working prototype of this model.

The first perceived step would be to look at the service provider side and design a practical integration with existing service providers. Taking into account that different service providers will store personal information differently, we will need to investigate the least-obtrusive means of integrating the collection part of our model into existing systems.

The second would obviously be to then look at implementing the collector itself along with its collator and classifier. This will involve basic principles of data identification and data sorting. Storing the data in the most efficient way will also be important – from a retrieval efficiency perspective as well as storage size perspective.

Finally, we would like to look at how we will allow users to query their personal information. Most importantly would be investigate means of identifying and authenticating the user.

5 Conclusion

In this paper, we investigated identity models and how they are composed of identity fragments consisting of Personally Identifiable Information (PII). Ultimately we introduced and discussed our PICoCo (Personal Information Collection and Collation) Model that collects personal information in such a way to allow it build a digital identity model for a natural person. PICoCo's components were discussed and how they operate together.

Finally, we performed a brief analysis on the model and looked at future work possible from this model.

References

1. Blauw, F.F., von Solms, S.H.: Towards quantifying and defining privacy metrics for online users. In: 2017 IST-Africa Week Conference (IST-Africa), pp. 1–9, May 2017
2. Boardman, R., Sasse, M.A.: Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 583–590. ACM (2004)
3. Council of European Union: What is personal data? https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en
4. Council of European Union: Council Directive 95/46/EC (EU) (1995). http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046
5. Council of European Union: Council Regulation (EU) 2016/679 (2016). http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016R0679
6. Krishnamurthy, B., Wills, C.E.: On the leakage of personally identifiable information via online social networks. In: Proceedings of the 2nd ACM Workshop on Online Social Networks, WOSN 2009, pp. 7–12. ACM, New York (2009). http://doi.acm.org/10.1145/1592665.1592668
7. McCallister, E., Grance, T., Scarfone, K.A.: Guide to protecting the confidentiality of personally identifiable information (PII). Technical report (2010). http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=904990
8. Milberg, S.J., Burke, S.J., Smith, H.J., Kallman, E.A.: Values, personal information privacy, and regulatory approaches. Commun. ACM **38**(12), 65–74 (1995)
9. Norberg, P.A., Horne, D.R., Horne, D.A.: The privacy paradox: personal information disclosure intentions versus behaviors. J. Consum. Aff. **41**(1), 100–126 (2007)
10. Oxford University Press: Definition of natural person. https://en.oxforddictionaries.com/definition/natural_person
11. Phelps, J., Nowak, G., Ferrell, E.: Privacy concerns and consumer willingness to provide personal information. J. Public Policy Mark. **19**(1), 27–41 (2000)

Towards an Ethical Security Platform

Scott Cadzow(✉)

Cadzow Communications Consulting Ltd., Sawbridgeworth, UK
scott@cadzow.com

Abstract. The world of **I**nformation & **C**ommunications **T**echnology (**ICT**) security has been dominated by the **C**onfidentiality **I**ntegrity **A**vailability (**CIA**) paradigm for several decades now and has been very effective in countering relatively simple document based security threats of masquerade, exposure of confidential data, and verification of integrity. Unfortunately real world security problems are not discrete or document based but are complex multi-domain, multi-value ones. In such environments the conventional CIA paradigm is no longer the ideal fit and in particular as we become more reliant on ICT for living support then hard security in the context of CIA needs to be reconsidered. This means taking into account issues that are traditionally "soft" such as Ethics and Dignity and making them "hard" and developing solutions that allow us to treat them. Our starting position is that humans design, operate and are the net beneficiaries of most systems. However humans are fallible and make mistakes. At the same time humans are adaptable and resourceful in both designing systems and correcting them when they go wrong. In contrast machines have in the main been designed to follow rules and are often constrained to produce the same output for the same input over and over again. As we move towards autonomous and intelligent machines the older models of ICT and ICT security based on the CIA paradigm, or deterministic code execution become more and more challenged. Into this mix we then bring a requirement for making ethical decisions.

Keywords: Security · Safety · Ethics · Artificial Intelligence
Machine Learning

1 Introduction - Why Ethics?

Ethical decisions require that different outputs arise from apparently identical inputs as the wider context for the decision has changed. Adaptive machines already appear to have made the switch from deterministic code and the rise of Artificial Intelligence will hasten this switch. The primary concern we ought to have in the long term of AI and M2M is that whilst humans make ethical decisions almost automatically as we move towards an increasingly machine led society those aspects of dignity, ethics and security which are managed by humans will be addressed by machines. The aim of this paper is to give an overview of the state of the art in security standardisation in machine to machine and IoT systems, for the use cases of eHealth and autonomous transport systems, in order to outline the new ethics and security challenges of the machine led

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 523–533, 2018.
https://doi.org/10.1007/978-3-319-91238-7_41

society. This will consider progress being made in standards towards the ideal of each of a Secure and Privacy Preserving Turing Machine and of an Ethical Turing Machine.

2 Human Fallibility

We start with a simple assertion: Humans design, operate and are the net beneficiaries of most systems. We can also assert as a consequence that humans are the net losers when systems go wrong. If that failure is in the security systems trust in the system can disappear.

Humans are fallible and make mistakes. It is also essential to recognise that humans are adaptable and resourceful in both designing systems and correcting them when they go wrong. These characteristics mean that humans can be both the strongest and the weakest link in system security. It also means that there is an incentive to manage the human element in systems such that those systems work well (functionality matches the requirement), efficiently (don't overuse resources), safely and securely. Thus human centric design, even for mostly machine based systems, is essential. However as we adopt more and more machines as human proxies we need to provide some of our intelligence into the systems but in doing so we need to be aware that the intelligence we are offering in machine systems will be different from that of humans. In part this is because the intelligence is by its nature artificial. The consequence may be that we actually design in fallibility.

The purpose of this paper is to mark those elements of the connected world and the publicised attacks on it, and to identify steps that security engineers should be taking to minimise the concerns raised. Addressing the fear of the threat model, promoting why good design works, relegating the "movie plot" threats to the fiction they belong in. The existing security design paradigms are those of "Secure by default" and "Privacy by design". It is not suggested that either of these paradigms is complete and that every product is both secure by default, and privacy protecting (privacy preserving) by design, however even when privacy is protected and security is assured the need for systems to act ethically and to treat their affected users with dignity needs to be assured too. The role of ethics—doing the right thing—in design is not yet clear as it is also not clear in real life. However as more and more decision making is moved into the machine world the need for machines and systems of machines to make the right decision is going to arise more and more. The consideration of dignity is perhaps even harder to quantify but again in machines interacting with humans there is often a need to treat the recipient with a certain degree of dignity, and furthermore to allow the human actor to hold their dignity intact.

The path to Artificial Intelligence, and to Machine Learning (the implied role of both is to gain wisdom in the use of data), is obviously key in this mode of development. The Ethics problem lies at the very top of the data transformation tree shown in Fig. 1. However much of our technology lies at the very lowest layers (networking, databases or data collections, some use of the semantic web).

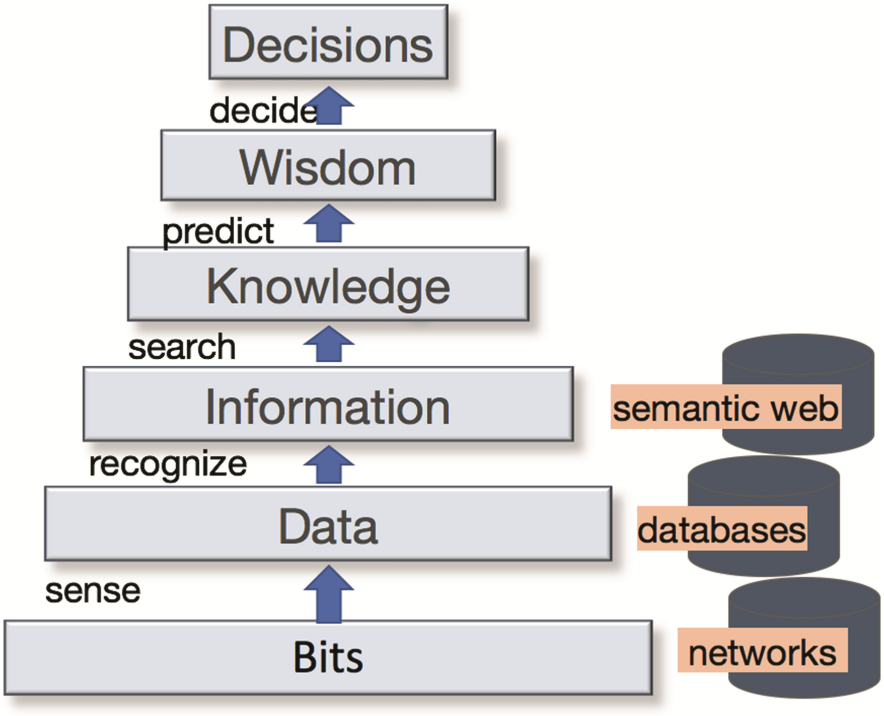


Fig. 1. Data transformation hierarchy from bits to decision

In terms of information processing experience it may be suggested that for the bulk of systems we lie somewhat to the left of the expertise scale (i.e., closer to Random decision making than at Expert decision making) (Fig. 2).

Fig. 2. The time dimension of learning

In looking to use cases there are two very obvious areas where machine ethics will be critical. In the domain of Intelligent Transport Systems (ITS) the operation of semi and fully autonomous vehicles will be increasingly divorced from human control, thus at the point when a crash is inevitable the vehicle has to be able to react in a way that minimises injury to both the occupants and to anyone or anything in the local area. This introduces the classical trolley problem at the ethical decision point: one decision may kill n people, the alternative will definitely kill 1 person. What is the right choice to make? Obviously neither answer is right. There is no rationale for the vehicle to disavow itself of all responsibility and pass control to the local human so the ethical processing has to be built into the machine. In practice the existence of ethical decision points are most visible in hindsight as the consequences of decisions are often out of sight at the decision point.

The second critical domain is that of health—the classical source of the Hippocratic Oath which is often simplified down to "do not harm".

3 Intelligent Gaming and Game Theory in Machine Ethics

Ethical decisions are often both time critical and time variant. What is "right" in one context may be "wrong" in another context, where context may include the players, the time, the location or any other variable? An ethical problem often needs solved at the time it arises—there can be no delay without the problem resolving itself or any solution being invalid. Thus the trolley problem is one oft cited example (see Fig. 3).

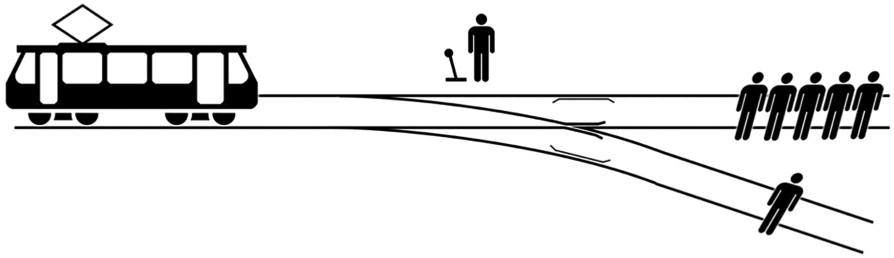


Fig. 3. The Trolley Dilemma (from McGeddon - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=52237245)

The problem is often phrased as a runaway train carriage at speed whilst ahead, on the track, there are five people tied up and unable to move. The train is headed straight for them. You are standing some distance off in the train yard, next to a lever controlled junction. If you pull this lever, the train will switch to a different set of tracks. However, you notice that there is one person on the side track. You have two options:

1. Do nothing, and the trolley kills the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

Which is the most ethical choice? If the choice is to be made by a machine how is the machine programmed? There is no correct choice of course and that is a problem of ethics—the right answer is almost wholly contextual and the deciding actor has limited perspective so can only see the 5 versus 1 conundrum. It is kind of assumed that all alternative avenues have either been tried and failed or are simply not available. How do you win and kill nobody? You can't without changing the problem and modifying the ethical argument.

An alternative view is that presented by the classical prisoner's dilemma but for the general case of co-operation. In moving away from the binary choice in the trolley dilemma the number of actors involved can be expanded such that actors can collude to define the ethically preferable outcome. In the trolley dilemma for example can the trolley itself become involved in the decision? Can it take actions that alter the set of possible outcomes? If we take the prisoner's dilemma where the temptation payoff (T) is greater than the Reward payoff (R) which is greater than the Sucker payoff (S) and which is greater than the Punishment payoff (P) we want to be able to get the actors to work in such a way that with or without collusion they always choose to receive R on the assumption that mutually beneficial strategies are better over the long term.

Game theory is suggested as one way in which ethical issues can be considered. However in order to make such tools work effectively there are a number of pre-conditions that need to be met. The assertion of this paper is that many of the pre-conditions require a commitment to standards to assure interoperability and this is explored more below. It is further contended that the AI that will underpin real time application of game theory itself needs to be standardised at least at the level at which AI systems can interconnect.

4 The Role of Standards

Standards are at the root of sharing a common syntactical and semantic understanding of our world. This is as true for security as it is for any other domain and has to be embraced.

The more flexible a device is the more likely it is to be attacked by exploiting its flexibility. We can also assert that the less flexible a device is it is less able to react to a threat by allowing itself to be modified.

The use of the Johari Window [JOHARI] to identify issues is of interest here (using the phrasing of Rumsfeld).

	Known to self	Not known to self
Known to others	Known knowns - BOX 1	Unknown knowns - BOX 2
Not known to others	Known unknowns - BOX 3	Unknown unknowns - BOX 4

The human problem is that the final window, the unknown unknowns, is the one that gives rise to most fear but it is the one that is not reasonable (see movie plot threats below). The target of security designers is to maximise the size of box 1 and to minimise the relative size of each of box 2 and box 3. In so doing the scope for box 4 to be of unrestrained size is hopefully minimised (it can never be of zero size).

We can consider the effect of each "box" on the spread of fear:

BOX 1: Knowledge of an attack is public knowledge and resources can be brought to bear to counter the fear by determining an effective countermeasure
BOX 2: The outside world is aware of a vulnerability in your system and will distrust any claim you make if you do not address this blind spot

BOX 3: The outside world is unaware of your knowledge and cannot make a reasonable assessment of the impact of any attack in this domain and the countermeasures applied to counter it
BOX 4: The stuff you can do nothing about as e.g., far as you know nothing exists here.

In the security domain we can achieve our goals both technically and procedurally. This also has to be backed up by a series of non-system deterrents that may include the criminalisation under law of the attack and a sufficient judiciary penalty (e.g., interment, financial penalty) with adequate law enforcement resources to capture and prosecute the perpetrator. This also requires proper identification of the perpetrator as traditionally security is considered as attacked by *threat agents,* entities that adversely act on the system. However in many cases there is a need distinguish between the threat source and the threat actor even if the end result in terms of technical countermeasures will be much the same, although some aspects of policy and access to non-system deterrents will differ. A *threat source* is a person or organisation that desires to breach security and ultimately will benefit from a compromise in some way (e.g., nation state, criminal organisation, activist) and who is in a position to recruit, influence or coerce a threat actor to mount an attack on their behalf. A *Threat Actor* is a person, or group of persons, who actually performs the attack (e.g., hackers, script kiddy, insider (e.g., employee), physical intruders). In using botnets of course the coerced actor is a machine and its recruiter may itself be machine. This requires a great deal of work to eliminate the innocent threat actor and to determine the threat source.

The technical domain of security is often described in terms of the CIA paradigm (Confidentiality Integrity Availability) wherein security capabilities are selected from the CIA paradigm to counter risk to the system from a number of forms of cyber attack. The common model is to consider security in broad terms as determination of the triplet {threat, security-dimension, countermeasure} leading to a triple such as {interception, confidentiality, encryption} being formed. The threat in this example being interception which risks the confidentiality of communication, and to which the recommended countermeasure (protection measure) is encryption.

The very broad view is thus that security functions are there to protect user content from eavesdropping (using encryption) and networks from fraud (authentication and key management services to prevent masquerade and manipulation attacks). What security standards cannot do is give a guarantee of safety, or give assurance of the more ephemeral definitions of security that dwell on human emotional responses to being free from harm. Technical security measures give hard and fast assurance that, for example, the contents of an encrypted file cannot, ever, be seen by somebody without the key to decrypt it. So just as you don't lock your house then hang the key next to the door in open view you have to take precautions to prevent the key getting into the wrong hands. The French mathematician Kerchoff has stated "A crypto system should be secure even if everything about the system, except the key, is public knowledge". In very crude terms the mathematics of security, cryptography, provides us with a complicated set of locks and just as in choosing where to lock up a building or a car we need to apply locks to a technical system with the same degree of care. Quite simply we don't need to bother installing a lock on door if we have an open window next to it—the attacker will ignore the locked door and enter the house through the open window. Similarly for a cyber system if crypto locks are put in the wrong place the attacker will bypass them.

It may be argued that common sense has to apply in security planning but the problem is that often common sense is inhibited by unrealistic threats such as the movie plot scenarios discussed below.

Standards are peer reviewed and have a primary role in giving assurance of interoperability. Opening up the threat model and the threats you anticipate, moving everything you can into box 1, in a format that is readily exchangeable and understandable is key. The corollary of the above is that if we do not embrace a standards view we cannot share knowledge effectively and that means we grow our box 2, 3, 4 visions of the world and with lack of knowledge of what is going on the ability of fear to grow and unfounded movie plot threats to appear real gets ever larger.

5 Standards and Interoperability

Let us take health as a use case for the role of standards in achieving interoperability. When a patient presents with a problem the diagnostic tools and methods, the means to describe the outcome of the diagnosis, the resulting treatment and so on, have to be sharable with the wider health system. This core requirement arises from acceptance that more than one health professional will be involved. If this is true they need to discuss the patient, they need to do that in confidence, and they need to be accountable for their actions which need to be recorded. Some diseases are "notifiable" and, again, to meet the requirement records have to be kept and shared. When travelling a person may enter a country with an endemic health issue (malaria say) and require immunisation or medication before, during and following the visit. Sharing knowledge of the local environment and any endemic health issues requires that the reporting and receiving entities share understanding.

Shared understanding and the sharing of data necessary to achieve it is the essence of interoperability. A unified set of interoperability requirements addresses syntax, semantics, base language, and the fairly obvious areas of mechanical, electrical and radio interoperability.

Syntax derives from the Greek word meaning ordering and arrangement. The sentence structure of subject-verb-object is a simple example of syntax, and generally in formal language syntax is the set of rules that allows a well formed expression to be formed from a fundamental set of symbols. In computing science syntax refers to the normative structure of data. In order to achieve syntactic interoperability there has to be a shared understanding of the symbol set and of the ordering of symbols. In any language the dictionary of symbols is restricted, thus in general a verb should not be misconstrued as a noun for example (although there are particularly glaring examples of misuse that have become normal use, e.g., the use of "medal" as a verb wherein the conventional text "He won a medal" has now been abused as "He medalled"). In the context of eHealth standardisation a formally defined message transfer syntax should be considered as the baseline for interoperability.

Syntax cannot convey meaning and this is where semantics is introduced. Semantics derives meaning from syntactically correct statements. Semantic understanding itself is dependent on both pragmatics and context. Thus a statement such as "Patient-X has a heart-rate of 150 bpm" may be syntactically correct but has no practical role without understanding the context. Thus a heart-rate of 150 bpm for a 50-year old male riding a bike at 15 km/h up a 10% hill is probably not a health concern, but the same value when the same 50 year old male is at rest (and has been at rest for 60 min) is very likely a serious health concern. There are a number of ways of exchanging semantic information although the success is dependent on structuring data to optimise the availability of semantic content and the transfer of contextual knowledge (although the transfer of pragmatics is less clear).

Underpinning the requirements for both syntactic and semantic interoperability is the further requirement of a common language. From the eHealth world it has become clear that in spite of a number of European agreements on implementation of a digital plan for Europe in which the early creation of 'e-health' was eagerly expected the uneven development of the digital infrastructure has in practice made for differing levels of initiative and success across the member states. These led to a confusing vocabulary of terms and definitions used by e-health actors and politicians alike. The meaning of the term e-health has been confused with 'tele-health' which in turn is confused with 'm-health' 'Telemedicine,' a term widely used in the USA has been rejected in Europe in favour of 'tele-health.' There is general agreement that for these terms to be effective we need to redefine them in their practical context. Without an agreed glossary of terms, it will be hard to improve semantic interoperability—a corner stone for the effective building of e-health systems. The vocabulary is not extensive but at present it fails to address the need for clarity in exchange of information in the provision of medical services.

Standards therefore enable and assert interoperability on the understanding that:

$$\text{Interoperability} = Semantics \cup Syntax \cup Language \cup Mechanics$$

Quite simply if any of the elements is missing then interoperability cannot be guaranteed. However we do tend to layer standards on top of one another, and alongside each other, and wind them through each other. The end result unfortunately can confuse almost as much as enlighten and unfortunately the solution of developing another standard to declutter the mess often ends up with just another standard in the mess.

In the security domain understanding that we need interoperability is considered the default but simply achieving interoperability is a necessary but insufficient metric for making any claim for security. As has been noted above the technical domain of security is often described in terms of the CIA paradigm (Confidentiality Integrity Availability) wherein security capabilities are selected from the CIA paradigm to counter risk to the system from a number of forms of cyber attack. The common model is to consider security in broad terms as determination of the triplet {threat, security-dimension, countermeasure} leading to a triple such as {interception, confidentiality, encryption} being formed. The threat in this example being interception which risks the confidentiality of communication, and to which the recommended countermeasure (protection measure) is encryption.

The very broad view is thus that security functions are there to protect user content from eavesdropping (using encryption) and networks from fraud (authentication and key management services to prevent masquerade and manipulation attacks). Technical security, particularly cryptographic security has on occasion climbed the ivory tower away from its core business of making everyday things simply secure.

6 Movie Plot Threats

Bruce Schneier has defined movie plot threats as "… *a scary-threat story that would make a great movie, but is much too specific to build security policies around*"[1] and rather unfortunately a lot of the real world security has been in response to exactly these kind of threats. Why? The un-researched and unproven answer is that movie plots are easy to grasp and they tend to be wrapped up for the good at the end.

The practical concerns regarding security and the threats they involve is that they are somewhat insidious, like dripping water they build up over time to radically change the landscape of our environment.

Taking Schneier's premise that our imaginations run wild with detailed and specific threats it is clear that if a story exists that anthrax is being spread from crop dusters over a city, or that terrorists are contaminating the milk supply or any other part of the food chain, that action has to be taken to ground all crop dusters, or to destroy all the milk. As we can make psychological sense of such stories and extend them by a little application of imagination it is possible to see shoes as threats, or liquids as threats. So whilst Richard Reid[2] was not successful and there is no evidence to suggest that a group of terrorists were planning to mix a liquid explosive from "innocent" bottles of liquid, the impact is that due to the advertised concerns the policy response is to address the public fears. Thus we have shoe inspections and restrictions on carrying liquids onto planes. This form of movie theatre scenario and the response ultimately diverts funds and expertise from identifying the root of many of the issues.

Again taking Schneier's premise the problem with movie plot scenarios is that fashions change over time and if security policy is movie plot driven then it becomes a fashion item. The vast bulk of security protection requires a great deal of intelligence gathering, detail analysis of the data and the proposal of targeted counter measures. Very simply by reacting to movie plots the real societal threats are at risk of being ignored through misdirection.

Movie plot derived security policy only works when the movie plot becomes real. If we built out bus network on the assumptions behind Speed we'd need to build bus stops for ingress and egress that are essentially moving pavements that don't allow for the bus to ever slow down, and we'd need to be able to refuel and change drives also without slowing the bus. It'd be a massive waste of money and effort if the attackers did a Speed scenario on the tram or train network or didn't attack at all.

A real problem is that for those making security policy, and for those implementing the countermeasures, they will always be judged in hindsight. If the next attack targets

1 https://www.schneier.com/blog/archives/2014/04/seventh_movie-p.html.
2 https://en.wikipedia.org/wiki/Richard_Reid => The "shoe bomber".

the connected vehicle through the V2I network, we'll demand to know why more wasn't done to protect the connected vehicle. If it targets schoolchildren by attacking the exam results data, we'll demand to know why that threat was ignored. The answer "we didn't know…" or "we hadn't considered this…" is not acceptable.

The attractiveness of movie plot scenarios is probably hard to ignore—they give a focus to both the threat and the countermeasures. In addition we need to consider the role of Chinese Whispers[3] in extending a simple story over time.

We can imagine dangers of believing the end point of a Chinese Whispers game:

- Novocomstat has missile launch capability
- Novocomstat has launched a missile
- Novocomstat has launched a bio weapon
- Novocomstat has launched a bio weapon at Neighbourstat
- Neighbourstat is under attack
- Neighbourstat is an ally and we need to defend them
- We're at war with Novocomstat because they've attacked with the nuclear option

As security engineers the guideline is to never react without proof. Quite simply acting on the first of these Chinese Whispers is unwarranted, and acting on the 6th is unwarranted unless all the prior statements have been rigorously verified, quantified and assessed. The various risk management and analysis approaches that exist (there are many) all come together by quantifying the impact of an attack and its likelihood. In recent work in this field in ETSI the role of motivation as well as capability in assessing risk has been re-assessed and now added to the method [E-TVRA]. The aim in understanding where to apply countermeasures to perceived risk requires analysis. That analysis requires expertise and knowledge to perform. In the approach defined by ETSI in TS 102 165-1 this means being able to quantify many aspects of carrying out a technical threat including the time required, the knowledge of the system required, the access to the system, the nature of the attack tools and so forth.

7 Where to Go?

How do you get rid of fear and get acceptance of the threat model? Shared knowledge, shared understanding and willingness to educate each other about what we know and what we may not know. This is the only real way forward. This result is close to zero in boxes 2 and 3 and a bounteous box 1.

8 Conclusions

As stated in Sect. 6 of this paper the approach to getting rid of fear and get acceptance of the threat model is in the wider acceptance of shared knowledge, shared understanding and willingness to educate each other about what we know and what we may not know.

[3] https://en.wikipedia.org/wiki/Chinese_whispers => A parlour game that passes a message round introducing subtle changes in meaning with each re-telling.

The role of standards in giving assurance of interoperability as the key to a solution where more than one stakeholder is involved is difficult to argue against. The nature of the standard is unimportant—it simply has to be accepted by the stakeholders. If the stakeholders are global and largely unknown then an internationally accepted standard is most likely to be the way forward. If, however, the stakeholders are members of a small local team the standard could be as simple as a set of guidance notes maintained on a shared file.

Spreading of fear through a combination of movie plot threats and Chinese Whispers is an inevitable consequence of human curiosity and imagination.

Standards are at the root of sharing a common syntactical and semantic understanding of our world. This is as true for security as it is for any other domain and has to be embraced.

Acknowledgements. Contributions made by the author in development of this paper have in part been supported by EU projects i-locate (grant number 621040), SUNSHINE (grant number 325161) and UNCAP (grant number 643555).

References

[E-TVRA] ETSI TS 102 165-1. https://portal.etsi.org/webapp/WorkProgram/SimpleSearch/QueryForm.asp

[JOHARI] Luft, J., Ingham, H.: The Johari window, a graphic model of interpersonal awareness. In: Proceedings of the Western Training Laboratory in Group Development, Los Angeles (1955)

PassHue: Introducing Analog Authentication

Jonathan Gurary(✉)

Cleveland State University, Cleveland, USA
jongurary@live.com

Abstract. Authentication, particularly in the mobile environment, has still failed to capture many useful types of human mental ability. In this paper, we present the concept of *analog authentication*, an idea revolving around the use of continuously variable information to authenticate users. Analog authentication shifts the memory task from pure recall to *estimation*, and taps into fundamental human intuition for differentiating items along a continuum. We design PassHue, a color-based analog authentication scheme, to demonstrate the memorability, usability, and security advantages that analog authentication can offer. Our in-the-wild user study demonstrates that PassHue can achieve average login times under 1.7 s, and remain memorable over a period of 2 weeks.

1 Introduction and Related Work

1.1 Mobile Authentication

Mobile devices present unique challenges for authentication. These devices have especially high usability requirements; an average smartphone user unlocks their phone 48 times a day [11]. A difference of one second between two authentication methods can cost the user hours in the long term. Many users forgo device authentication altogether to avoid hassle [8]. At the same time, mobile devices carry increasingly sensitive information such as banking and stock trading applications. In many cases, an unlocked mobile device capable of receiving text messages can serve as the sole authentication requirement to reset a password, for example to email accounts, which can in turn be used to access even more sensitive information. The lock screen can be the sole defense against an attacker that already has physical possession of the device.

According to a 2016 Pew survey of smartphone owners [17], numeric PIN (25%) and Fingerprint scanner (23%) are currently the most common methods of locking mobile devices, trailed by alphanumeric password (9%) and Google's Pattern Unlock (9%). While the fingerprint reader and other biometric methods are quickly gaining popularity, neither iOS nor Android allow fingerprint authentication as the sole lock method, with both requiring either a password or a PIN as a fallback measure. Additionally, many low-cost devices are still sold without fingerprint readers or other biometric scanners. At this time, knowledge-based

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 534–553, 2018.
https://doi.org/10.1007/978-3-319-91238-7_42

methods are still the primary method of authentication for most users, and even biometric approaches are still backed by knowledge-based authentication.

Alphanumeric text passwords are especially ill-suited to mobile devices [26] because these devices lack a full-sized hardware keyboard to facilitate fast and accurate typing.

The numeric PIN is notoriously insecure; there are only $4^{10} = 10,000$ possible 4-digit PINs, and PINs are easy to observe in a shoulder-surfing attack. Furthermore, the top 100 most popular PINs are chosen by almost a third of users [14]. Several authors have proposed improvements to the PIN scheme. Roth et al. [22] proposed to extend PIN's shoulder-surfing resistance by splitting the digits into two sets. Instead of picking a digit, users pick the set, and the PIN is eventually entered by the intersection of the sets. Von Zezschwitz et al. proposed SwiPin [25], a method to increase the shoulder-surfing resistance of PIN by using directional touch gestures for input. De Luca et al. propose ColorPIN [7], which increases the number of possible passwords to $531,441$ by adding a color dimension. Bianchi et al. [3] proposed several approaches for entering a PIN using haptic or audio cues.

Google's Pattern Unlock has $389,112$ possible passwords on a 3 by 3 grid, but a space of just 300 passwords may capture around 50% of users [24]. Pattern unlocks are also easy to observe in shoulder-surfing attacks and easy to replicate from the smudge left behind by the user's finger on the screen [1]. Kwon and Na's TinyLock [15] offers increased resistance against shoulder-surfing and smudge attacks with minimal usability penalty by leveraging a smaller drawing area and a special gesture which obfuscates the resulting smudge pattern.

The low security strength of existing mobile authentication has led to security-conscious people and organizations adopting a number of frustrating developments, for example long lockouts or even a permanent device wipe after several failed attempts, length requirements of 8+ digits for PINs and 6+ digits for passwords, system assigned passwords, and strict limitations on common passwords, repeating patterns, and password reuse. The usability and memorability of long PINs is significantly worse than the 4-digit base [12,14]. In other words, the poor security strength of existing methods has led to usability woes for many users.

The goal of this work is to create an authentication method that meets the usability standard of existing authentication methods but offers significantly improved security. We propose analog authentication as a framework for developing authentication based on a continuum. We demonstrate that PassHue, an example of analog authentication, can achieve high usability while offering better security strength than existing methods.

2 Authentication Using Continuous Information

Most traditional authentication methods ask users to remember information which is discrete, such as letters, numbers, or an ordered pattern like Pattern Unlock. Users remember a sequence of discrete information and *recall* that information back exactly. Some methods, such as RealUsers's PassFaces [21], ask users to remember discrete items such as faces or patterns and *recognize* them from a larger set of items later.

By discrete, in this paper, we mean that the information being remembered can be divided easily into a whole number of choices: there are 26 letters in the alphabet, 10 digits, and 3–8 possible directions that a user can pick from any given dot in a pattern. In practice, many sets which are treated as continuous may also be considered discrete, for example 3D space is sometimes argued to be discrete in terms of the plank length. The discussion of which sets are discrete vs continuous is outside the scope of this paper; any sufficiently large set which appears to be continuously variable to an average human will be considered continuous.

Analog authentication asks users to remember information from a continuum. That is, given a continuum such as loudness, an analog authentication scheme would ask the user to reproduce a specific volume or volumes. The memory task is effectively extended from *recall* to *estimation*, as the user must now not only remember the volume that was previously set but also estimate it accurately. By necessity, a tolerance must be given to the user for error. The password space of an analog authentication scheme is proportional to the size of the continuum divided by the tolerance. Thus, analog authentication has a direct tradeoff between usability (a larger tolerance so the user can authenticate more easily), and security (a smaller tolerance to increase the size of the password space).

We use the term analog authentication to convey that a continuum of information is being used rather than discrete information. The concept is not to be confused with continuous authentication, which refers to authentication that functions by analyzing user behavior in the background while the user is interacting with a device.

Intuitively, it is tempting to assume that humans will perform at vastly varying abilities depending on the estimation task. The well-known "seven, plus or minus two" rule [16] dictates that the average person can distinguish between about 7 pieces of continuous information at a time. When a continuous set is broken into n items, humans start having trouble discerning between items when n is larger than 7. The general rule holds for continuous sets such as pitch, saltiness, loudness, or points in a square. Humans are generally able to discern between no more than 7 unique items before accuracy beings to suffer considerably. Splitting the continuum any more finely leads to errors with rapidly increasing frequency.

Cowan [6] describes the limit as "The Magical Mystery Four" instead, arguing that working memory for the average young adult is limited more closely to 3–5 items.

It follows that an important concern in the design of an analog authentication scheme is to ensure that the user does not have to break the continuum down into more than 7 pieces, and fewer is better. Beyond that, the memorability of a particular continuum must be justified individually; there is no research suggesting that continuous data is always more memorable than discrete data or vice-versa. As Miller [16] demonstrates, even though memorability is similar between various types of continuous information, some are still more memorable than others.

Discrete information like letters is often bundled into higher-order information like words or sentences to improve the number of items a person can remember. Similarly, continuous information like pitch and tone can be bundled into higher order information like notes and songs, though this may also have the effect of making the information discrete.

2.1 Related Work in Analog Authentication

Currently, continuous information is seldom used for authentication. Even when potentially continuous information is used, it is often presented in a discrete manner. For example, if the user is asked to pick a color, for example as a banking security question, they are typically presented with a short list of options or asked to use a standard language-based description such as "blue" or "silver".

An exception is free-form gesture drawing, such as the work by Sherman et al. [23] and by Clark and Lindqvist [4]. Free-from gesture drawing uses a continuous 2d drawing to authenticate the user, placing it firmly in the realm of analog authentication. While these works have discussed the implications and advantages of utilizing continuous information as opposed to discrete information, none have formalized the concept of using continuous information outside the scope of free-form gesture drawing. Free-form gesture drawing can be considered just one type of analog authentication. Additionally, free-form drawing can be considered as an example of analog authentication which bundles low-level continuous information (2d positions), into higher order information (lines and shapes), while still preserving the analog nature of the method.

On the contrary, Google's Pattern unlock, and in fact any touch-based authentication on the mobile platform, can be considered examples of turning analog information (2d positions) into discrete information (connections between points). Buttons that users touch to input a PIN or password can be considered a type of tolerance: any 2d positions that fall inside the button count as the same input. We do not count these methods as analog because the memory task facing the user is discrete, only the input method is continuous.

Bianchi's [3] works falls into a similar category: continuous information such as vibration, beat, and hold time is ultimately used as a cue for discrete information like numbers. Remembering the cue is part of the memory task, so Bianchi's approaches can be considered partly analog, however analog cues such as vibration are broken down into discrete functions like "number of vibrations that have elapsed", an integer value which is plainly discrete.

We note that in biometrics, analog authentication is the norm, utilizing continuous data such as gait and typing speed. In this paper, when we discuss authentication methods, we refer specifically to knowledge-based methods, not to biometric methods.

In this paper, we introduce the concept of analog authentication, the idea of using continuous data for authentication. We developed PassHue, a mobile authentication scheme that uses a color continuum, to demonstrate the potential advantages of analog authentication. PassHue follows a PIN-like approach similar to the classic numerical PIN, SwiPIN [25], or ColorPIN [7]. It is designed to be immediately familiar to end users and offer login times and memorability on par with existing PIN-based approaches. PassHue improves on existing mechanisms by providing a much larger password space and moderate protection from shoulder-surfing. As an example of analog authentication, PassHue demonstrates that continuous information can be used for authentication just as well as discrete information. Our in-the-wild user study demonstrates that PassHue can achieve high usability and remain memorable over a period of 2 weeks.

3 The Design of PassHue

In this section, we present the design of PassHue, an example of analog authentication that utilizes color. We envision PassHue to be used as an unlock scheme, potentially integrated with the OS, or for in-app authentication in, for example, email or banking applications.

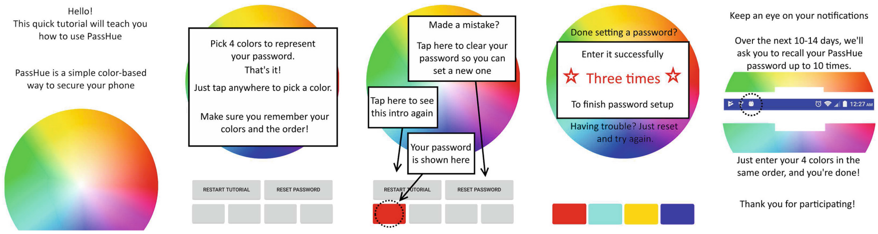


Fig. 1. Tutorial images shown on the store page (Color figure online)

PassHue is implemented on Android. Figure 1 shows the tutorial images that are presented on the Play Store listing for PassHue. Users are given no further guidance beyond these images.

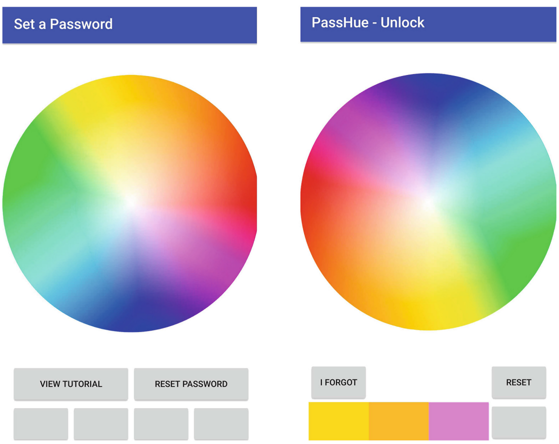


Fig. 2. (left) The password setup screen, (right) The login screen (Color figure online)

PassHue is designed to simultaneously use three continuous sets of color: red, green, and blue, referred to as "RGB" values. The RGB system is the most common method for representing colors in digital applications– a color is made up of one value from each set: R, G, and B. In general, each set has a range of 0–255, so the sets are not actually continuous, but sufficiently large so that they appear continuous to a human. The size of RGB color space is quite large; 256^3 yields approximately 16.8 million possible colors.

Users set a password by tapping 4 colors in order. The 4 RGB color values are stored as the user's password. The password setup screen is shown on the left in Fig. 2. Tapping the "View Tutorial button" allows users to see the images in Fig. 1 again. Before finishing password setup, the user must re-enter the same 4 colors an additional 3 times to verify that they remember the password. Until the 3 verification entries are complete, the password is not set. If the user decides that their password is too hard before verifying it 3 times, or wants to pick a different password for any other reason, they can reset it with no penalty using the "Reset Password" button.

To authenticate themselves later, users must pick the same 4 colors, within the tolerance, and in the same order. The login screen is shown on the right in Fig. 2. The user has already picked three colors: orange, yellow, and pink – those choices are tracked at the bottom of the screen. The fourth color is still awaiting user input. The "Reset" button can be used to clear the current input if a mistake was made, and the "I Forgot" button clears the user's password and allows them to set a new one if they wish to continue the experiment. We include this button so users can easily communicate to us that they do not remember their password.

Colors are picked by tapping the standard color wheel shown in Fig. 2. The wheel is identical to the color wheel found in many graphics applications such as Adobe Photoshop and Paint.net. There are three elements in RGB color, and it is difficult to express all three in a single 2D image while maintaining a continuously variable pattern. Ideally there should not be "jumps" in color when the user moves across the image, otherwise two very different colors may end up adjacent, and this can make picking a color accurately difficult.

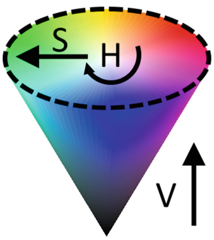


Fig. 3. Cone representation of HSV color space. (Hue is the primary color where red is 0°, Saturation is the strength or intensity of the color, and Value describes how dark the color is.) (Color figure online)

Additionally, the user should be able to locate colors quickly based on location, for example it is expected that orange falls between yellow and red. The color wheel accomplishes these requirements; movement in any direction around the wheel is associated with a gradual change in color, movement towards the center increases the "whiteness" of the color, and colors appear in classic order around the wheel: red, orange, yellow, green, blue (cyan), indigo (dark blue), violet, and purple.

A tradeoff to using a color wheel is that relatively few RGB colors are represented. The color wheel used by PassHue is often called an HSV (Hue, Saturation, Value) wheel, which typically features a Value slider in addition to the wheel. HSV is a simple transformation of RGB. Figure 3 shows how an HSV system addresses colors; the flat area at the top enclosed by the dotted line represents the portion of colors used by PassHue and demonstrates that PassHue uses only Hue and Saturation in the HSV scheme.

PassHue contains only the RGB colors where at least 1 of the 3 RGB values equals 255. This allows PassHue to display a variety of colors with a consistent gradient on a 2D display but sacrifices the ability to display the colors that reside in the rest of the HSV cone.

We design PassHue around color because color is a continuum that humans are relatively good at processing. There are at least 2.8 million colors discernible to normal humans [20], and some researchers suggest as many as 10 million [13]. Halsey and Chapanis [9] presented participants with 342 CIE colors, all of equal luminance, and asked them to match given colors exactly to one of the 342 presented colors. Participants could pick out over 11 unique colors – that is, colors with no overlapping matches to other colors, at the 5% error level – and over 15 colors at the 10% level, significantly better than the expected "magic number seven" [16]. Hamwi and Landis [10] found no relationship between time delay and color memory for delays of 15 min, 24 h, and 6 h, indicating that color may be good for long-term memory.

3.1 Comparison of Color Values

PassHue illustrates a potential difficulty in analog authentication: humans are often better at discerning values on one part of the continuum than on another. This can make it difficult to establish an exact estimate for the tolerance t, since t varies depending on which part of the continuum the user picks. In color, for example, humans are worst at discerning shades of green [9,18], so the tolerance should be greater for green colors. Euclidean distance between colors does not accommodate for different levels of performance with different colors.

PassHue compares colors using the CompuPhase algorithm [5], a commercially-used, simple, and efficient method for calculating the distance between two colors in a way that tries to emulate how a human would perceive the distances. A key advantage of the CompuPhase algorithm over an algorithm like CIE2000 is that it has significantly fewer mathematical operations and does not require conversion into another color space, potentially saving valuable overhead. Processing overhead is especially important on mobile devices with limited computing resources and battery life. The algorithm describes the difference between two RGB colors using the following equation:

$$\sqrt{(2 + \frac{\bar{r}}{256}) \times \Delta R^2 + 4 \times \Delta G^2 + (2 + \frac{255 - \bar{r}}{256}) \times \Delta B^2} \quad (1)$$

where $\bar{r}$ is the mean red level, i.e., $(R_1 + R_2)/2$, and ΔR, ΔG, and ΔB are the differences between the respective R, G, and B values of the two colors. The result of Eq. 1 will be referred to from here on as the similarity score – the lower the score, the more similar the colors.

A similarity score of 100 or lower is considered a match. All 4 colors in a user's password must match for authentication to succeed. That is, the similarity score for all 4 chosen colors vs the stored RGB values for that password must be 100 or less.

A score of 100 was chosen after a brief pilot test with 5 participants. A goal of the user study in this paper is to determine if the score should be raised or lowered for the average person. We hypothesize that for most users, a lower score will be sufficient, while a few may struggle without a higher score. In a full commercial application of PassHue, the similarity score may start high and gradually reduce if the user continuously meets a lower score, allowing users with better color-discerning ability to enjoy increased security.

4 Security Strength of PassHue

Current mobile authentication methods are limited in password space. A 4-digit PIN can generate only $10^4 = 10,000$ passwords, and pattern unlock offers $389,112$ possible passwords on a 3 by 3 grid [24]. In this section, we calculate the password space of PassHue. We will address hotspots and other considerations in password selection in Sect. 5.

The password space of a traditional discrete scheme is C^n, where C is the number of choices per item and n is the number of items chosen. The password space of an analog authentication scheme with one variable is $(C/t)^n$, where C is the size of the continuum, t is the tolerance, and n is the number of choices picked from the continuum. We can consider t to be number of items in C which are treated as identical for purposes of satisfying the password. In a traditional discrete scheme, we can say that $t = 1$.

PassHue uses 3 continuums: red, green, and blue, but only a single tolerance based on a score generated using all 3 values. For simplicity we can combine the colors and consider C to be a single continuum from (0, 0, 0) to (255, 255, 255). In practice, PassHue uses only the top circle of the HSV color cone, so at least one RGB value must always equal 255. We can choose any of the three colors R, G, or B to set equal to 255. While one color must be 255, there are 256 options for both other colors, leaving $(256 * 256)$ choices. The total size of C is therefore $3 * (256 * 256) = 196,608$ colors, which represents just over 1% of RGB color space. Users pick 4 colors, so n is equal to 4.

It is difficult to calculate a value for the tolerance. The similarity score, generated according to Eq. 1, weighs the values in each continuum differently, so the tolerance varies depending on the values of the colors in question.

To find an accurate estimate for t, we write a short script to process all RGB color pairs where at least one value in both colors is equal to 255, and the similarity score between the colors is between 99–100, yielding approximately 40 million pairs. The worst-case product of differences between two colors having a similarity score of 99–100 (i.e., $\Delta R \times \Delta G \times \Delta B$) is approximately 39,000, and the average product is approximately 3,400. That is, for any given RGB value, on average, there are 3,400 other RGB values that would be considered a match for purposes of authentication.

Using our average estimate for t, we find that PassHue has a password space of $(196,608/3,400)^4 = 1.1 * 10^7$ (appox. 23 bits), 1000 times larger than a traditional 4 digit PIN, and 28 times larger than pattern unlock on a 3 by 3 grid.

5 User Study

The experiment in this paper was approved by the IRB at Cleveland State University. Participation is anonymous.

We design our user study to determine the effectiveness of an analog authentication scheme in-the-wild. Users download the application on their own device, set a password, then recall that password several times over a period of 14 days to simulate a phone unlock scheme that is used daily. Participants are given little to no guidance about how to use the scheme, the entire tutorial is contained in Fig. 1, and viewing it is optional. To keep the time burden on our participants low, we notify them to authenticate just once per day, though a typical authentication scheme will be used far more frequently by most users.

5.1 Data Collection

Information is transmitted via php to a dedicated private server. Participants download the application from the Google Play Store.

After downloading the application, participants must verify they are over 18 and consent to our terms before they can continue. After, participants are asked to provide optional demographic information, such as age, which is encrypted and transmitted to our server. Participants are then taken directly to the password setup screen in Fig. 2 with no further guidance.

Passwords are stored on the device's local memory and also transmitted to our server upon creation. Information about authentication attempts is transmitted to our server after each attempt, including total entry time, entry time for each individual color, raw RGB values of each attempted color, and the similarity score between the attempted colors and the actual password.

After initializing a password, participants are notified once per day, at approximately the same time of day the password was originally set, to recall the password. Notifications last a total of 14 days. Participants can chose to ignore a notification.

After 14 days, the application notifies the user to complete an exit survey. Users answer basic questions about how they liked the scheme and are given the option to leave written feedback. This information is encrypted and transmitted to our server.

5.2 Participants

Participants were recruited with fliers around our university, social media posts, and word of mouth. We extend a special thank you to the /r/Android community on Reddit for providing a large number of our participants. To be considered as completing the experiment, participants needed to attempt recall on at least 4 different days, or forget their password and have completed at least 2 authentication attempts. A total of 38 participants completed the experiment. The drop out rate, based on participants who set a password but did not meet the above criteria, is 30% (16 participants).

Participation is anonymous, but participants are asked to provide some optional demographic information at the start of the experiment. Some information, such as Android version and country of origin, is collected automatically by Android. Of the 38 participants, 35 chose to provide their age and 37 provided their gender. The average age of participants was 25.5 (median = 21, std = 11) and our population was 22% female. Most of our participants (≥60%) were using Android 7.0 or higher. Approximately 70% of participants were from the United States, but we also found participants in Canada, Sweden, Russia, Australia, and the UK.

Because our experiment requires ownership of an Android device as well as some rudimentary abilities such as downloading the application from the Play Store, our experiment self-selects towards participants who are already skilled at using their device. Participants self-reported skill with using their device on a scale from 1 (worst) to 5 (best), with an average score of 4.7 (median = 5, std = 0.74). We asked participants what unlock method they currently use to lock their device, with the following options and scores respectively: PIN (1), Dot Pattern (7), Fingerprint (25), Alphanumeric Password (0), Other (2), Don't Lock Device (1), or Prefer not to Answer (2). We note that our population has an unusually high rate of locking their device, but this is expected in a population of people who were interested in an experiment about device authentication. We also note that the rate of fingerprint is quite high, notably because our experiment appealed to mobile phone enthusiasts who tended to have higher end devices which supported fingerprint authentication. As mentioned in Sect. 1, fingerprint authentication is not currently a replacement for knowledge based authentication, and users of fingerprint authentication are still required to set a separate PIN or password as well.

On the first application startup, participants were randomly assigned into one of the following two conditions for the remainder of the experiment:

Stationary – In this condition, the color wheel appears in the same orientation for every login attempt. The default orientation is shown on the left in Fig. 2.

Rotating – In this condition, the color wheel has a different rotation for every authentication session. Figure 2 demonstrates this condition: the color wheel on the right is oriented differently from the color wheel on the left. The wheel's rotation is determined only once, at authentication start; the wheel is *not* rotated again if the user authenticates incorrectly. When initially setting the password, users must confirm the password 3 times before it is set. In the rotating condition, the color wheel is rotated after each successful attempt.

Nineteen participants were assigned to each condition. Participants were not made aware that different conditions existed and received no guidance about rotation or lack thereof. We hypothesize that the rotating condition will perform slightly worse in terms of entry times and failed authentication attempts, but may offer better defense against shoulder-surfing.

5.3 Memorability of PassHue

Three participants forgot their passwords, for an overall memorability of 92%. Two participants belonged to the rotating condition while the third belonged to the stationary condition. We found no significant difference in memorability between the conditions ($\chi^2 = .36, p = .548$).

All three participants forgot their passwords within the first three days of the experiment. After resetting a new password, all three participants went on to complete the experiment successfully. We conclude that PassHue is highly memorable, even after a period of two weeks, but a small subset of users can have issues with initially memorizing a password.

5.4 Usability of PassHue

Figure 4 shows the amount of time, in seconds, needed to create a password, including the required 3 additional successful entries before a password is officially set. As expected, stationary users on average required less time to set a password, with the exception of three outliers. The data shows that with very little guidance, users are able to figure out how to use PassHue in about one minute.

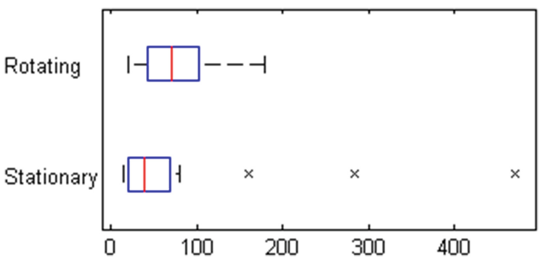


Fig. 4. Creation time

We recorded entry times for a total of 1192 authentication attempts. To obtain more realistic timing data, we filtered attempts that were likely "pocket dials" or random tapping. We categorized these as attempts where the sum of difference scores for the four colors was greater than 500. Most of these attempts had entry times lower than 1.5 s. We filtered out 112 attempts in this manner. We also filtered out attempts where the total login time was much longer than 60 s, indicating the user accidentally left the application open. Nine attempts were filtered in this manner, leaving a total of 1071 valid authentication attempts.

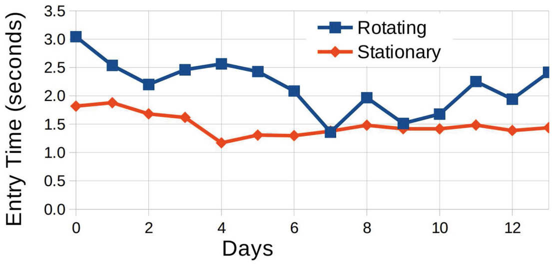


Fig. 5. Median entry time of PassGame users over time ("Days" represents the number of individual calender days the user has made an attempt at authentication.)

We hypothesize that PassHue users will improve in entry time over the course of two weeks as they practice the scheme, and that participants in the rotating condition will perform slightly worse than those in the stationary condition. Figure 5 shows the median entry time in seconds for PassHue users over a 14 day period. Figure 5 is based on *days attempted*, though several days may have elapsed for the user. For example, if the user attempted authentication on days 0, 4, 7, and 12, those attempts would be plotted as day 0, 1, 2, and 3. We organize attempts in this manner because the elapsed days between authentication attempts is not consistent for each user, many users opted to skip days. Participants attempted authentication on an average of 11 separate days.

The overall average time for a single authentication attempt is 2.63 s (median = 2.25, std = 1.99) for rotating PassHue and 1.67 s (median = 1.46, std = .86) for stationary PassHue. As expected, two-tailed Mann-Whitney testing indicates a significant difference between the entry times for the two conditions ($p < .0001$).

The data supports our hypotheses that entry times improve over a short practice period, and that the rotating condition is slower. The average entry time of both conditions is close to in-the-wild entry times reported by other research for traditional 4 digit PINs (1.5 s) and Pattern Unlocks (3.1 s) [27]. PassHue's in-the-wild entry times are superior to average lab entry times for similar PIN-based schemes: ColorPIN – 13.9 s [7], SwiPIN – 3.7 s [25], DOC – 25.7 s [22], and The Phone Lock – 12.2 s [2].

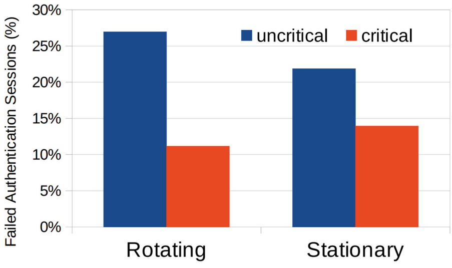


Fig. 6. Authentication sessions with failures

Error Rates. We hypothesize that Rotating users will experience more authentication errors. Figure 6 shows the percentage of authentication sessions that result in critical or uncritical failures. A critical failure is defined as 3 or more incorrect attempts consecutively, as this could traditionally lead to a device lockout. A session is defined as all the authentication attempts in a single instance of using the application. A uncritical failure is 1–2 incorrect attempts in the same session. Users required multiple attempts to authenticate in roughly 35% of authentication sessions in both conditions, but Stationary users face more critical failures. The average number of incorrect authentication attempts per authentication session is .90 and 1.34 for Rotating and Stationary respectively.

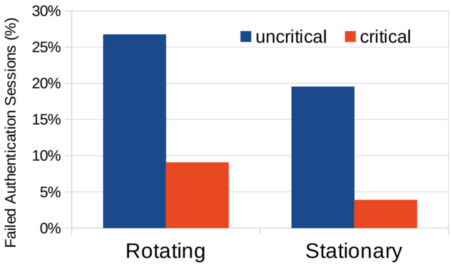


Fig. 7. Authentication sessions with failures (outliers removed)

We noted that most failed attempts originated from a small subset of users, particularly in the Stationary condition. We removed users who made errors more than two standard deviations above the mean error rate as outliers to generate Fig. 7. Even with outliers removed, we find that for most users, the error rate is worse than Pattern Unlock (14.6%, 1.6%) [27]. PassHue would require a high error tolerance before lockout to be viable for most users. With outliers removed, the average number of incorrect authentication attempts per authentication session is .76 and .46 for Rotating and Stationary respectively. The data supports our hypothesis that average Rotating users will make more errors than their Stationary counterparts.

Because PassHue is very fast, the time impact of incorrect authentication attempts is largely insignificant. Using timing results from the previous section, users can expect to spend an average of 1.99 s and .77 s making errors in Rotating and Stationary PassHue respectively. We hypothesize that some users simply preferred to go quickly rather than carefully since there was no punishment for multiple incorrect attempts. PassHue offers a greater chance to trade speed for precision than most discrete authentication methods.

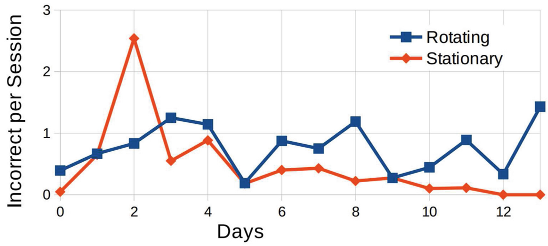


Fig. 8. Failed authentication attempts per session over time (outliers removed)

Figure 8 demonstrates the improvement over time in failed authentication attempts per session. A clear trend emerges in the Stationary condition, demonstrating that PassHue users become significantly less error prone after just 4 days of use, with diminishing gains in accuracy after one week. We hypothesize that Rotating users do not share the training effect because they do not have a chance to build muscle memory due to the color wheel being in a different position each time.

5.5 Color Selection and Hotspots

Figure 9 shows all colors selected by our participants, grouped roughly in ascending RGB order. We did not find any impact on color selection by condition or color order. Cyan, violet, and white hues are slightly under-represented, while

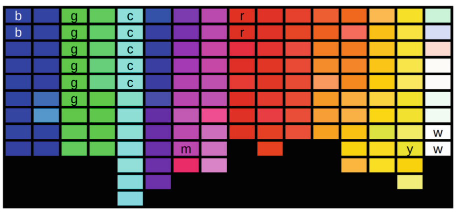


Fig. 9. Colors selected by participants (Colors, within $\Delta R + \Delta B + \Delta G \leq 10$ of the true value of that color, e.g., 0, 255, 255 for true cyan, are marked for blue, green, cyan, magenta, red, yellow, and white.) (Color figure online)

hues between yellow and red, are slightly over-represented. The mean RGB values are ($\bar{R} = 161$, $stdev = 116$; $\bar{G} = 133$, $stdev = 99$; $\bar{B} = 131$, $stdev = 112$). The expected mean assuming an even distribution is 128, indicating red is slightly over-represented.

A brute force attacker may gain some advantage guessing shades of red, orange, and yellow first. From Fig. 9, we note that roughly 40% of all color choices fall between true red (255, 0, 0) and true yellow (255, 255, 0), despite this section making up only one sixth (17%) of the color wheel. In other words, red-orange-yellow is selected roughly twice as frequently as expected. Only one participant (3%) relied on these colors exclusively.

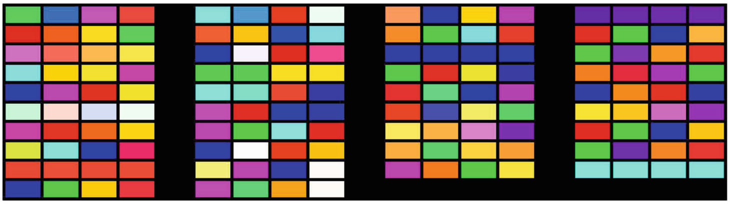


Fig. 10. All PassHues chosen by participants (Color figure online)

Figure 10 shows all the PassHues chosen by participants in our user study, each row represents one participant. Four participants (11%) generated a PassHue using the same color 4 times. By same, we mean that all 4 colors had a difference score of less than 10 based on Eq. 1. This is very similar to 4-digit PINs, where roughly 8% of PINs are comprised of 4 duplicated digits [14]. Two more participants (5%) generated a PassHue using the same color twice.

Notably absent in the data are repeating patterns such as couplets in the form XYXY, which comprise approximately 18% of PINs. Number based patterns, such as the years 1951–2000 (accounting for roughly 6% of all PINs [14]) are impossible in PassHue. We conclude that PassHue may have a substantial advantage in encouraging good password choices, simply because there are relatively few commonly occurring patterns based on color. In our future work, we plan to include more participants in order to determine if any patterns emerge in password selection.

6 Discussion

6.1 Color Blindness and Tetrachromacy

We hypothesized that some people with minor color blindness would still be able to use PassHue if they avoided colors that were difficult for them, by using some sort of relative position on the wheel, or by using only the most intense colors. In general, these techniques may not reduce the security strength of PassHue, since the attacker is unlikely to know if the user is color blind, and if they do, it is nearly impossible for the attacker to know *how* color blind the user is. Figure 11 demonstrates what the PassHue wheel would look like to someone with minor green color blindness (Deuteranomaly), the most common form of color blindness. Subjectively, it appears to be still be usable, and distinct line patterns are now clearly visible.

Fig. 11. The PassHue wheel seen with minor deuteranomaly (Color figure online)

As part of our demographic information, we asked participants to tell us if they were color blind, and if so, what type of blindness they had. One participant reported that they were red-green color blind. This participant was assigned to the Stationary condition. Password creation time was 50 s, and average entry time was 2.42 s, slightly below average. The average number of incorrect attempts per authentication session was 1.83, again slightly worse than average.

However, we note that entry time for this participant in the first 3 days was 3.62 s, declining to 1.9 s over the latter 11 days. Likewise, we note that the error rate for this participant in the first 3 days was 4.38 incorrect attempts per session, declining to .88 incorrect attempts per session for the latter 11 days of the experiment. We conclude that the participant was able to find a way to use PassHue despite not having perfect color vision.

Fig. 12. Color-blind participant's PassHue (Color figure online)

Figure 12 shows the PassHue for this participant. Notably, the colors are less saturated and relatively far from true color values for red, blue, yellow, and green. The participant gave the following response on their exit survey:

"Since I know I could never remember the colour I set my password according to a colour sequence which consists of the most obvious colours from each 4 groups namely red, blue, yellow, and green. Since in blue and yellow groups I can see very distinct lines of that colour I use them in the password sequence. The rest is muscle memory."

From this statement we hypothesize that color blind people in general will tend to pick colors that are very different from each other, since they have a harder time discerning similar colors. We plan to investigate this hypothesis with more color blind participants in our future work.

We conclude that PassHue, despite being based on the color continuum, can actually be used by someone with limited ability to discern parts of the continuum. The user can effectively substitute memorizing color with memorizing x-y location relative to some color or pattern hint.

Although roughly 8% of males of .5% of females suffer from color blindness and may not be able to use PassHue effectively, it is also estimated that 1% of the population has tetrachromatic vision, allowing them to see additional colors. By including more colors from the types that tetrachromes can better discern, a version of PassHue can be developed that is highly secure against anyone that doesn't have tetrachromatic vision. This system would require a tetrachromatic display, and finding participants with tetrachromatic vision is difficult.

6.2 Gender Bias

Because most kinds of color blindness occur more frequently men, and because research suggests that on average women remember color more easily and accurately than men [18], PassHue may be slightly biased towards women. We hypothesized that females on average would be faster and more accurate when entering their PassHue.

Surprisingly, we found that women were slightly slower at entering their PassHue, with average entry times of 2.69 and 2.44 s vs male average entry times of 2.53 and 1.68 s for Rotating and Stationary respectively. Mann-Whitney testing found a significant difference in entry times between genders for both conditions ($p = .006$ and $p = .016$ for Rotating and Stationary respectively). The difference in entry time may be due to motor proficiency, where studies have found men to have some advantage [19].

However, the data supported our hypothesis that women would be more accurate on average. For Rotating and Stationary respectively, women made an average of 0.5 and 0.24 incorrect authentication attempts per authentication session vs male error rates of 0.87 and 1.55 incorrect attempts per session ($\chi^2 = 5.27, p = .022; \chi^2 = 35.21, p \leq .00001$). Additionally, the lowest performers, discussed in Sect. 5.4, were all male.

The data shows that PassHue is generally suitable for both genders, but the error rate may be quite high for a small subset of males. This may suggest that tolerances for male users should be slightly higher by default.

6.3 Inclusion of Additional Colors

Adding a "Value" slider to PassHue would greatly increase the password space as it would allow use of the entire RGB color space. Alternatively, there are also 2d images containing the entire RGB color space. The size of C would become all of RGB color space, $(256^3) = 16.8$ million colors. Even assuming the worst-case tolerance of 39,000, the password space is $(256^3/39000)^4 = 3.4 * 10^{10}$, about the same as a 6-character case-sensitive alphanumeric password with no symbols ($62^6 = 5.8 * 10^{10}$). We sample 40 million random color pairs and find that the average product of distances is 15,000. Using this value as our tolerance, the password space is $(256^3/15000)^4 = 1.5 * 10^{12}$, roughly on-par with a 7-character alphanumeric password ($62^7 = 3.5 * 10^{12}$). In our future work, we plan to see if it is feasible to extend PassHue to the entire RGB color space without impacting usability.

6.4 Shoulder-Surfing Resistance

We hypothesize that PassHue will be significantly more shoulder-surfing resistant in the Rotating condition because the attacker cannot use position as a cue and must directly memorize the colors. The rotating condition should also be resistant to smudge attackers, since the wheel is not in the same orientation for different attempts, though a smudge can still give the attacker the relative spatial relationship between colors. A formal user study of PassHue's shoulder-surfing resistance is a plan for future work. PassHue is inherently resistant to black-and-white or poor resolution camera attacks, especially in the rotating condition.

7 Conclusion

In this paper, we introduced the concept of *Analog Authentication*, using continuous rather than discrete information for authentication purposes. PassHue, a proof-of-concept for analog authentication, is able to achieve a much higher security strength than PIN or Pattern Unlock. Our in-the-wild user study demonstrates that PassHue is highly memorable, PassHue is on-par or faster than existing authentication schemes, and PassHue is not significantly prone to hotspots. We conclude that PassHue, and by extension analog authentication, are viable avenues of exploration in finding a new method for mainstream mobile authentication.

References

1. Aviv, A.J., Gibson, K.L., Mossop, E., Blaze, M., Smith, J.M.: Smudge attacks on smartphone touch screens. In: USENIX 4th Workshop on Offensive Technologies (Woot), vol. 10, pp. 1–7 (2010)
2. Bianchi, A., Oakley, I., Kostakos, V., Kwon, D.S.: The phone lock: audio and haptic shoulder-surfing resistant pin entry methods for mobile devices. In: Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction, pp. 197–200. ACM (2011)
3. Bianchi, A., Oakley, I., Kwon, D.S.: Counting clicks and beeps: exploring numerosity based haptic and audio pin entry. Interact. Comput. **24**(5), 409–422 (2012)

4. Clark, G.D., Lindqvist, J.: Engineering gesture-based authentication systems. IEEE Pervasive Comput. **14**(1), 18–25 (2015)
5. CompuPhase: Color metric, April 2012. https://www.compuphase.com/cmetric.htm
6. Cowan, N.: The magical mystery four: how is working memory capacity limited, and why? Curr. Dir. Psychol. Sci. **19**(1), 51–57 (2010)
7. De Luca, A., Hertzschuch, K., Hussmann, H.: Colorpin: securing pin entry through indirect input. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1103–1106. ACM (2010)
8. Egelman, S., Jain, S., Portnoff, R.S., Liao, K., Consolvo, S., Wagner, D.: Are you ready to lock? In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 750–761. ACM (2014)
9. Halsey, R.M., Chapanis, A.: Chromaticity-confusion contours in a complex viewing situation. JOSA **44**(6), 442–454 (1954)
10. Hamwi, V., Landis, C.: Memory for color. J. Psychol. **39**(1), 183–194 (1955)
11. Harbach, M., Von Zezschwitz, E., Fichtner, A., De Luca, A., Smith, M.: It's a hard lock life: a field study of smartphone (un)locking behavior and risk perception. In: Symposium on Usable Privacy and Security (SOUPS), pp. 9–11 (2014)
12. Huh, J.H., Kim, H., Bobba, R.B., Bashir, M.N., Beznosov, K.: On the memorability of system-generated pins: can chunking help? In: Symposium on Usable Privacy and Security (SOUPS), pp. 197–209 (2015)
13. Judd, D.B., Wyszecki, G.: Color in Business. Science and Industry. Wiley, New York (1963)
14. Kim, H., Huh, J.H.: Pin selection policies: are they really effective? Comput. Secur. **31**(4), 484–496 (2012)
15. Kwon, T., Na, S.: Tinylock: affordable defense against smudge attacks on smartphone pattern lock systems. Comput. Secur. **42**, 137–150 (2014)
16. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81–97 (1956)
17. Olmstead, K., Smith, A.: Americans and cybersecurity, January 2017. http://www.pewinternet.org/2017/01/26/2-password-management-and-mobile-security/
18. Pérez-Carpinell, J., Baldoví, R., de Fez, M.D., Castro, J.: Color memory matching: time effect and other factors. Color Res. Appl. **23**(4), 234–247 (1998)
19. Plimpton, C.E., Regimbal, C.: Differences in motor proficiency according to gender and race. Percept. Mot. Skills **74**(2), 399–402 (1992)
20. Pointer, M.R., Attridge, G.G.: The number of discernible colours. Color Res. Appl. **23**(1), 52–54 (1998)
21. RealUser: Passfaces: Two factor authentication for the enterprise, October 2017. http://www.realuser.com
22. Roth, V., Richter, K., Freidinger, R.: A pin-entry method resilient against shoulder surfing. In: Proceedings of the 11th ACM Conference on Computer and Communications Security, pp. 236–245. ACM (2004)
23. Sherman, M., Clark, G., Yang, Y., Sugrim, S., Modig, A., Lindqvist, J., Oulasvirta, A., Roos, T.: User-generated free-form gestures for authentication: security and memorability. In Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, pp. 176–189. ACM (2014)
24. Uellenbeck, S., Dürmuth, M., Wolf, C., Holz, T.: Quantifying the security of graphical passwords: the case of android unlock patterns. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security, pp. 161–172. ACM (2013)

25. Von Zezschwitz, E., De Luca, A., Brunkow, B., Hussmann, H.: Swipin: fast and secure pin-entry on smartphones. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1403–1406. ACM (2015)
26. Von Zezschwitz, E., De Luca, A., Hussmann, H.: Honey, i shrunk the keys: influences of mobile devices on password composition and authentication performance. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction, pp. 461–470. ACM (2014)
27. Von Zezschwitz, E., Dunphy, P., De Luca, A.: Patterns in the wild: a field study of the usability of pattern and pin-based authentication on mobile devices. In: Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 261–270. ACM (2013)

Check for updates

A Hand Gesture-Based Method for Biometric Authentication

Satoru Imura and Hiroshi Hosobe(✉)

Faculty of Computer and Information Sciences, Hosei University, Tokyo, Japan
hosobe@acm.org

Abstract. With the spread of computers to ordinary people, computer security is becoming increasingly important. User authentication is one of the most important technologies for computer security. Although passwords are used in many personal computers, they are known to sometimes have problems. As an alternative to passwords, biometric authentication, such as fingerprint authentication and face recognition-based authentication, is becoming more widely used. In this paper, we propose a hand gesture-based method as a new kind of biometric authentication. It supports three-dimensional (3D) gestures that allow its user to move the user's hand without touching an input device. Using the motions of fingertips and finger joints as biometric data, the method improves the performance of authentication. Also, we propose seven 3D gestures that can be classified into three types. We implemented the method by using a 3D motion sensor called the Leap Motion controller. We present the results of an experiment that we conducted with nine participants to evaluate the method. For all the gestures, the true acceptance rates were more than 90%, and the equal error rates were less than 4%.

Keywords: Biometric authentication · Hand gesture · Motion sensor

1 Introduction

With the spread of computers to ordinary people, computer security is becoming increasingly important. User authentication is one of the most important technologies for computer security. Although passwords are used in many personal computers, they are known to sometimes have problems. For example, users sometimes employ passwords that are easy to memorize, which might enable dictionary attacks to succeed. Also, passwords might be stolen, e.g., by shoulder surfing.

As an alternative to passwords, *biometric authentication* [11,21,22] is becoming more widely used. In general, biometric authentication adopts physiological/behavioral characteristics of users (e.g., fingerprints, faces, irises, palm veins, pen pressures, keystrokes, and gaits) for user authentication. Such characteristics are contrary to passwords that users might forget and that other people might steal. Also, since such characteristics are unique to users, there are no common

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 554–566, 2018.
https://doi.org/10.1007/978-3-319-91238-7_43

Fig. 1. Hand gesture-based biometric authentication

patterns like easy-to-memorize passwords. In particular, fingerprint authentication is widely used for mobile devices such as smartphones and tablets. Also, face recognition-based authentication is lately becoming popular.

In this paper, we propose a hand gesture-based method as a new kind of biometric authentication. It supports three-dimensional (3D) gestures that allow its user to move the user's hand without touching an input device, as shown in Fig. 1. Using the motions of fingertips and finger joints as biometric data, the method improves the performance of authentication. Also, we propose seven 3D gestures that can be classified into three types. We implemented the method by using a 3D motion sensor called the Leap Motion controller [14]. We present the results of an experiment that we conducted with nine participants to evaluate the method. For all the gestures, the true acceptance rates were more than 90%, and the equal error rates were less than 4%.

This paper is an extended version of the poster paper that we previously published as [9]. In particular, this extended paper describes how the proposed method represents 3D hand gestures (Subsect. 4.2), how it registers template gestures (Subsect. 4.3), and how it performs gesture authentication (Subsect. 4.4).

The rest of this paper is organized as follows. After presenting related work in Sect. 2, we briefly describe the Leap Motion controller and performance measures for biometric authentication in Sect. 3. Then, in Sect. 4, we propose our biometric authentication method, and present its implementation in Sect. 5. We report the experiment that we performed to evaluate our method in Sect. 6, and discuss the method in Sect. 7. Finally, we describe conclusions and future work in Sect. 8.

2 Related Work

Many existing personal computers and mobile devices use text-based passwords and similar symbolic sequences such as PIN codes for user authentication. However, some users employ vulnerable passwords to easily memorize or enter the passwords. Google's Android introduced pattern lock [18], which allows users to draw patterns by connecting dots on touch screens instead of entering text passwords. However, since Android's pattern lock uses only a small number of dots, they are essentially almost as simple as passwords. In addition, Ye et al. [26] showed the possibility of attacking pattern lock; they were able to infer correct patterns by analyzing videos for the motions of fingertips even if the videos had not directly captured the screens.

As an alternative to passwords, there has been research on biometric authentication [11,21,22] employing characteristics of users. Methods for biometric authentication can be roughly divided into two categories, the first one using physiological characteristics and the second one using behavioral characteristics. Physiological characteristics include, for example, fingerprints, faces, irises, and palm veins. In particular, fingerprint authentication is widely used in smartphones and tablets, and face recognition-based authentication is lately becoming popular. There has also been research on the use of the physiological characteristics of hands for biometric authentication [2]. For example, Jain et al. [10] showed the possibility of identifying persons by the geometries of hands.

The method that we propose in this paper falls in the second category of biometric authentication. Behavioral characteristics used in this category include, for example, pen pressures, keystrokes, and gaits [11,15,21,22,25]. Kholmatov and Yanikoglu [12] proposed an online signature verification method that used not only the form of a signature but also the number and the order of the strokes and the velocity and the pressure of the pen. Kim et al. [13] used the pressures of fingertips as behavioral characteristics, in particular, to solve the shoulder surfing problem in the context of collaborative tabletop interfaces. Ataş [1] proposed a biometric authentication method using the tremors of hands, and implemented it by using the Leap Motion controller.

Our biometric authentication method uses hand gestures as behavioral characteristics. Such methods can be categorized into two, one using two-dimensional (2D) hand gestures, and the other using three-dimensional (3D) hand gestures [5]. Methods using 2D hand gestures for biometric authentication typically employ touch screens. Sae-Bae et al. [16] proposed multi-touch gestures using a thumb and four fingers on a multi-touch screen, and studied particular 22 gestures. Sherman et al. [19] studied the use of free-form gestures for authentication.

Biometric authentication methods using 3D hand gestures can be further categorized into sensor- and camera-based methods [5]. Sensor-based methods typically use accelerometers. Guerra-Casanova et al. [8] proposed the use of an accelerometer-embedded mobile device to capture a 3D hand gesture like an in-air signature. Sun et al. [20] developed a biometric authentication system for smartphones that used on-phone accelerometers to capture 3D hand gesture signatures. It should be noted that such sensor-based methods do not consider the motions of fingers.

Camera-based biometric authentication methods perform image processing. Fong et al. [7] used the stationary images of 3D hand gestures that represented sequences of alphabets in a hand sign language.

Recently, there has been research on camera-based methods using the Leap Motion controllers. Chan et al. [3] used the Leap Motion controller for biometric authentication using a circle-drawing gesture with one finger. Saritha et al. [17] also used the Leap Motion controller to treat circle-drawing, swipe, screen-tap, and key-tap gestures. Xiao et al. [24] used the Leap Motion controller to handle 3D hand gesture signatures. Wong and Kang [23] proposed stationary hand gestures that used the free motions of fingers without the motions of hands, wrists, and arms; they implemented a biometric authentication method by using the Leap Motion controller.

In some sense, these camera-based methods using the Leap Motion controllers share the same motivation as our method. However, our method is more focused on the use of a wider variety of 3D hand gestures. The 3D hand gestures used by Chan et al. [3], Saritha et al. [17], and Xiao et al. [24] were relatively simple because they did not utilize gestures with the complex motions of fingers. Although stationary hand gestures proposed by Wong and Kang [23] were more complex, they still did not use the motions of hands. In this sense, our method is closer to the one proposed by Sae-Bae et al. [16] that defined and studied 22 2D hand gestures in the context of a multi-touch screen.

3 Preliminaries

This section briefly describes the Leap Motion controller and the performance measures for biometric authentication that we use in our work.

3.1 Leap Motion Controller

The Leap Motion controller [14] is a motion sensor that captures the positions of hands and fingers in a 3D space by using two infrared cameras and an infrared LED. It has an effective range of 25 to 600 mm upward, a field of view of 150° in the side direction and of 120° in the depth direction, a tracking rate of 120 fps, and a tracking accuracy of 1/100 mm. It internally has a standard model of human hands, and treats various motions of hands by associating sensor data with the internal model. However, it cannot accurately sense fingers that are hidden by other fingers or by palms. It is a small device of a size of $80 \times 30 \times 11$ mm, and is as low as about 100 US dollars. It is easily available, and even laptop computers with built-in Leap Motion controllers are being sold.

3.2 Performance Measures for Biometric Authentication

There are various measures for evaluating the performance of biometric authentication [4,6,11]. In this paper, we use two performance measures. The first

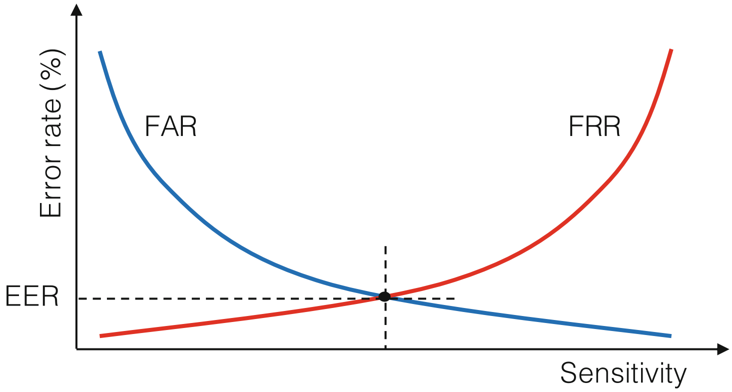


Fig. 2. The false rejection rate (FRR), the false acceptance rate (FAR), and the equal error rate (ERR)

measure is the *true acceptance rate* (TAR, also known as the true match rate). This is a simple measure that indicates the rate of how often an input is correctly accepted.

The second measure is the *equal error rate* (EER). This is a measure commonly used to more precisely evaluate the performance of an authentication method. The EER is computed from the false rejection rate (FRR, also known as the false nonmatch rate)[1] and the false acceptance rate (FAR, also known as the false match rate) by considering an appropriate threshold. The EER is defined as the rate at which the FRR and the FAR cross (Fig. 2). This is based on the following: a tighter threshold makes the authentication more sensitive by more often causing false rejection (i.e., rejecting a right input); a looser threshold makes the authentication less sensitive by more often causing false acceptance (i.e., accepting a wrong input). Thus there is a trade-off between the FRR and the FAR, and the EER is the error rate corresponding to the best threshold.

4 Proposed Method

This section proposes a hand gesture-based method for biometric authentication.

4.1 3D Hand Gestures

Our method executes biometric authentication using the geometries and the 3D gestures of a hand. It adopts a 3D motion sensor to obtain the geometry and the position of the hand in real time (Fig. 1), from which it computes the similarities of the input gesture with the template gestures stored in its database beforehand.

[1] It should be noted that $FRR = 1 - TAR$ holds.

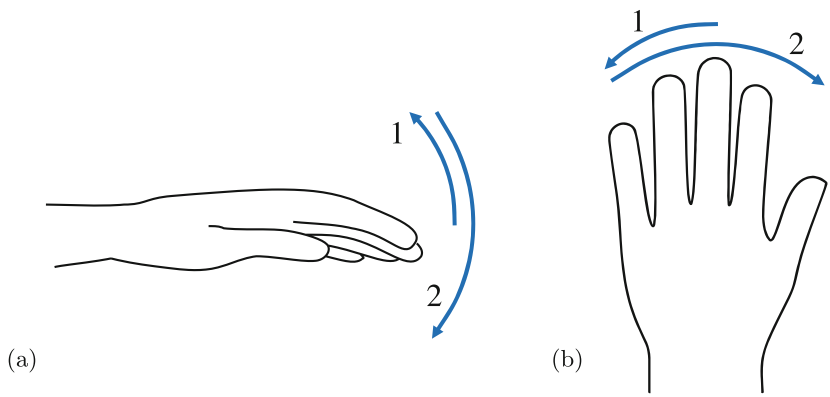


Fig. 3. 3D hand gestures called (a) WUD and (b) WLR

We propose the following three types of 3D hand gestures, namely, the fingertip, the wrist, and the complex type.[2]

The fingertip type defines gestures that start from an open posture of fingers and then move fingers in certain directions. It consists of the following two gestures.

FCO: All fingertips close and then open.

FBO: Fingertips bend and extend one after another.

The wrist type defines gestures that move the hand in certain directions with the position of the wrist fixed. It consists of the following four gestures.

WUD: The wrist is bended up and down (Fig. 3(a)).

WLR: The wrist is bended to the left and to the right (Fig. 3(b)).

WCWY: The hand and the forearm are rotated clockwise with the wrist as its center and along the vertical axis (which is parallel to the y-axis shown in Fig. 4).

WTR: The hand is turned over and back.

The complex type consists of user-defined gestures.

UDS: The user writes the user's signature in the air.

We use the Leap Motion controller for 3D motion sensing. Users position their hands about 20 cm above the Leap Motion controller as shown in Fig. 4. In our experiment that we report in Sect. 6, the users used the left hands for the gestures other than UDS. For UDS, they used their dominant hands since they needed to write their signatures.

[2] Initially, we included another gesture type called "hand" that consisted of four gestures such as moving the left hand up and down. However, we found that this type of gestures was difficult for users to precisely perform again because they needed to move their arms as well as their hands. Therefore, we excluded this type of gestures.

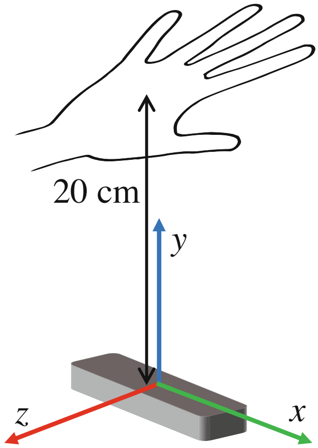


Fig. 4. Setting of the Leap Motion controller for a hand

4.2 Representation of 3D Hand Gestures

Our method internally treats a 3D hand gesture as a time-based sequence of multi-dimensional vectors. For a posture of a hand at a particular time point, it considers the 3D positions of the joints of the fingers as well as the 3D positions of the fingertips. More specifically, a posture of a hand at a time point is represented as a 75-dimensional vector that can be divided into five 15-dimensional vectors for the thumb and the four fingers.

Figure 5 shows the parts of a hand that we use for biometric authentication. The motion sensor measures the 3D positions of the fingertips and the joints of the thumb and the four fingers. For each of the four fingers, there are four joints at the ends of the bones called a distal phalange, an intermediate phalange, and a proximal phalange, and a metacarpal. For the thumb, there are three joints because one bone is missing. In the case of the Leap Motion controller, it internally generates an extra joint for the thumb by additionally considering a bone of a zero length.

Therefore, we can represent each of the thumb and the four fingers at a time point t as the following 15-dimensional vector:

$$\mathbf{f}_{i,t} = (x_{i,t}^{\text{tip}}, y_{i,t}^{\text{tip}}, z_{i,t}^{\text{tip}}, x_{i,t}^{\text{dis}}, y_{i,t}^{\text{dis}}, z_{i,t}^{\text{dis}}, x_{i,t}^{\text{int}}, y_{i,t}^{\text{int}}, z_{i,t}^{\text{int}}, x_{i,t}^{\text{pro}}, y_{i,t}^{\text{pro}}, z_{i,t}^{\text{pro}}, x_{i,t}^{\text{met}}, y_{i,t}^{\text{met}}, z_{i,t}^{\text{met}}),$$

where $i \in \{1, 2, 3, 4, 5\}$ indicates the thumb or one of the four fingers, and each component of $\mathbf{f}_{i,t}$ represents a coordinate of the thumb or one of the four fingers. Using these vectors, we represent a hand posture at t as the following 75-dimensional vector:

$$\mathbf{h}_t = (\mathbf{f}_{1,t}, \mathbf{f}_{2,t}, \mathbf{f}_{3,t}, \mathbf{f}_{4,t}, \mathbf{f}_{5,t}).$$

Finally, we represent a gesture as the following time-based sequence G of hand postures from time point 1 to T:

$$G = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T].$$

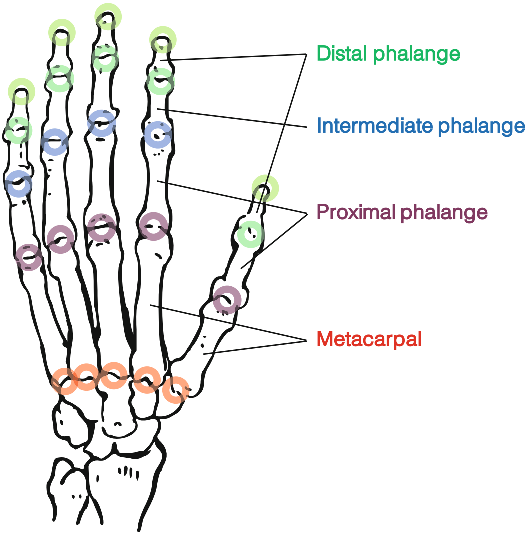


Fig. 5. Parts of a hand used for biometric authentication

We use such a sequence G of 75-dimensional vectors for both an *input gesture* and a *template gesture*. An input gesture is the data that is obtained by measuring a user's actually performing a certain 3D hand gesture. A template gesture is the data that is associated with a certain 3D hand gesture of a user and that is stored in a database.

4.3 Registration of Template Gestures

Before gesture authentication, we need to construct a database of template gestures. Since it is often difficult for a user to re-perform a sufficiently similar 3D hand gesture, we use the average of multiple input gestures to obtain a template gesture. Specifically, our method requires the user to repeat a certain gesture ten times to register its template gesture on the database. The average of the ten input gestures is calculated as a template gesture, and then is stored in the database.

We also associate a *preset threshold* with a template gesture. We use it to judge whether an input gesture is sufficiently similar to the template gesture associated with the preset threshold. We obtain the preset threshold of a template gesture by computing the average similarities between this template gesture and the ten used input gestures, where the similarities are calculated as presented in the next subsection. We use such preset thresholds to compute true acceptance rates in our experiment.

4.4 Gesture Authentication

Given an input gesture, our method of gesture authentication reports whether it accepts or rejects the input. This process is performed as follows.

1. Find the template gesture that is the most similar to the input gesture.
2. Do the following:
   (a) If the similarity between the input gesture and the found template gesture is smaller than the preset threshold associated with the template gesture, it accepts the input gesture;
   (b) Otherwise, it rejects the input gesture.

To compute the similarity between an input and a template gesture, our method calculates the sum of the Euclidean distances between the corresponding 75-dimensional vectors from the sequence G_{inp} of the input gesture and the sequence G_{tem} of the template gesture. Formally, with $G_{\text{inp}} = [\mathbf{h}_1^{\text{inp}}, \mathbf{h}_2^{\text{inp}}, \dots, \mathbf{h}_T^{\text{inp}}]$ and $G_{\text{tem}} = [\mathbf{h}_1^{\text{tem}}, \mathbf{h}_2^{\text{tem}}, \dots, \mathbf{h}_T^{\text{tem}}]$, the similarity $S(G_{\text{inp}}, G_{\text{tem}})$ between G_{inp} and G_{tem} is defined as follows:

$$S(G_{\text{inp}}, G_{\text{tem}}) = \sum_{t=1}^{T} d(\mathbf{h}_t^{\text{inp}}, \mathbf{h}_t^{\text{tem}}),$$

where each $d(\mathbf{h}_t^{\text{inp}}, \mathbf{h}_t^{\text{tem}})$ is the Euclidean distance between 75-dimensional vectors $\mathbf{h}_t^{\text{inp}}$ and $\mathbf{h}_t^{\text{tem}}$.

The simple calculation of this similarity is usually impossible due to the different lengths of G_{inp} and G_{tem}, which occurs because their lengths depend on the time lengths of the gestures performed by the user. To solve this problem, we extend the shorter sequence to the length of the longer one. We use the following simple method: we first determine the appropriate locations of the shorter sequence where new 75-dimensional vectors should be inserted; then we insert each of the new vectors that we obtain by calculating the midpoint of its previous and next vectors. For example, consider two sequences of lengths 103 and 100. Then we extend the sequence of length 100 by adding new three vectors. Let this sequence be $G = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{100}]$. We insert new three vectors between $\mathbf{h}_{25}$ and $\mathbf{h}_{26}$ and between $\mathbf{h}_{50}$ and $\mathbf{h}_{51}$ and between $\mathbf{h}_{75}$ and $\mathbf{h}_{76}$. We compute the new three vectors as $(\mathbf{h}_{25} + \mathbf{h}_{26})/2$, $(\mathbf{h}_{50} + \mathbf{h}_{51})/2$, and $(\mathbf{h}_{75} + \mathbf{h}_{76})/2$.

5 Implementation

We developed a prototype hand gesture-based biometric authentication system based on the method proposed in the previous section. We used the Leap Motion controller for 3D motion sensing. We implemented the system in Java using the Leap Motion SDK. It consists of approximately 3300 lines of code.

6 Experiment

This section reports the experiment that we conducted to evaluate the proposed method.

Table 1. Results of the experiment

Type	Gesture	TAR (%)	EER (%)
Fingertip	FCO	90.4	3.8
	FBO	90.6	0.0
Wrist	WUD	94.0	2.4
	WLR	97.6	2.9
	WCWY	94.1	2.9
	WTR	93.0	2.3
Complex	UDS	100.0	0.0

6.1 Procedure

We recruited nine participants who all were male and who were 21.8 years old on average.[3] We asked them to use their left hands to perform the gestures other than UDS. In the case of UDS, the participants used their dominant hands since they needed to write their signatures. The participant pressed a key with the other hand when he started and finished a gesture.

We first constructed a database of template gestures. First, the participants practiced each gesture several times before registering its template. For the registration, each participant performed each of the seven gestures ten times (i.e., 70 times in total).

After all the participants registered their template gestures on the database, we conducted an experiment on gesture authentication. In the experiment, each participant performed each of the seven gestures ten times again. Every time after the participant performed a gesture, the system notified him whether it was accepted or rejected. For this purpose, the preset thresholds were used.

6.2 Results

To show the performance of the proposed method, we use two measures TARs and EERs that we explained in Subsect. 3.2. Table 1 shows the resulting TARs and EERs for the seven gestures. The TARs were more than 90% for all the gestures, and the average of the TARs for all the gestures was 94.2%. The averages of the TARs for the fingertip, the wrist, and the complex type were 90.5%, 94.7%, and 100% respectively.

We computed the EERs for the seven gestures by finding appropriate thresholds. All the EERs resulted in smaller than 4%, and the average of the EERs for all the gestures was 2.0%. The averages of the EERs for the fingertip, the wrist, and the complex type were 1.9%, 2.6%, and 0.0% respectively.

[3] Since we wanted a stricter experimental setting, we did not recruit female participants; otherwise, the larger variance of the geometries of the participants' hands could have more affected the experimental results.

7 Discussion

The results of the experiment indicate that the proposed method almost always distinguished the gestures of the participants. It should be emphasized that, in the cases of the fingertip and the wrist type, the participants performed the same 3D hand gestures. This suggests that it is difficult for users to imitate the gestures of other users in such a way that the method cannot distinguish them, which is an ideal property for biometric authentication. It also should be noted that, unlike passwords, 3D hand gestures are not vulnerable to shoulder surfing because of the difficulty of gesture imitation.

Many of the participants gave comments that they had suffered hand fatigue because it had been time-consuming for them to position their hands about 20 cm above the motion sensor. To solve this problem, we need to implement the facility that automatically adjusts the vertical positions of input gestures. Also, it will be better for us to implement the facility that automatically adjusts the angles of input gestures, which can be expected to further improve the performance by reducing errors caused by hand positions.

To compute the similarity between an input and a template gesture, we used a simple definition of the similarity based on the sum of Euclidean distances between time-based sequences of 75-dimensional vectors. Also, we used a simple method for treating an input and a template gesture with different lengths. However, we can consider other alternatives. For example, we could use a distance between two high-dimensional vectors by treating a time-based sequence of 75-dimensional vectors as a single high-dimensional vector. Alternatively, we could use a cosine distance instead of an Euclidean distance. Also, we could use a more sophisticated interpolation method to handle an input and a template gesture with different lengths. We need to explore such alternatives to improve the performance of the gesture-based authentication.

As an experiment for evaluating biometric authentication, our experiment was small in the number of participants. However, the results of the experiment indicated that our method is insufficient than state-of-the-art methods such as fingerprint- and iris-based ones that obtain EERs of less than 0.1%. Therefore, before conducting a larger experiment, we need to improve the performance of our method by doing things described above. For this purpose, there are also other directions such as the combination of multiple gestures for one trial of authentication.

8 Conclusions and Future Work

We proposed a new biometric authentication method based on 3D hand gestures. We used, as biometric data, timed-based sequences of 3D positions of fingertips and finger joints. Also, we proposed seven 3D hand gestures that are classified into three types. We implemented the method by using the Leap Motion controller as a 3D motion sensor. To evaluate it, we conducted an experiment of gesture authentication with nine participants. As a result, for all the gestures, the TARs were more than 90%, and the EERs were less than 4%. This indicates that, even if different users perform the same gestures, the method can almost always distinguish such gestures. Also, this suggests that, even if a gesture is imitated by another person, the method is not likely to accept it.

Compared to state-of-the-art authentication methods such as fingerprint- and iris-based ones, our method still has room for improvement in performance. For example, we need to implement the facility that automatically adjusts the vertical positions and the angles of input gestures. Also, we could improve the performance by combining multiple gestures for one trial of authentication. We believe that we should pursue a biometric authentication method that is easy-to-use for ordinary people.

Acknowledgment. This work was partly supported by JSPS KAKENHI Grant Number JP15KK0016.

References

1. Ataş, M.: Hand tremor based biometric recognition using Leap Motion device. IEEE Access **5**, 23320–23326 (2017)
2. Bača, M., Grd, P., Fotak, T.: Basic principles and trends in hand geometry and hand shape biometrics. In: New Trends and Developments in Biometrics, pp. 77–99. InTech (2012)
3. Chan, A., Halevi, T., Memon, N.: Leap motion controller for authentication via hand geometry and gestures. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2015. LNCS, vol. 9190, pp. 13–22. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20376-8_2
4. Cherifi, F., Hemery, B., Giot, R., Pasquet, M., Rosenberger, C.: Performance evaluation of behavioral biometric systems. In: Behavioral Biometrics for Human Identification: Intelligent Applications, pp. 57–74. IGI Global (2009)
5. Clark, G.D., Lindqvist, J.: Engineering gesture-based authentication systems. IEEE Pervasive Comput. **14**(1), 18–25 (2015)
6. El-Abed, M., Charrier, C.: Evaluation of biometric systems. In: New Trends and Developments in Biometrics, pp. 149–169. InTech (2012)
7. Fong, S., Zhuang, Y., Fister, I., Fister Jr., I.: A biometric authentication model using hand gesture images. Biomed. Eng. Online **12**(111), 1–18 (2013)
8. Guerra-Casanova, J., Sánchez-Ávila, C., Bailador, G., de Santos Sierra, A.: Authentication in mobile devices through hand gesture recognition. Int. J. Inf Secur. **11**(2), 65–83 (2012)
9. Imura, S., Hosobe, H.: Biometric authentication using the motion of a hand (poster). In: Proceedings of ACM SUI, p. 221 (2016)
10. Jain, A.K., Ross, A., Prabhakar, S.: A prototype hand geometry-based verification system. In: Proceedings of International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA), pp. 166–171 (1999)
11. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circ. Syst. Video Technol. **14**(1), 4–20 (2004)
12. Kholmatov, A., Yanikoglu, B.: Identity authentication using improved online signature verification method. Pattern Recogn. Lett. **26**(15), 2400–2408 (2005)

13. Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J., Nicholson, J., Olivier, P.: Multi-touch authentication on tabletops. In: Proceedings of ACM CHI, pp. 1093–1102 (2010)
14. Leap Motion. Leap Motion for Mac and PC. https://www.leapmotion.com/product/desktop/
15. Mahfouza, A., Mahmouda, T.M., Eldinc, A.S.: A survey on behavioral biometric authentication on smartphones. J. Inf. Secur. Appl. **37**, 28–37 (2017)
16. Sae-Bae, N., Ahmed, K., Isbister, K., Memon, N.: Biometric-rich gestures: a novel approach to authentication on multi-touch devices. In: Proceedings of ACM CHI, pp. 977–986 (2012)
17. Saritha, L.R., Thomas, D., Mohandas, N., Ramnath, P.: Behavioral biometric authentication using Leap Motion sensor. Int. J. Latest Trends Eng. Technol. **8**(1), 643–649 (2017)
18. Shabtai, A., Fledel, Y., Kanonov, U.: Google Android: a comprehensive security assessment. IEEE Secur. Priv. **8**(2), 35–44 (2010)
19. Sherman, M., Clark, G., Yang, Y., Sugrim, S., Modig, A., Lindqvist, J., Oulasvirta, A., Roos, T.: User-generated free-form gestures for authentication: security and memorability. In: Proceedings of MobiSys, pp. 176–189. ACM (2014)
20. Sun, Z., Wang, Y., Qu, G., Zhou, Z.: A 3-D hand gesture signature based biometric authentication system for smartphones. Secur. Comm. Netw. **9**(11), 1359–1373 (2016)
21. Wayman, J., Jain, A., Maltoni, D., Maio, D.: An introduction to biometric authentication systems. In: Biometric Systems, pp. 1–20. Springer, London (2005)
22. Wayman, J.L.: Fundamentals of biometric authentication technologies. Int. J. Image Graph. **1**(1), 93–113 (2001)
23. Wong, A.M.H., Kang, D.-K.: Stationary hand gesture authentication using edit distance on finger pointing direction interval. Sci. Program. **2016**(7427980), 1–15 (2016)
24. Xiao, G., Milanova, M., Xie, M.: Secure behavioral biometric authentication with Leap Motion. In: Proceedings of ISDFS, pp. 112–118. IEEE (2016)
25. Yampolskiy, R.V., Govindaraju, V.: Taxonomy of behavioural biometrics. In: Behavioral Biometrics for Human Identification: Intelligent Applications, pp. 1–43. IGI Global (2009)
26. Ye, G., Tang, Z., Fang, D., Chen, X., Kim, K.I., Taylor, B., Wang, Z.: Cracking Android pattern lock in five attempts. In: Proceedings of NDSS Internet Society (2017)

HCI Patterns for Cryptographically Equipped Cloud Services

Thomas Länger[1(✉)], Ala Alaqra[2], Simone Fischer-Hübner[2], Erik Framner[2], John Sören Pettersson[2], and Katrin Riemer[3]

[1] Université de Lausanne, Lausanne, Switzerland
thomas.laenger@unil.ch
[2] Karlstad University, Karlstad, Sweden
[3] XiTrust Secure Technologies GmbH, Graz, Austria

Abstract. Recent cryptographic research has devised several new algorithms and protocols with a potential of mitigating several of the most ardent security and privacy threats, existing in currently available public cloud services. Nevertheless, such cryptographic schemes often exhibit counterintuitive functionality to end users, or they work differently to other already established traditional schemes with which users are already familiar. A practical solution to address these problems involves a human centered design approach, deriving Human Computer Interaction (HCI) requirements from consultations and extensive testing with experts, prospective end users, and other stakeholders. The European Horizon 2020 project PRISMACLOUD "Privacy and Security Maintaining Services for the Cloud" uses such an approach and provides HCI patterns as part of its proper cloud service development methodology CryptSDLC to communicate HCI requirements to cloud service designers and user interface implementers. In this article, we present several new cryptographic cloud services, e.g. for redacting digitally signed data, and for redundant storage and sharing of confidential data in a public cloud scenario, together with three example HCI patterns for specific interactions of end users with these services. We show how these patterns were elaborated and validated in practice to prove the suitability for their intended purpose. To summarize, we give an account on our practical experience during the actual prototype development and implementation and show how they constitute an essential element of the CryptSDLC development methodology.

Keywords: Cloud computing · Cryptography · HCI patterns · End-user security End-user privacy

1 Use of Cryptography in the Cloud

1.1 Current Security and Privacy Situation

For end users of public cloud systems, be they individuals, corporations, administrations, or other entities, a central feature is that data is given to someone else for storage and processing. This is assumed as being cost effective, enabling sharing of data and applications among devices and other cloud users, and providing protection against data loss. But

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 567–586, 2018.
https://doi.org/10.1007/978-3-319-91238-7_44

as regards protection of confidentiality, and also of integrity, and availability of the data, the cloud provider (the "controller" and "processor", in the terms of the GDPR [1]) has to be trusted to protect the data against all kinds of attacks by malicious hackers and other outsiders. In many cases, the cloud provider itself has full access to the end user data—plus the metadata arising from usage and access to the data. The cloud provider may be honestly defending end user data against outsiders, but be curious, and untrustworthy with respect to confidentiality. In several currently available public cloud offerings for individual end users, it is the very business model of the cloud provider to collect and use the user-generated content and the knowledge of user interaction for their own business.

Hence, end users have to face threats against the confidentiality of their data, and that the data remains available through the cloud service. In specific cloud applications, the integrity of private or personal data may also be under threat. However, cloud controllers, and especially now also the processors that perform the actual storage and processing in the cloud of data related to subjects in the European Union, have strong reasons to combat such threats: From May 25, 2018, when the GDPR will apply, they are facing significant fines if there are no appropriate means of security in place.

1.2 Suitable Cryptographic Primitives and Protocols

A preferred method to address several of these security and privacy concerns would be to use cryptographic protection from end to end, together with specific cryptographic functionality to reduce the amount of metadata being generated in secure and authentic transactions. However, in public cloud systems currently available on the market, cryptography is mostly only used for protecting the data between the end user and the cloud, while in the cloud the data is completely entrusted to the protection capabilities and the benevolence of the cloud provider. Most cloud services provide cryptographic protection of the data only in motion between end users and the cloud, and some provide encryption of data at rest in a simple use case, precluding further sharing and processing of the data in the cloud plus requiring from the end user a fully-fledged cryptographic key management system with all its consequences.

Nevertheless, cryptographic research carried out by researchers from universities, research centers, and corporations, is currently proposing cryptographic primitives and protocols for addressing several of these threats, and demonstrating technology readiness of its solutions with several application demonstrators in realistic cloud environments. The H2020 project PRISMACLOUD proposes the use of secret sharing for distributed storage and archiving of data among multiple cloud providers, with no single cloud provider aware of the plain data. In fact, a *secret sharing algorithm* can give more information to storage providers that are more "trustworthy". The secure data archive also enables the implementation of a secure and securely private data sharing service in a cloud of clouds, without having to rely on the need to trust one cloud provider. We demonstrate such a service in an infrastructure provider for administrations of municipalities of a region in Europe, providing secure and reliable archiving services, in a mix of private clouds and public clouds that are rented on demand.

A *selective authentic exchange service* enables end users to have verified and digitally signed information hosted in a cloud service with the intention to have precisely selected verifiable information items disclosed to a third party. This particular service wants to ensure that only an exactly specified subset of the data is revealed, the authenticity of which nevertheless can be cryptographically verified. This counters the unfavorable consequence in current certificate based authenticity systems: That always the entire certificate, the entire information need to be revealed, even when only e.g. one particular piece of authentic data (e.g. the age, the name to a key…) is required. The selective authentic exchange service is demonstrated by a prototype in the health domain. In a hospital, doctors digitally sign medical data, e.g. diagnosis data, or lab data. The patients, if they need to show some of the medical data to third persons (employer, dietician etc.), can select several data to be "blackened out", i.e. redacted from the document, without the other, remaining items losing their valid signatures.

Nevertheless, cryptographic schemes such as the ones used in PRISMACLOUD often provide "crypto-magic" and thus exhibit counterintuitive functionalities to end users, or they work differently to other already established traditional schemes with which users are already familiar [2]. Also for this reason, the HCI (Human Computer Interaction) research in PRISMACLOUD has played an important role, and the HCI patterns discussed in this paper are reflecting the results of this HCI work.

1.3 HCI Patterns as Promoter of Cryptography Diffusion

An analysis of promoters and inhibitors of cryptography diffusion in the (public) cloud context [3] yielded several results, among these that it is favorable to have during the cloud service development process the proper instruments and procedures in place to communicate requirements and capabilities across the domains of experts, involved in the development process on the different layers. Design patterns in their different expressions as *security and privacy patterns*, as well as *HCI patterns* are such communication instruments—of which we will show three instantiations below. The HCI patterns are being developed with feedback from end users, and codify requirements and design decisions for several HCI aspects of the application of cryptographic cloud services. As such, the HCI patterns support the creation of usable and accepted end user applications.

Other Promoters Identified. Another promoter for the diffusion of cryptography in cloud applications, as identified in the context of the PRISMACLOUD project, is the *usage of a service and tool based approach,* where the cryptography, with its complexity, is hidden inside a tools layer, and thus can more safely and securely be used by cloud service developers. *Standardization of cryptographic primitives and protocols* (algorithms and parametrization) *and compliance to regulation* are also beneficial to increase trust in cryptographic cloud services and make them more widely used. Compliance, in the European context compliance with the GDPR, will require from controllers and processors (aka providers) that they use appropriate technical means of protection, which could include for privacy-sensitive applications, such as eHealth applications, strong cryptography to (provably) protect entrusted personal data—otherwise, the GDPR foresees severe financial fines for cases where a breach happens with no adequate data protection in place.

Other Inhibitors Identified. The *complex and frequently rudimentary specification* (at least for a practical use case) of a cryptographic algorithm or protocol in the scientific literature makes transformation and secure implementation in a real service difficult and expensive. Notations are often very formal and the security assumptions and the correct parametrization not derived easily. Other identified inhibitors include the existence of *artificial requirements*, where cryptographic primitives are being brought forward, satisfying more requirements of cryptographic research, and of academic beauty—than being practically applicable in a real world application.

2 HCI Patterns as Integral Part of a Cloud Service Development Methodology

2.1 The PRISMACLOUD CryptSDLC Method

The construction of cryptographically equipped cloud services is a huge undertaking and requires contributions by and collaborations among many involved disciplines on different layers. To structure that process, and to enable a secure development process, the project PRISMACLOUD proposes a fourfold architecture [4], plus a proper methodology for the research and development activities required during cloud application development an all of these layers.

The architecture layers are:

- The **applications layer**—of the applications using (public) cloud services.
- The **services layer**, providing the cloud services to the end user applications. The services use cryptographic tools of the tools layer to implement security and privacy functions.
- The tools of the **tools layer** completely encapsulate the cryptographic primitives and protocols, including the correct parametrization, thus supporting a secure and effective use.
- The **cryptographic primitives' layer** containing the cryptographic primitives and protocols. Here the cryptographic research is being carried out.

The PRISMACLOUD CryptSDLC ("Cryptographic Software Development Lifecycle") defines the activities connected with traversing the architecture during cloud service development [5]. From the applications layer down, high-level requirements are derived and translated to more formalized language; such requirements are mapped to cryptographic models on the layer of cryptographic primitives and protocols. Cryptographic research is being carried out to fill gaps and provide the required functionalities. The algorithms and protocols are built into software, which is being structured as a tool to be used by cloud services on an upper layer. Security and privacy is built into the tool *by design and by default* and in an optimal world, the practical security can also be quantified and (formally) proven. The tool is deployed, and the cryptographic capability provided to high-level applications through a cloud service. The CryptSDLC method is based on conventional software development lifecycles, like Microsoft SDL, but augmented with steps specifically dealing with designing cryptographic systems [5].

2.2 Experts Involved in the CryptSDLC

Table 1 lists for each architecture layer, which group of individuals needs expertise on that particular layer–plus at least in the adjoining layer above and below (if there is a layer above or below) during the development process [5]. For example, a tool designer with main expertise in the tools layer needs knowledge of the capabilities of the cryptographic primitives developed and configured in the primitives layer, as well as of the requirements postulated by the experts of the Services Layer. A service designer, on the other hand, only needs knowledge of the tools but no longer the detailed cryptographic knowledge of the cryptographic primitives and protocols layer.

Table 1. Experts, engaged on architecture layers

Primitives layer	Cryptographers
Tools layer	Tool designers, specialized software engineers, HCI experts
Services layer	Service designers, usability and HCI experts, cloud service providers and sub-providers (GDPR: “controllers” and “processors”)
Applications layer	Business model developers, general domain experts
On several or all layers	Project communicators, IT security specialists

PRISMACLOUD maintains specific communication tools and mechanisms to support the layered development process governed by the CryptSDLC, as well as to support the diffusion of new paradigms and capabilities among prospective providers and end users of the proposed tools and services. These communication tools and mechanisms are applications of design patterns: *Cloud security and privacy patterns*, and *HCI–human computer interaction–patterns*.

2.3 Role of HCI Patterns

Design patterns on several levels provide communication functions during the development process. The patterns support the layered development process governed by the CryptSDLC. Cloud security and privacy patterns codify and explain the new paradigms and capabilities from cryptographic researchers to prospective service providers and end users. Specific HCI patterns guide the implementation of interfaces guiding the human computer interaction.

Cloud security and privacy patterns and HCI patterns are applications of design patterns. In similar structures, they are used to codify expert knowledge and requirements within a specific scope in a way that the information remains accessible across domains of involved actors. The main idea is that a design pattern shall “describe(s) a problem which occurs over and over again (…) and then describe(s) the core of the solution to that problem, in such a way that you can use this solution a million times over (…)” [6]. This is done by describing the (empirical) background of the pattern, i.e. the “problem”, and giving instructions for the “solution” in natural language in a framework of categories.

The concept was invented in Berkeley, CA, in the 1970s for application in architectural design [6] and has later on been modified for application in several information technology subdomains. The first application of design patterns in information technologies was in software architecture in the 1990s when object oriented design and reusability required efficient communication of complex issues across different domains of involved people [7]. Later on, the concept was used for the specification of security and privacy concerns in security and privacy patterns [8, 9], as well as for human computer interaction aspects in HCI patterns [10]. Since several years, there exist collections and catalogues of cloud security and privacy patterns specifically for modelling threats and solutions in the cloud context.

2.4 HCI Patterns Methodology

We will follow the structure of the HCI patterns as presented in [10], and additionally embrace the pattern with an overview section and a motivation section at the beginning, and a section testing and validation, showing how the pattern was elaborated and validated. In Table 2 we give a short definition of each of the categories used to describe the HCI patterns.

Table 2. HCI pattern categories

Overview	Title of the pattern, including information on its maturity (i.e. how intensely it has been tested so far). For pattern HCI.P2 the overview contains a description of the underlying cryptographic primitive of redactable signatures to an extent required for the comprehension of the following HCI pattern
Motivation	Description, why the necessity arose for the particular pattern
Problem	Outline of existing problem; description of the context and applicability of the pattern
Solution	Describes the elements necessary to solve the given problem. Describes how the elements need to be arranged to achieve this goal. “The solution describes the elements that make up the design, their relationships, responsibilities, and collaborations” [7]
Use when	Outline of the situation and context when the pattern is best applied in
Use how	Provide detailed insight into the way the solution is being achieved; provide detailed information for the developer and implementer (steps needed to achieve the solution)
Use Why	Rationale as why the pattern is needed and where the benefit for the end user lies
Related patterns	Related patterns (in this collection, in other catalogues)
Testing and validation	Description on how the testing was carried out and how the pattern was validated

The HCI pattern categories grid is not to be seen as orthodox and fixed—for specific presentations and communication need, categories may be omitted, or other categories (as e.g. “GDPR context”, or “Standardization status”) could be added.

3 Example HCI Patterns for Cryptographic Applications in the Cloud

3.1 PRISMACLOUD Use Cases

In the PRISMACLOUD project, several end user applications were developed to demonstrate the capabilities and security benefits of the proposed cryptographic tools and services. A *Health Care Data Sharing Platform* uses a selective authentic exchange service to enable the minimization of data to be shared with third parties to the items actually required. The *e-Government IaaS (Infrastructure as a Service) Cloud* provides a redundant and highly secure backup solution in a hybrid cloud scenario. The three presented patterns were tested and are being applied in the *Health Care Data Sharing Platform* (HCI.P1, HCI.P2) and the *e-Government IaaS Cloud* (HCI.P3). The *Privacy Enhanced Simon* equips an existing application of the FP7 SIMON project (an implementation of a mobile application for prioritized parking for people with disabilities) with capabilities for effective privacy protection and data minimization of the involved end users. An *Evidence Sharing Platform* capable of deployment to a public cloud protects its information against curious cloud providers. The three HCI patterns are being applied in the Health Care Data Sharing Platform, as well as the e-Government application.

3.2 HCI.P1 Digital Signature Visualization

Overview. Digital Signature Visualization is a relatively mature pattern and is already practically applied in XiTrust's commercial MOXIS solution[1]. It was user-tested in the PRISMACLOUD eHealth use case "Healthcare Data Sharing Platform". The tests included two sets of users, the signers of the document (medical staff) and the redactors (users of the medical document: patients). Several results of the user test were incorporated into that pattern.

Motivation. XiTrust has been active in the field of digital signatures for more than 15 years and experience shows that users need a graphical visualization for digital signatures. Consequently, the graphical representations have been revised constantly over the last years. In principle, a digital signature does not need to be visualized, but then it would not be visible to the human eye. One option would be to simply apply the hash value of the signature to the document. However, this option has found only little acceptance. Thus, we implemented the possibility to upload an image into the XiTrust MOXIS digital signature solution, which shows a scan of the handwritten signature. In times when more and more people are doing their business on tablets and smartphones, a more flexible solution was requested. Therefore, we provide now also the possibility to create the signature visualization directly with a stylus or finger. Furthermore, it is

1 Online (8.2.2018): https://www.xitrust.com/en/products/xitrust-moxis/. The MOXIS solution is currently available for qualified digital signatures, but not for redactable signatures. The identities are provided with trust service provider A-Trust, online (8.2.2018): https://www.a-trust.at/%C3%BCber-uns/en/.

possible to display the name of the signer, the date, the time and a short user-defined message under the visualization.

Problem. Prospective digital signers of documents need to see which document they are just going to digitally sign in an electronic document flow. Then signers need to be sure whether a digital document they are seeing on a computer display has already a valid digital signature attached to it. This is particularly of importance right after the process of signing a document (to confirm that the operation has been successfully carried out), but also in review of a past action (e.g. that the document has already been signed). Signature verifiers, i.e. the people receiving the signed document, need a straightforward and intuitive way to check the signature's validity (invalid signatures need visual representation). Additionally in the eHealth use case, when viewing the final document there are two types of signatures: signing the document with redactable signatures, and signing the redaction for accountability with a digital signature. Although the second (the digital signature of the redactor) could be valid, it does not necessary mean that the redactable signature (the first) is still valid; that is due to the dependency of redactable signatures on redaction rules (validity of the signature remains when redacting only allowed fields from the document).

Solution. Provide the user the required information in a way that resembles the common process of handwritten signatures on paper documents. Digital documents are rendered on the screen black on white, resembling printed documents. Digital signatures are represented by images of signers' handwritten signatures, and on the location where signatures would be expected on a paper document, i.e. at the bottom of the last page.

Use when. Use it in the generation and verification process of digital signatures in a digital documents flow.

Use how. Prospective signers shall be presented with a placeholder, indicating that a signature may be placed on a certain document. Such a placeholder has the form of a frame or box, which is empty (i.e. contains no image of a written signature) but contains information on who is entitled to sign the document (see Fig. 1). This can be a list of natural persons' names, or the name of a group, of which any individual member may sign (e.g. "medical doctor"). The placeholder shall be positioned at the bottom of a document, as this is mostly the case for handwritten signatures on paper. If there are

Fig. 1. Signature placeholder mock-up. The QR code encodes metadata associated with the signer (e.g. identity) and the application (e.g. position of visualization in document, descriptor of business process related to document) in mixed electronic/paper document environments.

more than one signatures required on a particular digital document, several such placeholder boxes should be placed next to each other. Enabling the placing of the placeholder by the signer is not advised, as this tends to confuse the signer.

After the signature process itself has been triggered (and the proper authentication of the signer has been established) an image of the signer's manual signature is displayed in the frame (see preview in Fig. 2. Handwritten signature visualization). Therefore, the signer has immediate feedback that the signature process was successfully carried out. Likewise, a verifier can have an efficient signalization that a document is signed, and by whom.

Fig. 2. Handwritten signature visualization

The signer shall be given the opportunity to revoke a signature on a document that was erroneously signed, e.g. when digital signatures have to be created in a stressful working environment, e.g. in a hospital. The revocation should be easily achievable by the signer immediately after the signing process (probably within a defined time limit, or while one particular session is ongoing) and should not involve a lengthy and difficult procedure, as this is common in current digital signature applications.

Use Why. Although digital signatures are intended as equivalent to handwritten signatures, they are merely bits of digital data that are not easily identifiable as a signature. A similar discrepancy exists between a document on paper to be signed, and electronic data, representing some document. Consequently, there is the problem of how both the digital document to be signed and the digital signature itself are visualized. We are here concerned with the latter problem, the visualization of the digital signature. The visualization of the space where a signer may 'place a digital signature (i.e. at the bottom of a 'virtual document'), provides for the signer an intuitive way to assess the quantity of data that the signer is about to digitally sign. The visualization of a valid digital signature by a box containing an image of the signer's handwritten signature provides, for both the signer and the verifier, an intuitive way to understand that a document is validly signed.

Related patterns. This pattern is used in several other HCI patterns having to do with digital signature applications, e.g. HCI.P2 Stencil for Digital Redaction (see below).

Testing and Validation. The visualization of signatures was tested as a part of the eHealth use case scenario walkthroughs on two sets of users: the medical staff who are the signers of the medical documents with redactable signatures, and the prospective patients who will be redacting the document, signing the redaction with digital signatures, and using their medical document for further purposes. We used low fidelity mockups[2] for the interface testing. In total there were 13 medical staff interviewed individually for testing the signing of redactable medical documents, and 5 focus groups for testing the redaction of signed documents and signing the redaction. The groups consisted of 32 participants of patient-users: 2 experts groups and 3 lay user groups.

Medical staff appreciated the visualization of the digital signatures, however raised their concerns regarding multiple signatures process across different departments in the medical facility. The signature placeholder was confused for the actual signature by some lay users, however it was noted by the participants that it is due to first time learnability encounter. Among the focus groups participants, views varied from appreciating the visual representation (non-expert users).

3.3 HCI.P2 Stencil for Digital Document Redaction

Overview. Redactable signatures, also known as malleable signatures, provide a means for, within given boundaries, redacting parts or field blocks from digitally signed digital documents, without the signature losing its validity [11]. In applications employing redactable signatures, one signer signs a digital document, from which a second user is able to redact some information (redacting in the sense of suppressing or "blacking out"), keeping in mind that redaction rules apply, i.e., not all fields are redactable. Meanwhile a verifier still can check the authenticity (versus the first signer) of the remaining information, as well as the authenticity of the redaction made by the second signer. HCI.P2 is also being applied in PRISMACLOUD in the eHealth use case "Healthcare Data Sharing Platform".

Redactable signatures are based on relatively recent cryptographic primitives, and there is (by the time of this document) not much end user experience available with the use of redactable signatures in digital signature applications. HCI tests have revealed that people tend to have problems to grasp the correct functionality and implications of a redactable signatures application and often preconceptions are uttered. A potential explanation of this phenomenon may be twofold: First, digital signatures in current digital signature applications are conventionally connected to the property that any modification destroys the validity of a connected digital signature. That situation is acerbated by the very name of "redactable signatures", which, being a technical term in the field of cryptography research, is misleading from an end user application point of view. This is because from an end-user point of view, it is not the signature that is being redacted, but rather the signed electronic document. From an end-user perspective, e.g. "redactable authentic documents" would be a more suitable name.

Motivation. In our previous user studies, we have elicited and evaluated HCI and user requirements for redactable (malleable) signatures [12, 13]. It was clear that there is a need to communicate and facilitate the functionalities (redacting of documents) to the user using suitable and understandable user interfaces and metaphors. Furthermore requirements for redaction called for suitable metaphors and support for the user. Therefore, we have chosen the stencil metaphor for the process of redaction and developed

[2] Balsamiq Mockups 3 by Balsamiq Studios LLC. online (8.2.2018): https://balsamiq.com/.

mock-ups user interfaces for visualizing redaction of signed documents. To improve usability and ease human computer interaction, a practical application for the redaction of digital documents shall at any time provide an immediate feedback on which of the visible elements on the screen will be visible, or redacted from the final document.

Default settings for redactions complying with the privacy principle of data minimization and data protection by default (Art. 25 GDPR) are needed.

Problem. During redaction of digital documents, the redacting end user may lose control of which parts of an electronic document are redacted (and which parts will remain visible for a potential verifier). Some parts of the document may not be redacted without the digital signature losing its validity. An end user may redact too little or too much information, so that the remaining document either does not disclose minimal data or may no longer fit for its intended use.

Solution. The elements that are redactable within the predefined framework, i.e. the elements that can be redacted by a user without the initial signature losing its validity, shall be clearly indicated; users shall be given templates that propose redactions for a specific purpose, and indicate which fields may be redacted, without the document losing its suitability for the intended purpose in situations where it may be applicable.

In Fig. 3, the icon depicts the "blacking out" based on the stencil metaphor, which can be used in the user interface for choosing the functions to redact documents. However, for the process of redaction "greying-out" of fields should be used instead of "blacking out" for leaving the text to be redacted visible and thus helping the user to verify which parts of the text will be redacted and which will remain. Hence, greyed-out fields highlight the parts of the text that the user chooses to redact and thereby limit the information they would like to share.

Fig. 3. Icon for redacting fields in the document

In Fig. 4, mock-ups (for a redaction template that the users first have to choose) depict an overview of how a document would look like after redaction, where users get to see the greyed-out blocks of text to be redacted. Two different views marking either the text to be redacted or the text to be kept are offered. The actual final document without the markings is shown to the user ('Document After' view).

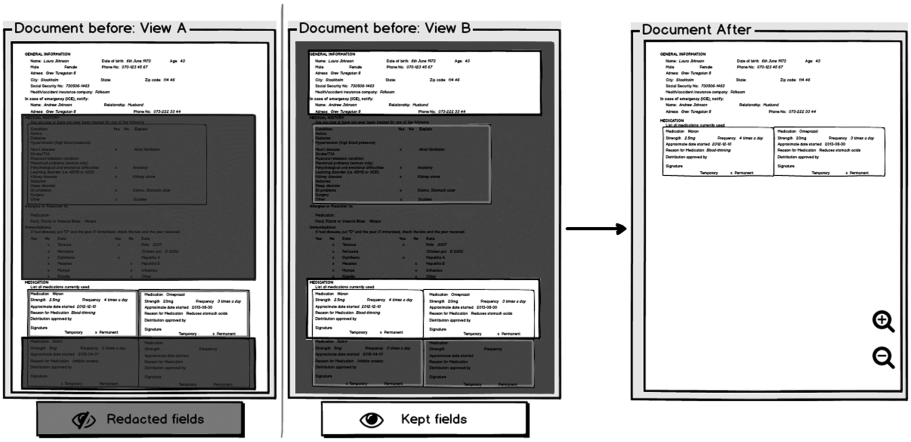


Fig. 4. Selecting redaction views

Use when. The Stencil for Digital Redaction should be used for the user interface during the redaction process of digital documents in redactable signatures applications, i.e. the Stencil for Digital Redaction shall support the redactor during the redaction process. Moreover, icons based on the stencil metaphor should be used for allowing the users to easily choose the redaction functions.

Use how. The digital document shall be presented on the screen resembling the printed document. At any time, and for all parts of the document visible on the screen, all potentially redactable elements shall be clearly indicated by displaying them inside a frame or box. Of these elements, not redacted elements, i.e. visible for a verifier, shall be displayed inside boxes with transparent background, while redacted elements, i.e. not visible for a verifier, shall be displayed in a box with grey background ('greyed-out'). Non-redactable text shall be displayed without a frame or box around it. At the beginning of redaction, users are given two options: one that users are clicking on the fields to be redacted and greyed out (View A in Fig. 4), the second is clicking on fields to be highlights and kept (View B). Both views will have the same end document result; it is mainly the mental model and preference of people selecting either hiding or showing information (Fig. 4). Each redactable section (i.e. each frame) shall have an adjoining button for toggling the redaction status (visible/not visible for a verifier) of that specific section (Fig. 5) User tests have proven an eye symbol as button as being convenient and effective for conveying the meaning of 'hiding' or 'making visible again' to the redactor.

The redactor shall be presented a choice of templates for the redaction of documents for specific purposes. E.g. if the original document would be results of a lab test in an eHealth portal (signed by the lab or responsible medical doctor), a template could be provided for the end user passing on information to his or her dietician. The templates shall be designed to enforce Privacy by Default, i.e. in the lab use case the template should propose to the redactor to redact all test results not primarily needed for dietary counselling, i.e. the data proposed by the template to be redacted shall be put in already

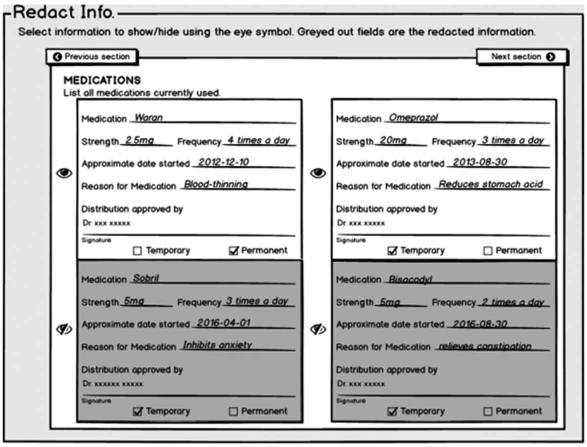


Fig. 5. Hiding/showing more information than the template.

greyed-out boxes. The information on the template currently in use shall be indicated on a separate portion of the screen at any time when the end user does the redaction.

Use why. HCI.P3 Stencil for Digital Document Redaction in combination with redaction templates guide a redacting user through the difficult process of digital redaction of signed documents. It helps the end user to grasp the consequences and implications of his or her actions during redaction and to enforce data minimization.

Related patterns. The accountability of a redactor for the effected redactions is usually implemented by a digital signature of the redactor. Use HCI.P1 Digital Signature Visualization for the digital signature process of the redactor.

Testing and Validation. The focus group study mentioned in HCI.P1 (Sect. 3.2) included the testing of redaction process by patient-users. Participants of all 5 focus groups understood the greying-out of fields in redaction process. In the overall view of document redaction, almost all participants chose View A (Fig. 4) as the suitable view for redaction, where they prefer to select further information to be greyed out rather than selecting information to be shown from the greyed-out view. Templates were perceived to be important part of redaction, since they act as a guide of redacting sensitive information by default; many participants indicated that they would rely on the templates for redaction. Therefore, it was concluded that there is a need for data protection certification (as promoted by the GDPR) for the template redaction specification for data minimization rights as well as setting the boundaries for data requirements of recipients, e.g., employers who would request more information than legally needed from their employees' documents.

3.4 HCI.P3 Secret Sharing Configuration Preferences

Overview. ARCHISTAR is a framework for secure distributed data storage and sharing in the cloud, it constitutes a system that applies Secret Sharing to a multi-cloud setting, meaning that the user's data is divided into "chunks" which are distributed to separate clouds/storage nodes [14]. This implies that an incident at a single cloud/node will not cause the data to be lost, stolen or tampered with. Hence, it intrinsically protects data privacy and availability. When adopting a Secret Sharing (or Secret Splitting) scheme, there are two fundamental parameters that need to be considered: (1) $\boldsymbol{n}$, representing the number of "chunks" that data should be divided into. (2) $\boldsymbol{k}$, which constitutes the threshold of chunks required to reconstruct the data into its original state. A higher threshold makes data more protected against data privacy breaches because a higher number of cloud servers (least $\boldsymbol{k}$) would have to collude against the user (i.e., data owner) to restore the data without permission. However, a higher threshold also makes the data less accessible for not only unauthorized individuals, but also for the user themselves, as a data recovery would require the availability of a higher number of servers. As illustrated in [15], different configurations of $\boldsymbol{n}$ and $\boldsymbol{k}$ affect the availability of data chunks in a multi-cloud setting. The availability of cloud services is commonly expressed in percentage of uptime – or number of "leading nines" [15]. A typical cloud service availability rate is 99.9%, or three nines [14], which constitutes a downtime of 8.76 hours per year. From that point on, three main categories for setting up the configuration preferences are set and are focusing on: "Cost Minimization", "Data Confidentiality Maximization" and "Data Availability Maximization", which constitute the 3 packages/categories of this pattern. The use of the word "Maximization" should clearly indicate to the user that even though all kinds of configurations will, through the use of secret sharing, already protect the confidentiality and availability of the data, there are options to even strengthen or "maximize" these protections.

Motivation. 16 structured interviews were conducted to derive suitable ARCHISTAR configurations and guidelines for organizational/private use, and to identify trust factors, unique advantages and risks of ARCHISTAR that should be communicated to different user groups. The respondents were IT experts who are familiar with the notion of cloud storage and had previous experience of organizational and/or private cloud storage use. Based on the interview results, it was noted that setting up configurations of data chunks and locations required guidance and support. Compared to encryption, secret sharing was generally perceived as less secure against breaches, while some reasoned that it would constitute a greater protection against data loss. The majority of respondents argued that secret sharing would not single-handedly be a sufficient security measure for sensitive data in the cloud and a layer of encryption would be valued or required in addition, as they would not trust or would not want to rely on the non-collusion assumption of secret sharing and/or as they anyhow would internally use encryption for protecting sensitive data.

When asked how many chunks (n) their data should be divided into and what the threshold for reconstruction (k) should be, most respondents did not appear to understand (or put much thought into) how different values on k would influence the level of security and availability of data. In other words, the implications of different combinations of n and k might not be clear to the user and arbitrary values may therefore be selected, resulting in a less suitable configuration.

The perceived importance of cost was mixed among the respondents. Some argued that they would not use the solution if the expenditure would be too high, while others though cost was a less crucial factor. However, regardless of the perceived significance, it was acknowledged that cost could have an impact on other key factors (such as security and reliability). This indicates that cost still may have be taken into consideration throughout the configuration process. However, other interviewees rather opted for data protection or for data loss preventions as their highest preferences.

Problem. Secret Sharing is a security measure for protecting both the availability and privacy/confidentiality of data. Various types of data may be stored/backed up in the cloud, all of which may involve different requirements in terms of the degrees to which Confidentiality, Integrity and Availability (CIA) are protected, which in turn can be influenced by the secret sharing configurations. Particularly the sensitivity of data and the frequency in which it needs to be accessed by the user may determine which Secret Sharing configuration is appropriate. However, during the interviews some respondents found it difficult to give their personal/organizational data a particular classification in regards to the CIA triad. This suggests that the requirements and priorities should be specified in a different manner.

Solution. The user should be presented with three configuration preferences –"Cost Minimization", "Data Confidentiality Maximization – High Data Protection", and "Data Availability Maximization – High Data Loss Prevention"– which should be prioritized from most important to least important. Based on the priority, the user will be provided with recommended default settings and configuration options. The aforementioned categories have the following implications and trade-offs:

Cost Minimization: The expenses should be kept small by selecting cheaper cloud storage options. The pre-selection of providers should be dictated by the price of the cloud storage offering, rather than locations in different regions or jurisdictions. However, if "Data Protection" is the second-highest priority, storage servers should be located in the European Union (EU) or within the organization's private cloud for guaranteeing data protection in compliance with the GDPR. If "Data Loss Prevention" is the second-highest priority, some pre-selected providers might reside in areas with a low risk of natural disasters to ensure that more than ($\boldsymbol{n}$-$\boldsymbol{k}$) chunks cannot be hit by the same natural disaster at the same time. Even if data protection is not the first or second preference, the data may still include personal information, and therefore out of the chunks needed to restore the data, at least one should per default be located in the EU.

Data Confidentiality Maximization – High Data Protection: The data is sensitive and requires high confidentiality. Accordingly, encryption in addition to secret sharing should be a mandatory feature. Compared to the total number of chunks ($\boldsymbol{n}$), a relatively high threshold for reconstruction ($\boldsymbol{k}$) should be recommended to the user in order to minimize the risk of collusion attacks. The (pre-selected) cloud storage providers should be geographically located in EU and follow EU privacy legislation (and particularly the GDPR), or should even if possible be located in the organization's private cloud. If the second-highest priority is "Data Loss Prevention", the configuration of $\boldsymbol{n}$ and $\boldsymbol{k}$ should be adjusted so that a high availability rate still will be achievable with a high threshold (i.e. increase the total number of chunks). Also, the choice of providers will additionally be determined by the geographical distance between them to minimize the risk that more than (***n***-***k***) chunks can be simultaneously destroyed or be inaccessible due to the same natural disaster. If "Cost Minimization" is the second-highest priority, the choice of providers will rather be influenced by the charged costs for the cloud storage offering.

Data Availability Maximization-High Data Loss Prevention: The data has high availability requirements. The option to add encryption should not be provided by default since it increases the risk of data being lost or inaccessible due to key loss issues. The recommended number of chunks (***n***) should be significantly bigger than the threshold for reconstruction (***k***) to ensure that the user will be able to restore the data if incidents occur at several storage nodes. That is, the user interface should suggest a configuration for a high availability rate (i.e. > 99.9%). The pre-selected locations should have a sufficient distance between them to ensure that a single disaster will not cause multiple chunks (i.e. > (***n***-***k***)) to become inaccessible. Storage nodes in high risk areas for natural disasters should not be available options in the interface. If "Data Protection" is the second-highest priority, the threshold should be slightly increased to ensure that the data will be protected against a higher number of breaches. Locations that are compliant with EU privacy laws should be suggested, which might limit the distance between storage nodes. If "Cost Minimization" is the second-highest priority, the choice of providers will again rather be influenced by the costs charged for the offered storage. Again, even if data protection is not the preferred second choice, the data may still include personal information, and therefore out of the chunks needed to restore the data, at least one should per default be located in the EU.

Use when. Throughout the ARCHISTAR configuration process of data backups that the user intends to protect in the cloud with Secret Sharing.

Use how. The first step in the configuration process should involve data classification to indicate the user's/organization's needs. The user should be presented with three main categories "Cost Minimization", "Data Confidentiality Maximization" and "Data Availability Maximization" (see Fig. 6), which should be prioritized from most important to least important.

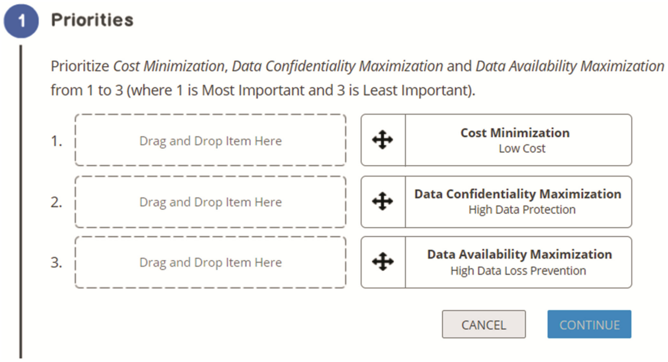


Fig. 6. Selection of the three categories according to priority.

Default settings should be suggested by an interface which can be manually adjusted by the user if desired. In particular, the user could change values for n and k, as well as the selection of storage nodes (to which data chunks should be geographically distributed) on a map. The configuration process would subsequently be completed by proceeding to an "Overview" and a "Confirmation" page.

Use Why. To avoid any ambiguity regarding the Secret Sharing mechanism and to assist the user in creating the most suitable configuration for their intended data backup.

Testing and Validation. A first iteration of the interface was evaluated during 5 preliminary walkthroughs/interviews. The respondents constituted 1 Administrative Director at a municipality's IT department, 1 IT Security Coordinator at a university, and 3 IT experts of which one had several years of experience in an IT security consultant company.

While some respondents appeared to perceive the categories as sufficient for describing the user needs/requirements of the intended data backup, some questioned the category names. The distinction between "Data Protection" and "Data Loss Prevention" was not totally clear for all respondents. Moreover, one respondent desired more information about what "Cost minimization" implied (i.e. to what extent are the expenses reduced?).

Providing recommended default settings based on the user's priorities appeared to be seen as an appropriate solution. While some respondents still would like the *option* to change the total number of chunks ($\boldsymbol{n}$) and the threshold for reconstruction ($\boldsymbol{k}$), most of them seemed to prefer using default values provided by the system. Some respondents even argued that parameter $\boldsymbol{n}$ and $\boldsymbol{k}$ should not be presented by the user interface at all, since their implication would not be clear to the user. The notion of selecting cloud storage providers based on location was also received with mixed views. Some respondents thought that the user should be able to select specific data centers, others thought locations should be selected on a higher level of abstraction (e.g. country or continent) – or even be pre-selected by the system based on already signed contracts with providers.

In correspondence to these evaluation results, we have, as described above, introduced default settings (in particular for the values $\boldsymbol{n}$ and $\boldsymbol{k}$) for the next UI iteration that the user can adapt if they would like to. Besides, the names of the three categories were slightly changed or amended (see Fig. 6).

4 Assessment and Lessons-Learned of Practical Application

In a research project of 3.5 years duration, it was only after 3 years, that our first HCI pattern came into practical use during the project internal service and application development process, governed by the CryptSDLC methodology. The service architecture was roughly available one year into the project, while the CryptSDLC method was available as first draft after 1.5 years, and fully specified after 2.5 years. It was also then, that the first three HCI patterns were published as part of the HCI Guidelines[3]. Currently, the patterns are used in feedback cycles to application developers to adjust the user interfaces with results from the tested HCI patterns. To our experience, it would have been better to have the HCI patterns available at an earlier stage of the application development process (so that they might have led to more *initial design* and less retroactive *adjustment*) but the obviously sequential processes of service definition/user interface development and testing/presentation of results as HCI patterns explains to some extent why the patterns came into play so late. So one result for similar projects would be to look into HCI pattern in an early as possible phase of a development process, and thus probably also rely on existing catalogues of HCI patterns, as e.g. given in [10], or in the patterns resulting from the PRISMACLOUD project[4].

In the PRISMACLOUD project, we are using design patterns also as cloud security and privacy patterns for the communication of requirements and capabilities during cloud cryptographic tools and cloud services development. Already in this area of application (earlier in the project), design patterns helped to communicate across domains of experts and stakeholders, as they later supported communication for improved HCI. The actual development of patterns requires a detailed study of the proposed cryptographic tools and services from several perspectives (from implementers, from cryptographers, from different end-user views, like e.g. these of doctors and patients, or organizations and customers, etc.). The pattern development helped to draw the focus from a technical approach to a user centered approach, which supports more the production of usable and accepted cryptographic applications.

As regards research on HCI concepts supporting usability and trust of cryptographically secured cloud services, the process of developing the HCI patterns supported empirical work on user experiences and perceptions of new paradigms, like redactable

[3] The respective PRISMACLOUD deliverable D3.2 "HCI Guidelines" is unfortunately marked *confidential* and thus not publicly available. An iteration D3.3 "HCI Research Report" with classification *public*, containing all the HCI patterns developed in the project, will be available by project end 31 July 2018 on the project homepage https://prismacloud.eu.

[4] Ibid. PRISMACLOUD D3.3.

signed documents, social secret sharing, and privacy preserving authentication. The tested and evaluated HCI patterns provide the requirements for metaphors influencing the mental models, suitable to support end user acceptance and ease in the difficult field of cryptographically secured cloud services.

Acknowledgments. The authors' work is supported by the European Union Horizon 2020 research project No. 644962 PRISMACLOUD "Privacy and security maintaining services in the cloud"; (2/2015-7/2018); online (8.2.2018): https://prismacloud.eu.

References

1. European Commission: Regulation (EU) 2016/679 of The European Parliament and of the Council, of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016). Accessed 20 July 2017
2. Wästlund, E., Angulo, J., Fischer-Hübner, S.: Evoking comprehensive mental models of anonymous credentials. In: Camenisch, J., Kesdogan, D. (eds.) iNetSec 2011. LNCS, vol. 7039, pp. 1–14. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27585-2_1
3. Lorünser, T., Krenn, S., Striecks, C., Länger, T.: Agile cryptographic solutions for the cloud. e & i Elektrotechnik und Informationstechnik, September 2017. ISSN: 0932-383X (printed version) ISSN: 1613-7620 (electronic version)
4. Lorünser, T., Slamanig, D., Länger, T., Pöhls, H.C.: PRISMACLOUD tools: a cryptographic toolbox for increasing security in cloud services. In: Proceedings of the International Conference on Availability, Reliability and Security (ARES 2016). IEEE (2016)
5. PRISMACLOUD: Improved Guidelines and architectures for Secure Service Composition. Public deliverable D7.6 of the PRISMACLOUD H2020 project (2017). https://prismacloud.eu
6. Alexander, C., Ishikawa, S., Silverstein, M.: A Pattern Language: Towns, Buildings, Construction. Oxford University Press, New York (1977)
7. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns Elements of Reusable Object-Oriented Software. Addison-Wesley, New York (1994). ISBN 0-201-63361-2
8. Schumacher, M., Fernandez-Buglioni, E., Hybertson, D., Buschmann, F., Sommerlad, P.: Security Patterns - Integrating Security and Systems Engineering. Wiley, West Sussex (2006)
9. Länger, T., Pöhls, H.C., Ghernaouti, S.: Selected cloud security patterns to improve end user security and privacy in public clouds. In: Schiffner, S., Serna, J., Ikonomou, D., Rannenberg, K. (eds.) APF 2016. LNCS, vol. 9857, pp. 115–132. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44760-5_8
10. Fischer-Hübner, S., et al. (ed.): HCI Pattern Collection–Version 2. PrimeLife Deliverable D4.1.3 (2010). http://primelife.ercim.eu/images/stories/deliverables/d4.1.3-hci_pattern_collection_v2-public.pdf
11. Derler, D., Pöhls, H.C., Samelin, K., Slamanig, D.: A general framework for redactable signatures and new constructions. In: Kwon, S., Yun, A. (eds.) ICISC 2015. LNCS, vol. 9558, pp. 3–19. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30840-1_1
12. Alaqra, A., Fischer-Hübner, S., Pettersson, J.S., Wästlund, E.: Stakeholders' perspectives on malleable signatures in a cloud-based eHealth scenario. In: HAISA, pp. 220–230 (2016)
13. PRISMACLOUD public deliverable D3.3 HCI Research Report. The report, containing the HCI patterns developed in the PRISMACLOUD project will be available by project end, 31 July 2018. https://prismacloud.eu

14. Loruenser, T., Happe, A., Slamanig, D.: ARCHISTAR: towards secure and robust cloud based data sharing. In: 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), pp. 371–378. IEEE, November 2015
15. Happe, A., Wohner, F., Lorünser, T.: The archistar secret-sharing backup proxy. In: Proceedings of the 12th International Conference on Availability, Reliability and Security, p. 88. ACM, August 2017

Why Users Ignore Privacy Policies – A Survey and Intention Model for Explaining User Privacy Behavior

Manuel Rudolph(✉), Denis Feth(✉), and Svenja Polst(✉)

Fraunhofer IESE, Kaiserslautern, Germany
{manuel.rudolph,denis.feth,svenja.polst}@iese.fraunhofer.de
http://www.iese.fraunhofer.de

Abstract. Privacy is a vital aspect of IT systems and services, and it is demanded from users and by law. Thus, most data-processing services provide interfaces for users to support transparency (e.g., privacy notices) and self-determination (e.g., privacy settings). In this paper, we present evidence that users do not make use of these privacy interfaces—although they generally would like to. Based on our findings, we present an intention model in order to explain this behavior. The model combines aspects such as privacy demands, motivation and barriers in order to argue about the resulting intention of the user regarding the application of privacy interfaces. We show the applicability of our model by instantiating it to a concrete use case.

Keywords: Human centered design and user centered design
Psychological application for user interface · Adaptive and personalized interfaces
Privacy · Motivation · Intention

1 Introduction

Every day, users share information while using digital services. As this data is typically person related, it is of high value for users and service providers. Users benefit from data sharing by highly customized and easy-to-use services. On the other hand, providers use collected data for user profiling, personalized advertisements and other lucrative analyses. Thus, many users have a variety of privacy concerns regarding the use of these data-centric services. In order to protect users, authorities passed legal regulations to empower the users to take protective measures for personal data according to their privacy needs. For instance, the European legislature passed the General Data Protection Regulation (GDPR) [6], which imposes (among others) these two requirements:

- *Transparency:* Users must be able to understand how companies collect, use and share data in order to have a basis for decision-making.
- *Self-determination:* Users must be able to configure their own privacy needs in an easy-to-use way in order to stay in control of their personal data.

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 587–598, 2018.
https://doi.org/10.1007/978-3-319-91238-7_45

Since similar laws are already in place and GDPR becomes effective in 2018, many service providers already provide corresponding privacy controls for the users. However, we are facing a so-called privacy paradox [10]: Users frequently do not make use of these means, even if they say they want to [18] and have the opportunity to do so. For example, only about 20% of the European population fully read privacy notices and claim to understand how service providers use their data [5]. Unfortunately, the reasons for this paradox, their interrelations, and their underlying causes are not yet completely researched, and so far, the user's intention has not been addressed in a systematic manner [10].

1.1 Ideas and Contributions

In this paper, we present a study that investigates the reasons for not taking appropriate privacy actions. We asked more than 1,000 persons about their usage behavior regarding privacy settings and privacy notices, including the burdens they are facing. The results confirm that users rarely take available actions to protect their privacy. Half of the participants check their privacy settings only sporadically or never. Moreover, half of the participants state that they never read privacy policy notices at all, and only eight percent are reading them carefully for each service. The main reasons are similar in both cases: It takes too long to perform the privacy tasks, and the tasks are too complicated.

Before we can find solution strategies to mitigate these issues, we need to identify and understand the obstacles faced by the user. Based on the study results and previous investigations, we propose a generic intention model that contributes to the explanation of the privacy paradox. This model borrows concepts from psychology to explain the user's behavior regarding privacy interfaces. The core element in the model is intention, which is a combination of the user's motivation and different kinds of barriers. As privacy is a very individual need, we focus our work on the private user and assume that extrinsic motivation barely plays a role for privacy decisions regarding personal data. Thus, we focus on intrinsic motivation. In addition, the barriers depend on the individual user, since they arise if the user's resources do not meet the requirements emerging from the properties of privacy enhancing technologies (such as privacy settings, privacy notices). Obviously, the prerequisite for using privacy interfaces is that the user's own resources (e.g., security knowledge, cognitive load capacity and available time) exceed the required resources. If the required resources exceed the available resources, the resulting barriers will prevent the user from actually performing the actions. We identified various relevant resources for users and privacy interfaces, which we discuss in the paper. However, intention is not only a matter of available resources. Additionally, the user's motivation (i.e., cost-benefit ratio) is an important factor that needs to be considered. We claim that even if the user has a strong motivation to perform privacy actions, he frequently squanders the potential to optimize his privacy.

In this paper, we describe and interpret the user study about the user's behavior with respect to privacy-related actions in Sect. 2. Based on these results, we propose an intention model in Sect. 3. The model is exemplarily instantiated in Sect. 4. The paper ends with a discussion about related work in Sect. 5 and a conclusion in Sect. 6.

2 Usage Behavior Regarding Privacy Settings and Notices

As stated above, we mainly focus on two aspects: transparency and self-determination. While transparency focuses merely on information provision, self-determination enables users to actively control data usage. In practice, privacy notices are the most common means to ensure transparency, and many services provide privacy settings to give the user (some kind of) control. However, it remains an open question whether users really use these means and what the major burdens are for them.

2.1 Setup

One of the main goals of our study was to cover a cross section of society—i.e., to include also people without any special expertise in security and privacy. To this end, we integrated a survey in a public museum exhibition about privacy and data protection in Kaiserslautern, Germany. This gave us access to a wide range of people with different backgrounds that have at least a basic interest in security and privacy. The survey was included into an interactive security awareness quiz in order to provide an interesting exhibit. The exhibit setting prevented us from requesting text input, which limited us to questions with multiple choice answers. We used German language in the exhibit. In addition, we had little control over the participants, as they were not supervised when visiting the exhibit. For example, we cannot rule out the possibility that some visitors have participated multiple times, or dropped out early. Although this poses a threat to validity, the number of participants (1,391 within five month) minimizes the risk of invalid results.

2.2 Usage of Security and Privacy Settings

Our first question targeted the usage of security and privacy settings, as found in many online services. We asked the participants: "How often do you check your security and privacy settings?" As Fig. 1 shows, the results vary—however, only 41% 1,391 participants state that they check their security and privacy settings regularly (always or multiple times per year).

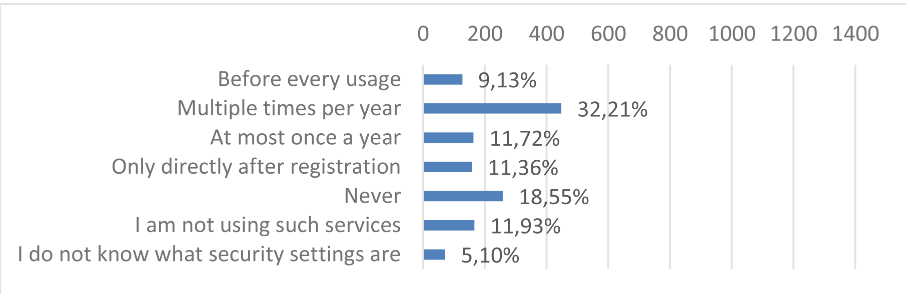


Fig. 1. How often do you check your security and privacy settings? (n = 1,391; one answer allowed)

As we were also interested in the reasons why users do not use security and privacy settings, we asked those participants that use these settings less than once a year: "Why don't you use security and privacy settings more often?" We did not ask this question to those participants who are updating privacy settings multiple times per year or more often, as we consider this behavior as acceptable. Figure 2 shows that most of the 558 participants are interested in general, but either do not think it is necessary to take action or find the provided tools too time consuming or complicated. Participants were allowed to choose multiple answers from the given five options.

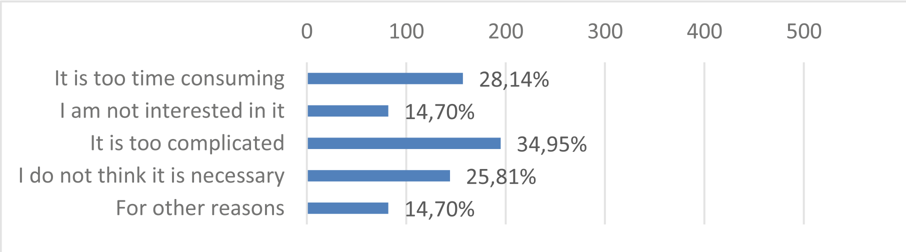


Fig. 2. Why don't you use security and privacy settings more often? (n = 558; multiple answers allowed)

This leads us to the conclusion that many users apply security and privacy settings in general—but only sporadically. The two major reasons are either that users do not think it necessary to do it more often or that the provided tools are too complicated and time consuming.

2.3 Usage of Privacy Notices

The second aspect we analyzed is the usage of privacy notices. They are an inherent part of most websites and services (also because they are partially required by law) and provide information about how a provider collects, processes and shares personal-related information. Thus, our question was "How often do you read online privacy notices?" Some participants stopped the survey before this question. As shown in Fig. 3, more than half of the 1,195 participants never read privacy notices at all. Another 25% reads them only in at most fifty percent of the cases. This means that—although privacy notices

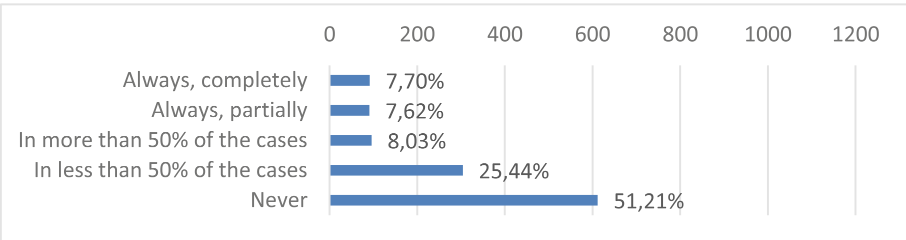


Fig. 3. How often do you read online privacy notices? (n = 1,195; one answer allowed)

are often the only information source regarding privacy—less than a quarter of the users actively use them.

As this result is quite unequivocal, we again asked for the reasons: “Why don’t you read online privacy notices more often?” Although the majority stated that they don’t read privacy notices, only 10% stated that they are not interested in them. The reasons for their inadvertence seem to be clear, as Fig. 4 shows: 72% of the 1,006 participants that do not regularly read privacy notices perceive privacy notices as too long. 43% also stated that privacy notices are too complicated. Multiple answers were allowed.

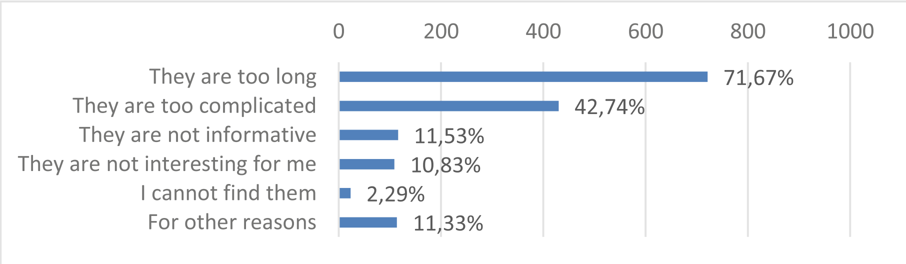


Fig. 4. Why don’t you read online privacy notices more often? (n = 1,006; multiple answers allowed)

2.4 Conclusion

The study shows that many users have a basic interest in transparency and privacy control measures. However, only a minority of users can make use of these measures. Privacy settings are only applied rarely, and they are perceived as too time consuming and complicated. The same applies to privacy notices, although the gap is even more severe. Although the majority of the users is interested, almost nobody reads the notices, as they are perceived as too long and complicated.

Overall, the findings underpin the privacy paradox: Users want to protect their privacy, but they do not take action regarding this respect. We want to understand and explain this effect in order to mitigate it as part of our future work.

3 An Intention Model Explaining the Usage Behavior

The study described above confirms previous studies [14, 15], according to which users, on average, take only moderate efforts to improve their privacy settings or to retrieve information on the use of their own data in the privacy notice. In many cases, this contradicts the user’s own need for privacy, which is one of the key drivers for performing privacy related activities. We consider the need for privacy as part of the humans’ basic needs in terms of safety and security [12]. We concentrate on those users who are not able to carry out these tasks (i.e., configuring privacy settings and reading privacy notices) appropriately despite their existing needs. Thus, we ignore potential unawareness of privacy issues (i.e., the lack of privacy needs). Lacking need for privacy could be compensated by awareness measures.

We developed an intention model (see Fig. 5) that abstracts existing problems (e.g., privacy paradox, too high complexity, too much time necessary) to a generic level. The model explains the discrepancy between the user's demand for the protection of his privacy (desired result) and the reality of the user ignoring his options of interaction (actual behavior).

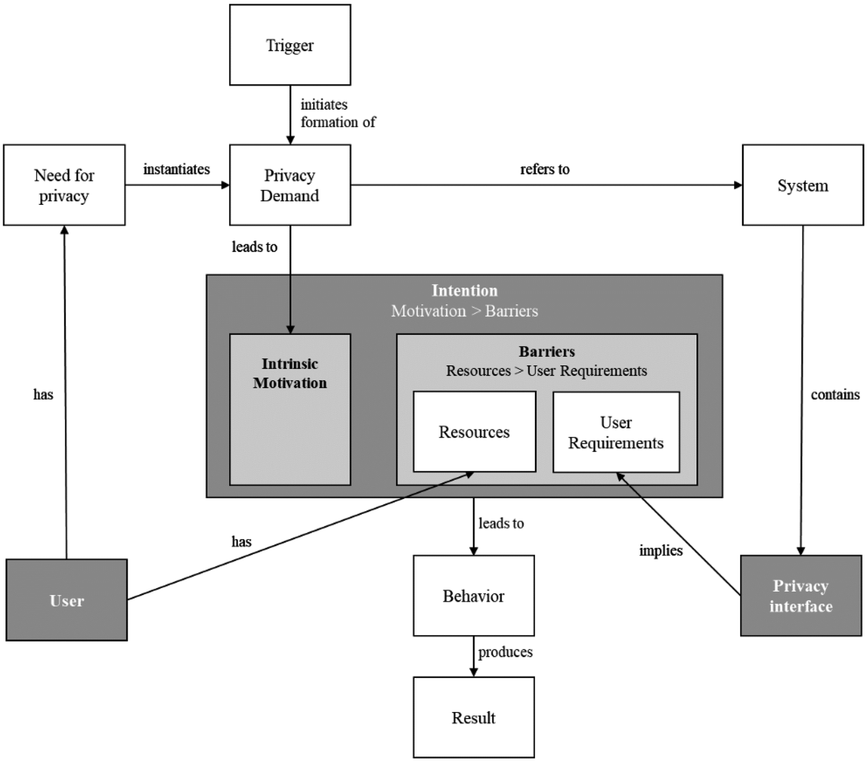


Fig. 5. Intention model

The baseline for our discussions is a private user who uses a system that processes personal-identifiable information. Processing includes collection, usage, distribution and sharing. As required by law and demanded by end-users, the system includes privacy interfaces. These privacy interfaces enable privacy-related information exchanges between the system and its users (e.g., privacy settings, privacy notices) and target transparency and self-determination. However, the utility of these interfaces depends on the behavior of the user. If the user does not use or does not want to use privacy interfaces, transparency and self-determination will not be achieved. Thus, we want to achieve a specific user behavior (i.e., usage of privacy interfaces) in order to obtain a result (e.g., specified privacy settings, understood privacy notices).

The actual behavior depends on the user's intention. The intention and its relationship to behavior is the focus of our work. Thus, we do not consider the quality of the result, as it is not directly depending on the intention.

In an ideal world, the user's intention is a direct consequence of his motivation. As the personal-identifiable information the system processes directly belongs or relates to the user, he typically has in intrinsic motivation to protect it.

Unfortunately, pure motivation is not the only factor influencing the intention. In addition, barriers come into play as a counterpart of motivation. Intention arises when the user's motivation exceeds the barriers he faces. The intention leads to the behavior of executing privacy-relevant actions. We will refine the barriers later and focus on the motivational part first.

The motivation for using privacy interfaces typically stems from situation-dependent privacy demands. These concrete demands are based on a general need for privacy and arise when the user experiences a certain trigger. The privacy demand could be, for instance, the desire to protect his personal data from abuse in a social network or to gather information about the data usage by third parties. In comparison to the need for privacy, the privacy demand does not describe a holistic need, but it refers to a certain system. Examples for a trigger are the use of a new service, a change in functionality or in the privacy notice of an existing service, or a request for additional personal data.

Table 1. User requirements vs. user resources

User requirement/ User resource	Description
Domain knowledge	Required vs. actual knowledge of the service's use cases and the personal data provided to the service necessary in order to be capable of making privacy-related decisions
Security & privacy knowledge	Required vs. actual knowledge of potential and actual use of personal data by the service and potential threats that arise from this use necessary in order to be capable of making privacy-related decisions
Technical knowledge	Required vs. actual knowledge of the functionality of the service and its privacy interfaces
Available time	Required vs. available time to apply the privacy interface
Cognitive capacity	Amount of privacy related information the user needs vs. is capable of processing simultaneously
Physical capacity	Required vs actual accessibility to a device that allows the use of a privacy interface in the respective system

As described above, barriers influence the intention. They emerge from the interrelation of the resources available to the user and the user requirements of the privacy interface. If the user has sufficient resources, he does not experience barriers. However, if the user's resources do not meet the user requirements, he experiences barriers towards using the privacy interface. As described above, the size of barriers does not directly determine the intention, but has to be exceeded by the motivation. The instantiation of the user resources and the user requirements and thus the identification of barriers strongly depends on the concrete system or privacy interface, respectively. In response to the question regarding the reasons for the moderate use of privacy settings, users responded by about 30% each that these are too complicated and time-consuming (cf. Sect. 2.2). Both reasons represent barriers to setting the privacy settings.

We identified multiple categories for requirements, resources and barriers resulting from a discrepancy between user resources and user requirements: Domain knowledge, security and privacy knowledge, technical knowledge, available time, cognitive capacity and physical capacity. In Table 1, we explain the trade-offs between user requirements and resources for each category.

Summarizing, our intention model explains the behavior of people who have a general need for privacy, but do not take appropriate actions to enforce it. Thus, the model approaches the privacy paradox. In the following, the model is instantiated for a specific application.

4 Case Study

We instantiated the behavior model for the two main privacy interfaces on Twitter: privacy notices and settings. Twitter provides different options that are relevant from privacy perspective. Most content (e.g., tweets, likes, shares) is public by default, and there are many privacy-relevant options to connect your contact book (e.g., from Gmail), get SMS notifications, and so on. Although Twitter's primary purpose is interaction with other users, and thus, the general need for privacy might be comparably low, profiling, tracking and customized advertisements can be strong motivators for privacy. Concrete

Table 2. Potential user barriers on Twitter

Barriers	Description
Domain knowledge	The user does not know or does not remember the provided personal information and does therefore not know what to specify
Security & privacy knowledge	The user does not understand how the personal data can be used by third parties in order to decide on the individual privacy settings
Technical knowledge	The user does not know about technical possibilities for tracking his usage behavior, for example via sensors on smartphones
Available time	As it is unclear which settings should be checked how often, the user would need to check all settings on every use, which is time consuming Privacy notice has approx. 4,000 words, and is not categorized according to user tasks
Cognitive capacity	The (privacy) settings overwhelm the user with many options and much textual information Information in privacy notices are distributed over the whole text and they do not relate to concrete user tasks (e.g., what happens when you tweet)
Physical capacity	Privacy settings can be done on mobile apps and browsers and are synchronized for all devices, which could be misleading (although explicitly stated) Privacy policies are hidden in app and not optimized for navigation on mobile devices

triggers for privacy demands can stem from the usage itself (e.g., visibility of sensitive tweets), reminders by Twitter (e.g., to update your phone number after login) and external triggers (e.g., press articles about Twitter).

On the other hand, we have barriers. The privacy notice is quite long (approx. 4,000 words), formulations are vague and information is distributed throughout the document. This increases the burdens for the users regarding cognitive load, needed time, etc. and prevents them from reading the privacy notice. In addition, the settings are distributed over 15 categories, which makes it time and effort consuming to maintain them. In response to the question regarding the reasons for the moderate use of privacy settings, users responded by about 30% each that these are too complicated and time-consuming. Both reasons represent barriers to setting the privacy settings. In Table 2, we show examples for burdens we identified in the categories presented in Sect. 3.

Of course, this instantiation is not a comprehensive evaluation and lacks certain details, as we could not perform large-scale user studies regarding Twitter's privacy interfaces. This is part of our future work, in which we analyze the applicability and completeness of our model in depth.

5 Related Work

Studies regarding the frequency of use of privacy notices have been performed by Moallem [14] and Obar and Oeldorf-Hirsch [15]. Our study confirms the results of these studies, but has eight/2.5 times more participants, respectively, cross-sectional through society. To improve the acceptance of privacy notices, there exists work targeting readability [4, 13], understandability [17] and design [20] of privacy policies. All these improvement aspects are important and could benefit from our intention model as a baseline for requirements. The consequences of lacking acceptance of privacy notices have been analyzed in different surveys [15, 18, 19]. This is relevant for our work insofar as the (expected) consequences affect on the user's intention towards privacy notices.

Boyd investigated reasons for users not to configure privacy settings in Facebook [3]. She found that both frequency and type of Facebook use as well as Internet skill influence the user behavior regarding privacy settings configuration. This underpins the prominence of the barriers in our model, as the increase of knowledge may influence the user's behavior. Research was also carried out to improve the usability of privacy settings and policy specification, respectively. Johnson, Karat, Reeder et al. proposed guidelines for the implementation of usable policy authoring interfaces [8, 16]. Their work includes amongst others the following guidelines, which we consider as relevant input for the user requirements of privacy interfaces: Limitation of expressivity, consistent terminology and communication of threats and risks. Ben-Asher found out that users behave differently with respect to system usage, if they are prompted with security warnings [2]. They behave more cautiously and adjust security settings more frequently if they are triggered appropriately. Liu et al. investigated the discrepancy between desired and actual privacy settings in Facebook [11], i.e., problems users are facing when their positive intention already made them specify their privacy requirements.

Our intention model was inspired by theories and models that try to explain human behavior, but are not focused on privacy. The key element 'intention' was inspired by the theory of planned behavior (TPB) by Ajzen [1] and by the behavioral model for persuasive design by Fogg [7]. The elements 'perceived behavioral control', 'intention' and 'behavior' of the TPB are included in our model. The perceived behavioral control is part of what we call 'barrier', intention is equivalent to 'motivation' and the term 'behavior' is used in the same way. Fogg's behavioral model inspired the interrelation of motivation and barriers. Fogg's model and especially its graphical representation illustrates that motivation need to be higher than the so-called simplicity factors. These simplicity factors are a positive formulation of barriers [8]. The element 'need for privacy' was inspired by the well-known hierarchy of needs by Maslow [12]. According to Maslow, "The organism is dominated and its behavior organized only by unsatisfied needs [12] ". We consider the need for privacy to be a subset of Maslow's need for safety/security. In our model, we assume that users whose need for privacy is satisfied will not take action for protecting their privacy.

The Privacy Paradox describes the dichotomy between the need for privacy and the actual behavior of users with respect to taking privacy-relevant actions. Kokolakis et al. conducted a meta study in order to summarize all findings from the state of the art regarding the privacy paradox [10]. They outlined multiple explanations for this phenomenon, but not the challenge of mastering barriers that we claim in our paper and which is underpinned by our study results.

6 Conclusion

In this paper, we presented the results of a study regarding the situation of users dealing with privacy-related actions, proposed an intention model to explain the findings of our study and applied this model to an example application. In our study, we asked more than one thousand visitors of a museum exhibition on security and privacy how strictly they carry out privacy-related actions. Half of the participants check their security and privacy settings in online services only once a year or less—mainly because those settings are too time consuming or too complicated. Regarding the attention to privacy notices of online services, only half of the participants read them at all, although only 10% claimed that they are not interested in them. Similar to settings, participants perceived privacy notices as too time consuming/long (70%) and too complicated (41%). We conclude that many users have a basic interest in privacy-related actions (privacy settings and privacy notices). However, a significant number of participants encounter barriers when it comes to taking actions regarding privacy.

Therefore, we wanted to find out what those barriers are and how they influence and explain the user's behavior of not taking appropriate privacy actions. To this end, we developed our intention model. The model explains the relationships between the user's need for privacy, his motivation, his intention and the resulting behavior of performing privacy actions. Mainly, we reason about how discrepancies in the user's resources and the usage requirements of the privacy interfaces lead to barriers that prevent users from performing privacy-related actions, regardless of his motivation. We defined the user's intention to use privacy interfaces as his motivation being high enough to overcome potential barriers. Our model is based on observations and experience regarding users carrying out privacy actions, as well as on the results of the study described in this paper. It respects and partially reuses terms and relations from other psychological models.

In order to obtain first evidence for the applicability of the model, we instantiated it for Twitter's privacy interfaces. The case study shows that the model can explain correlations between user behavior, the users' need for privacy and their intentions. Obviously, this instantiation is not yet a comprehensive evaluation of our model. We will apply the model in a large-scale study in order to validate its correctness in near future. We plan to let different user types who have different sets of resources use different types of privacy interfaces. Next, we will correlate user behavior, user acceptance of the privacy interfaces and correctness of results in order to obtain evidence of potential barriers as well as to derive mitigation strategies for preventing barriers for specific user types.

Acknowledgements. The research presented in this paper is supported by the German Ministry of Education and Research projects "Nationales Referenzprojekt für IT-Sicherheit in der Industrie 4.0" (IUNO) (grant number 16KIS0328). The sole responsibility for the content of this document lies with the authors.

References

1. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**(2), 179–211 (1991)
2. Ben-Asher, N., Meyer, J., Moller, S., Englert, R.: An experimental system for studying the tradeoff between usability and security. In: International Conference on Availability, Reliability and Security, pp. 882–887 (2009)
3. Boyd, D., Hargittai, E.: Facebook privacy settings. Who cares? First Monday **15**(8) (2010)
4. Ermakova, T., Fabian, B., Babina, E.: Readability of privacy policies of healthcare websites. In: Wirtschaftsinformatik (2015)
5. European Commission: Special Eurobarometer 431 - Data Protection (2015). http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_en.pdf. Accessed 15 Feb 2018
6. European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016). http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679. Accessed 15 Feb 2018
7. Fogg, B.J.: A behavior model for persuasive design. In: Proceedings of the 4th International Conference on Persuasive Technology, p. 40. ACM (2009)
8. Fogg, B.J.: What causes behavior change? (2016). http://www.behaviormodel.org/index.html. Accessed 15 Feb 2018
9. Johnson, M., Karat, J., Karat, C.-M., Grueneberg, K.: Optimizing a policy authoring framework for security and privacy policies. In: Cranor, L.F. (ed.) Proceedings of the Sixth Symposium on Usable Privacy and Security. The Sixth Symposium, p. 1, Redmond, Washington, New York, NY. ACM (ACM Digital Library) (2010)
10. Kokolakis, S.: Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. Comput. Secur. **64**, 122–134 (2017)

11. Liu, Y., Gummadi, K.P., Krishnamurthy, B., Mislove, A.: Analyzing Facebook privacy settings: user expectations vs. reality. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 61–70. ACM (2011)
12. Maslow, A.H.: A theory of human motivation. Psychol. Rev. **50**(4), 370 (1943)
13. Milne, G.R., Culnan, M.J., Greene, H.: A longitudinal assessment of online privacy notice readability. J. Public Policy Mark. **25**(2), 238–249 (2006)
14. Moallem, A.: Do you really trust "privacy policy" or "terms of use" agreements without reading them? In: Advances in Intelligent Systems and Computing, vol. 593, pp. 290–295 (2018)
15. Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services. In: The 44th Research Conference on Communication, Information and Internet Policy (2016)
16. Reeder, R.W., Karat, C.-M., Karat, J., Brodie, C.: Usability challenges in security and privacy policy-authoring interfaces. In: Baranauskas, C., Palanque, P., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4663, pp. 141–155. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74800-7_11
17. Reidenberg, J.R., Breaux, T., Carnor, L.F., French, B., Cranor, L.F., Grannis, A., Graves, J.T., Liu, F., Mcdonald, A., Norton, T.B., Ramanath, R., Russell, N.C., Sadeh, N., Schaub, F.: Disagreeable privacy policies: mismatches between meaning and users' understanding. Berkeley Technol. Law J. **30** (2014)
18. Symantec: State of Privacy Report (2015). https://www.symantec.com/content/en/us/about/presskits/b-state-of-privacy-report-2015.pdf. Accessed 15 Feb 2018
19. Tsai, J., Egelman, S., Cranor, L., Acquisti, A.: The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study (2007)
20. Waldman, A.E.: Privacy, Notice, and Design (2016)

Digital Breadcrumbs: A Lack of Data Privacy and What People Are Doing About It

Carley Ward(✉), Dana Ellis, Lisa A. D'Ambrosio, and Joseph F. Coughlin

Massachusetts Institute of Technology AgeLab, Cambridge, MA, USA
carleyw@mit.edu

Abstract. With the seemingly ubiquitous presence of technology, interactions and transactions constantly take place over websites, apps, text, and email. Despite the convenience and efficiency of these modes of communication, these online interactions give many parties access to personal information and can leave individuals' information vulnerable to misuse and attack. As people use electronic devices for more and more tasks, they leave behind evidence of their activities, a trail of digital breadcrumbs providing behavioral and demographic information about where they go, whom they talk to, what they do, and what they believe. Consumers are often unaware of who has access to these data and how they might be used. In this mixed methods study, researchers conducted a series of surveys and focus groups in the United States to examine the everyday digital breadcrumb trail people leave behind when using technologies on a daily basis, people's awareness of this trail of data, and what measures they take to protect their digital information and prevent it from being collected or misused. Participants discussed their attitudes about data privacy, sharing, and trust. Implications for research, business, and policy are provided.

Keywords: Data privacy · Digital breadcrumbs · Mixed methods

1 Introduction

With the seemingly ubiquitous presence of technology, interactions and transactions constantly take place over websites, apps, text, and email. As of 2017, 77% of American adults owned a smartphone [1]. In one study of smartphone users, the average participant used their phone 76 separate times per day [2]. Despite the convenience and efficiency of these modes of communication, online interactions can leave individuals' information vulnerable to misuse and attack.

As people use electronic devices for more tasks, they leave behind evidence of their activities, a trail of digital breadcrumbs providing information about where they go, whom they talk to, and what they do. The fear of credit card information, transaction history, location history, home address, and email and bank account passwords being accessed or used unethically is not unjustified. In one study, 46% of smartphone users reported that they have had a company take advantage of their data and use it for something that they had not agreed to [3]. The present study examined the everyday digital breadcrumb trail people in the United States leave behind when using

© Springer International Publishing AG, part of Springer Nature 2018
M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 599–612, 2018.
https://doi.org/10.1007/978-3-319-91238-7_46

technologies on a daily basis, people's awareness of this trail, and attitudes about data privacy, sharing, and trust.

2 Literature Review

2.1 Use of Technology and the Trail It Leaves Behind

As the technology business has boomed, people have come to use their computers, tablets, and smartphones for an increasing number of tasks. In a 2015 Pew Research Center study, smartphone owners in the US had used their phones in the past year to look up information about a health condition (62%), do online banking (57%), search real estate listings (44%), research information about a job (43%), look up government services or information (40%), and get educational content (30%) [1]. In addition, smartphone owners used their phone at least occasionally to follow breaking news (68%), share pictures or video about community events (67%), learn about community events (56%), get turn-by-turn navigation while driving (67%), and get public transit information (25%) [1]. Dscout (2016) reported that the average Android user touched their phone screen 2,617 times per day, with the most touches occurring on Facebook (15%), Messages (11%), the Home Screen (9%), and Chrome (5%) [2].

The Internet of Things provides abundant opportunities for convenience and connectivity, but it also creates new cybersecurity issues. The Internet of Things includes a variety of smart home or connected home devices including smart refrigerators, doorbell cameras, keyless locks, thermostats and smoke detectors, lights and blinds, and televisions. In one study, 90% of consumers cited personal or family safety as the top reason for smart home technology adoption [4]. The same study reported that 71% of consumers feared that their personal information may be stolen, and 64% feared that their data would be collected and sold [4]. These fears are not without warrant; most devices do not provide a secure mode of operation. In one study of 50 smart home devices, none of them enforced strong passwords, used mutual authentication (where individuals prove their identity to the server, and the server proves its identity to the individual), and almost 20% of apps used to control said devices did not encrypt communications to the cloud [5]. In 2014, hackers took advantage of these security gaps; there was a large-scale attack on smart home devices, including TVs and refrigerators, in which hackers were believed to have accessed more than 100,000 devices [6].

There are opportunities to collect data about the way people move about their communities as well. Drivers frequently pass cameras while driving, which record timestamps and license plate numbers. Additionally, many drivers encounter toll roads and use registered electronic toll passes. Similarly, many vehicles have built-in GPS which monitor drivers' progression on a route. Other drivers use smartphone navigation or map apps for directions. For example, the traffic and navigation app Waze collects and shares data with local governments [7]. Some cars collect and transmit data on driver speed, location, and braking. Tesla records video in their vehicles on the road to improve their autopilot system. Tesla's privacy policy states that they can share data with business partners, service providers, and other parties [8]. Outside of the car, public transit pass usage can be recorded, and some passes are registered with name and credit card information connected. Similarly, some public transit passengers use apps to identify transit arrival times, and some pedestrians use GPS or map apps to identify efficient walking directions.

In the context of health, digital data are collected on fitness trackers or other wearable devices (e.g., Fitbit). Organizations can use health data collected from fitness trackers and apps including location, activities, sleep patterns, biometric data, and reproductive health data for their own purposes [9]. For example, the Australian supermarket Coles encourages consumers using health trackers to link them to their loyalty cards by providing "loyalty points for walking 10000 steps per day" [9]. Insurance companies and employers may also see uses for health data.

All of these activities using electronic devices leave evidence behind. Most service providers, apps, and online companies collect and use the information people leave behind while using their product. For example, Facebook, which received the most touches of any app [2], records information from users' posts, messages, photos, videos, birthdays, relationship statuses, schools attended, hometowns, check-ins, and everything they have viewed, liked, or shared. Each touch is a decision which provides information about the user. Many more companies also require people to allow them access to other digital data in order to gain access to the product. For example, go to download the popular social media app Snapchat and a box will pop up to notifying the user that Snapchat requires access to in-app purchases, identity, contacts, location, SMS, photos/media/files, camera, microphone, Wi-Fi connection information, Bluetooth connection information, and device ID and call information. Accept?

2.2 Data Privacy or Lack Thereof

Privacy is an individual, group, or institutions' right to determine when, how and to what extent information about themselves is shared with others [10]. By engaging with online entities and accepting privacy agreements, individuals often unknowingly give up their right to privacy. Sixty percent of millennials report that they would be willing to share data about their preferences and behaviors with marketers, and 30% would be willing to provide even the most private data in return for discounts [11]. Not all age groups are equally willing to share their data, however. Only 13% of baby boomers would be willing to share private data in exchange for discounts. Additionally, 63% of US adults reported that they feel concerned about their privacy and security when they access the internet on their cell phones [3]. Likewise, 78% believe that it is difficult to trust companies when considering how they use consumer data and feel that service providers have too much information about consumer preferences and behaviors [12]. Only 44% of millennials believe companies they interact with keep their personal information private [13].

Once the apps, services, and websites collect user data, they are often able to give others access to these data. A study of apps in the US, Australia, Brazil, and Germany found that between 85% and 95% of free apps and 60% of paid ones connected to third parties that collected personal data [14]. In a study of 200,000 participants who visited 21 million web pages, third party trackers were present on 95% of those web pages [15].

Furthermore, retail, travel, media, telecommunications, and finance companies also collect and use digital data to develop products or services to improve marketing and sales and to personalize advertising and discounts. In addition to using these data for their own purposes, organizations frequently sell or share these data with other parties. Google (2017), for instance, reported that it captures "approximately 70% of credit and debit card transactions in the US" through third-party organizations [16]. Companies and data brokers use information such as demographics, credit history, medical history, and browser history to create profiles, which can be used for classification, estimation, or prediction [17]. Companies can use these profiles to target individuals with certain characteristics, influence behavior, predict their health risks or credit risk, or determine if they are desirable employees, good tenants, or valuable customers [18]. For example, US consumer reporting agencies provide reports to employers, landlords, insurance companies, and government agencies. Reports for employers may contain information about an individual's employment, salary, education, driving history, health, drug testing information, and fingerprints. Generally speaking, anyone can purchase lists of personal data of nearly every demographic. For example, ExactData.com has premade lists of "Americans with Bosnian Muslim Surnames" and "Unassimilated Hispanic Americans" [19].

Even when data are intended to be confidential and secure, hacking and stolen data are commonplace. For example, the data breach of the consumer credit reporting company Equifax, which began in May 2017 and continued until it was discovered in July, was not made public until September. Over the course of these two months, more than 145 million records were accessed, with information such as names, Social Security numbers, birth dates, addresses, driver's license numbers, credit card numbers, and certain dispute documents [20]. Further, Equifax experienced smaller breaches in 2016, 2012, 2010, and 2006 [20].

3 Methods

The present study examined the digital breadcrumbs that people leave behind, people's awareness of this trail, and attitudes about data privacy. Data were collected in two parts. First, we conducted a multipart survey (N = 36) mapping participants' technology use over a 24-h period, as well as collecting information on their technology adoption and use, methods they use to protect their data, and attitudes toward data privacy and sharing. Next, we conducted focus groups (N = 14) exploring what people know about the privacy of their data, how much they trust companies collecting digital data, and what factors come into play when deciding to use these companies.

3.1 Participants

Survey respondents were 36 (47.2% women) self-selected adults from across the US, ages 20–65 ($M = 42.9$), recruited from the MIT AgeLab Volunteer Database. Respondents were 83.3% White, 13.9% Asian, 2.8% Black, 2.8% Latino, and 2.8% other. Respondents ranged in technology savviness from 2.6 to 5 ($M = 3.9$) on a 1–5 scale where 5 indicated a high level of technology savviness. All respondents had at least some college education and over half of all participants (52.8%) had a total annual household income of at least $100,000 (see Table 1).

Table 1. Survey respondent demographics (N =36).

Category	Characteristics	Percent
Age	20–29	22.2
	30–39	25
	40–49	11.1
	50–59	22.2
	60–69	19.4
Gender	Men	52.8
	Women	47.2
Race	White	83.3
	Asian	13.9
	Black	2.8
	Latino	2.8
	Other	2.8
Education	Some college	11.1
	Trade/technical school or associates degree	2.8
	College Degree	22.2
	Some post-graduate work	16.7
	Post-graduate degree	47.2
Household income	Less than $25,000	13.9
	$25,000–$49,999	11.1
	$50,000–$74,999	13.9
	$75,000–$99,999	8.3
	$100,000–$149,999	22.2
	$150,000 or more	30.6

Focus group participants were 14 (71.4% women) Boston-metro area residents ages 23–59 (M = 48.9) recruited through local website postings. Participants' technology savviness scores ranged from 2.2 to 5 (M = 3.5).

3.2 Measures and Procedure

Surveys. Respondents completed a total of six online questionnaires over three days. The first questionnaire was an introductory survey with items about technology ownership and use (smartphone, cellphone, landline, computer, tablet, e-reader, GPS, credit or debit card, rewards or loyalty card, transit or toll pass, smartwatch, Fitbit or wearable, personal assistant device), smart home device ownership (smart thermostat, home security system, wireless doorbell camera, keyless entry, smart smoke/carbon monoxide detector, smart outlets/plugs, automatic lighting), and demographic information. Next, respondents completed one questionnaire every six hours over a 24-h period. These four surveys were identical and were used to document when respondents engaged in certain digital activities throughout the day, on which devices (smartphone, computer, tablet, smartwatch, and Amazon Alexa), using which software. Activities tracked included texting, social media use, making or receiving calls, email use, visiting websites, viewing maps, listening to music, reading, doing online banking, taking or viewing photos, video chatting, and online shopping. Finally, respondents completed a follow up survey exploring their attitudes, beliefs, and knowledge about data sharing, how they attempt to protect their data, and to what extent they trust in the companies and organizations that collect and use their data.

Focus Groups. Participants first completed a short online questionnaire providing information about demographics, technology savviness, and technology use. We conducted two focus groups over two days. Each focus group lasted two hours and included discussion led by two facilitators around the following topics: what kinds of data they believed were shared or vulnerable; what their data were being used for; what measures they took to protect their data; benefits and drawbacks of data sharing; attitudes about the sharing economy; and trust in organizations and brands. Focus groups were audio and video recorded.

4 Results

4.1 Technology Use and Data Sharing

Survey participants reported using their smartphones, computers, tablets, smartwatches, and smart thermostats throughout the 24-h data collection period (see Table 2). These devices were used for a variety of activities including texting, calls, social media, email, websites, maps, music, reading, banking, photos, video chatting and purchases. Smartphones were the most frequently used devices across time periods, with the most common activities being texting, social media, calls, visiting websites, and emails (see Table 3).

Table 2. The percentage of respondents who owned each device and who used each device in each time period.

Device	Sample percent (n)	Time period			
		12pm–6am	6am–12pm	12pm–6pm	6pm–12am
Smartphone	100(36)	36.1	86.1	88.9	80.6
Computer	94.4(34)	14.7	61.8	76.5	58.8
Tablet	50(18)	5.6	22.2	27.8	22.2
Thermostat	16.7(6)	16.6	16.6	33.3	16.6
Smartwatch	13.9(5)	20.0	20.0	20.0	20.0
Amazon Alexa	11.1(4)	0.0	25.0	0.0	25.0

Table 3. The percentage of respondents who used smartphones for each activity in each time period (N = 36).

Activity	Time period			
	12a–6am (n = 13)	6a–12pm (n = 31)	12p–6pm (n = 32)	6p–12am (n = 29)
Texts	53.8	83.9	96.9	82.8
Social media	30.8	45.1	56.3	51.7
Calls	0.0	54.8	68.8	41.4
Emails	38.5	71.0	56.3	44.8
Websites	46.2	64.5	31.3	37.9
Maps	7.7	32.3	21.9	17.2
Music	7.7	0.0	12.5	17.2
Reading	7.7	32.3	21.9	10.3
Banking	0.0	3.2	3.1	6.9
Photos	0.0	3.2	12.5	3.4
Video chat	0.0	0.0	3.1	3.4
Purchases	0.0	7.7	0.0	0.0

The constant use of these devices leaves an abundance of digital breadcrumbs with information about individuals' activities, locations, purchases, interests, and contacts; such information may be used by the company collecting it or sold to other companies. In addition to using personal electronic devices, respondents were sharing data while driving, taking public transportation, using rewards or loyalty cards, and making non-cash purchases (see Table 4).

Table 4. The percentage of respondents who engaged in each activity in each time period (N = 36).

Activity	Time period			
	12pm–6am	6am–12pm	12pm–6pm	6pm–12am
Driving	5.6	38.9	47.2	30.6
Public transit	0.0	5.6	2.8	5.6
Rewards card	0.0	13.9	13.9	2.8
Non-cash purchases	0.0	30.6	22.2	5.6

Similarly, most focus group participants reported using their cell phones constantly; as one 65-year-old man noted, "I never turn it off." One 25-year-old woman described herself as being on her phone "Most minutes of the day" for music, podcasts, and texting and using her tablet for watching Hulu and Netflix. Another (25-year-old woman) explained, "At least every hour of the day that I'm awake I'm using it [smartphone] in some capacity." A 39-year-old woman used her phone even when she was sleeping to play a white noise app from Spotify. Only three participants said that they were not constantly on their cell phones.

4.2 Beliefs About Data Sharing

Ninety-seven percent of survey respondents reported that they believed information about what they did online is collected, used, or viewed by others on a typical day. They believed that several different groups were collecting their digital breadcrumbs, including companies whose websites they visit, companies they purchase from, their internet provider, the US government, and hackers (see Table 5).

Table 5. Organizations which respondents believed to be collecting their digital data on a typical day (N = 36).

Organization	Percent
My internet service provider	77.8
Companies whose websites I visit	100
Companies I purchase from, have accounts with, subscribe to	94.4
Companies I don't purchase from, have accounts with, subscribe to	58.3
US Government	77.8
US Non-Governmental Organizations	52.8
Hackers	75.0

A majority of survey respondents reported that they believed many types of data were being collected about them, including their browsing history (100%), location (100%), purchase history (100%), email (97.2%), name (97.2%), social media activity (94.4%), IP address (88.9%), physical address (83.3%), phone number (83.3%), birthdate (80.6%), credit card information (69.4%), photos (66.7%), and passwords (61.1%).

Focus group participants also perceived that data about their digital behavior were being collected. Many made statements such as: "It's the new reality that everything is being tracked" (39-year-old woman); "Everything, literally everything [is being tracked]. Any possible way they can make money (39-year-old, woman); and "Everything [website] you visit is going to be collecting data on you" (23-year-old man). Participants listed many devices other than their smartphones that collected data on them throughout the day, including security cameras, Internet of Things devices, loyalty cards, GPS, cars, red light cameras, and credit or debit cards. A 39-year-old woman summed it up by saying, "I'm trying to think of where in modern life, if you're living in an urban area, where you could go where you're not being tracked."

Pros and Cons of Data Sharing. Survey respondents and focus group participants believed there were pros and cons to data sharing. Seventy-five percent of survey respondents said they felt that there were benefits to the collection and use of their digital data. The advantages they saw included helping companies develop better products and services, creating better website experiences, and receiving personalized information, such as customized ads, marketing campaigns and offers for things they might be more interested in purchasing. Further, many reported that they would feel comfortable with companies using their data if those companies offered them discounts for products and services (42%) or if it made things easier or more convenient (39%) (see Fig. 1).

IIn addition to these benefits, focus group participants discussed a variety of others. Many focused on the fact that data sharing offers convenience and saves them time and money. A 69-year-old woman in a focus group commented, "I started using Waze a few years ago even to go a few miles because it navigates you around the traffic. I tend to use it fairly frequently, and I guess I don't mind the tracking that goes with it." A 65-year-old man explained, "I guess it's a small price to pay for the convenience that if I'm looking for something, I can find the lowest price and go buy it, rather than go to the store and go to another store and the time spent is unbelievable!" A 64-year-old woman added that companies would send her discounts for items that she had viewed on their websites, but had not purchased, so now when she does online shopping, she puts items in the cart, and then waits for an email with a coupon to arrive.

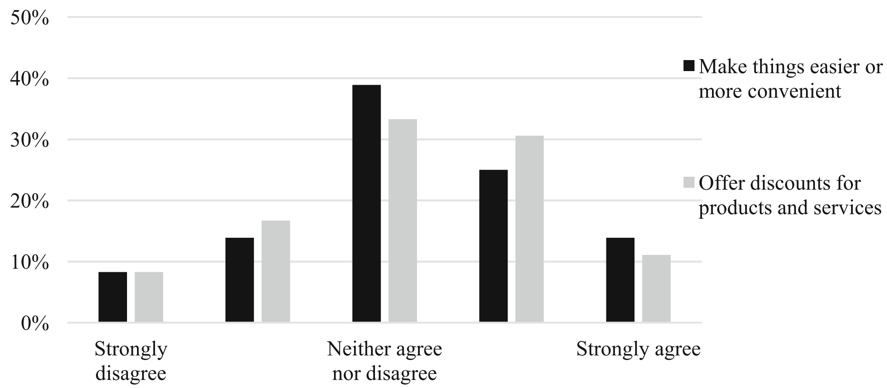


Fig. 1. Survey respondents: I would be comfortable with companies using my data if they (N = 36).

Focus group participants also identified a peace of mind that came with having the information that accompanies a digital transaction. One 26-year-old woman said, "If something bad happens in a cab or if I lose something in a cab, the chances of me ever being able to find that driver, to identify them, are pretty slim, whereas in an Uber you're being tracked by GPS, you know the name and contact information of the driver right away". The same theme was echoed around finances. Several participants remarked that because their banks and credit card companies track their spending patterns, they would automatically freeze participants' accounts if any red flags occurred and call them to make sure their information had not been compromised. As one 39-year-old woman explained, "I know that my bank has my spending information, if that's the protection that I'm gonna get, I'm comfortable with them having my information".

In addition to personal benefits, one group discussed the societal benefits of collecting both individual and aggregated data. For example, a 65-year-old man mentioned how law enforcement and the legal system use evidence and the digital traces from electronic devices in criminal cases. He said, "I don't have a problem with it. In fact, sometimes I think it keeps us safer… I mean, unless I commit a crime, so what?" Another 50-year-old woman identified potential health benefits: "they can actually monitor disease progression and outbreaks by what people are googling and searching". She continued to note potential benefits around traffic flow, explaining how the Massachusetts Turnpike could determine the ideal time to schedule road construction based on EZPass data showing "when there's the fewest people using this particular stretch of road".

Despite the benefits identified, all of the survey respondents and focus group participants agreed that there were disadvantages to data sharing. Some of the disadvantages identified by survey participants included companies exploiting them for financial gain, fraud, risk of a breach of confidential information, identity theft, loss of privacy, and receiving junk advertising. Only 22.2% believed that companies used their data responsibly, and 86.1% believed that companies made money from their data. A majority reported that it was very or extremely important to protect their social security number (100%), passwords (100%), credit card information (100%), bank account information (100%), address (86.1%), phone number (75.0%), photos (75.0%), birthdate (69.5%), IP address (63.9%), purchase history (58.4%), and geographic location (55.6%).

Focus group participants also felt exploited or thought that having data constantly collected about them was creepy. One 39-year-old woman passionately exclaimed, "they're taking literally all of the information, every parameter they possibly can, and shooting it out any possible way they can to make money!" Others expressed discomfort with the fact that many privacy agreements allow companies, once they have collected data, to sell it to any number of other unidentified organizations. A 47-year-old man put it succinctly: "I feel like I'm being big brother-ed." A 69-year-old woman provided an example, "I was eating in some obscure restaurant, and then up came something, how did you enjoy your experience at this place, and I thought 'How do they know I'm here?' You know, that kind of thing is just creepy to me because I didn't put in that I was there, but it knew".

4.3 Data Security and Stolen Data Experiences

Focus group participants discussed using a range of methods to try to protect their information. Some reported using physical protections such as covering the keypad when entering their pin number or placing tape over their webcam. One 39-year-old woman said, "I have the tape right over it [webcam] and then when I saw the thing that said Zuckerberg has his microphone cut off somehow I was like ooh, how do I do that?" Others used process protections such as a VPN or private browsing, turning off location services on their devices, and creating strong passwords. One 23-year-old man said, "I'm careful what websites I go to avoid viruses. For example, I'll use a VPN if I'm on public wifi". Still others said they did not do anything. Participants who did not make efforts to protect their data either were not aware of what to do or felt their actions were futile. As one summed it up, “It’s a crap shoot. If it’s gonna get hacked, it’s gonna get hacked” (65-year-old man).

Survey respondents were similarly not very intentional with data security. Twenty-five percent of respondents used a service to monitor fraud or identity theft. The majority changed their passwords only when a website required it (33.3%) or when they couldn’t remember their password (27.8%), and most cleared their browsing history or internet cache on their phones if they happened to think of it (55.6%) or had never done so at all (16.7%). In addition, only 1.9% reported using a VPN sometimes or always when they connected to public wifi networks. However, nearly all respondents reported that it was extremely important to protect their passwords (91.7%), bank account information (88.9%), credit card information (88.9%), and social security number (83.3%). To this point, the majority of respondents did report taking a few common digital security measures such as using dual authentication methods (52.8%) and Spyware or anti-virus software on their computers (65.7%). They were much less likely to have Spyware or anti-virus protections on their phones (19.4%) or tablets (16.7%).

Stolen Data and Identity Theft. Unfortunately, having digital information stolen or misused is common. Many survey respondents reported having their email (44.4%), credit card (36.1%), or debit card (19.4%) information hacked, or even their identity stolen (22.2%). Overall trust in organizations to keep personal data secure was very low, with only 36.1% trusting companies they use, 30.5% trusting their internet service provider, 25.0% trusting the US government, and 11.1% trusting companies whose websites they visit. When asked to think about specific companies they used, however, reported trust was often much higher. A majority of respondents trusted their digital information with their bank (88.9%), credit card company (77.8%), healthcare provider (77.8%), home or auto insurance company (63.8%), pharmacy (58.3%), and Amazon.com (55.6%).

Nearly all focus group participants also reported their data had been compromised at least once through large data breaches, hacked ATMs, identity fraud, phone scams, or stolen credit cards. A 55-year-old man explained, “One time it was identified as BJ’s Wholesale Club, another time it was TJX where there were massive breaches. You can get it stolen anywhere”.

When asked how worried they felt about having their data stolen, focus group participants and survey respondents diverged. Fifty-three percent of survey respondents reported they were very worried about their personal information being stolen or hacked and 55% of said that they did not have control over what information they shared digitally. Interestingly, despite the frequency of the misuse of digital data, many focus group participants were not overly concerned about being hacked. In one group, when asked directly, *do you worry about being hacked*, not a single participant was more worried than “In the middle”. Participants made statements such as “Can’t waste your time worrying about it” and “No point in worrying, there’s nothing you can do”. Additionally, participants reported that they had not changed their purchasing or technology use habits since experiencing hacks. A 26-year-old woman who had fraudulent tax returns filed in her name said that subsequently “literally nothing has changed”. This may be because when data are stolen, issues are often quickly resolved by other entities, namely banks and credit card companies. A 39-year-old woman concluded that "part of the reason why I don't stress out about it anymore. is that banks seem to really be on top of it."

5 Discussion and Conclusion

The present study examined the everyday digital interactions of people in the United States, their awareness of how data can be collected and used, and attitudes about data privacy, sharing, and trust. In review, people used electronic devices for tasks in a variety of domains all day and often through the night, leaving a long trail of data about their behaviors. Many knew that this phenomenon was occurring, and they were able to identify types of information being collected and some of the agents collecting said data. Individuals used a variety of methods to attempt to protect their data, but most measures were not used consistently or across devices, and there remained feelings of futility around these actions.

Respondents' use of their smartphones mirrored that of participants in other studies, primarily using their devices for social connection and messaging, engaging most frequently with texts, calls, and email [1, 2]. While much research examines how we use our smartphones, this study sheds new light to how we use other devices, how these other devices (including computers, tablets, smartwatches, and other smart home technologies) add to the density of digital data disseminated, and how people sought to protect their personal data from being used for unknown purposes by unknown agents.

As technologies continue to evolve and become ever more sophisticated, a tradeoff for the convenience and connectivity such electronic devices offer often means sharing data with other parties, resulting in decreased privacy and increased risks for individuals. These risks are not simply due to the exposure of financial data; adoption rates of smart home devices, fitness trackers, and other electronics that track everyday behaviors are on the rise. In 2016 there were 5.4 million multifunction or whole smart home systems in homes in North America [21], 72% of North Americans were Facebook users [22], and 10% of Americans owned a fitness tracker [23]. As technologies that can track individual actions and behaviors become ubiquitous, it is increasingly important that people be aware of the extent to which their every action is collected and shared, by whom and with whom, so that they can make informed decisions about whether and how they protect their data. It may be difficult to persuade individuals to adopt security measures if they believe that their efforts do not substantially lower the likelihood of their data being accessed or misused – even though there may be risks around allowing others to collect information about them. In particular, in the US, where data privacy laws tend to be less stringent than those in the EU, public educational campaigns or school curricula about personal data collection by third parties and types of precautions individuals could be taking to protect themselves could be crucial. Although participants in the present study were aware that their data were easily accessible and being collected by many parties, they primarily focused on how it was being used for advertising, shopping, and credit card fraud. Understanding the specific uses of their data and the consequences they have, now or in the future, may encourage individuals to be more attentive to their data security.

Acknowledgment. The authors gratefully acknowledge research support from Liberty Mutual.

References

1. Smith, A.: Record shares of Americans now own smartphones, have home broadband (2017). http://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/
2. Dscout Mobile touches: Dscout's inaugural study on humans and their tech (2016). https://blog.dscout.com/mobile-touches
3. Husson, T., Khatibloo, F.: How will people trust you? Technical report (2015). https://www.forrester.com/report/How+Will+People+Trust+You/-/E-RES117920
4. Icontrol Networks 2015 State of the Smart Home Report. Technical report (2015). https://www.slideshare.net/iangertler/2015-state-of-the
5. Barcena, M.B., Wueest, C.: Security response: Insecurity in the internet of things (2015). https://www.symantec.com/content/dam/symantec/docs/white-papers/insecurity-in-the-internet-of-things-en.pdf
6. Marketwatch: Proofpoint uncovers internet of things (IOT) cyberattack. Marketwatch (2014). https://www.marketwatch.com/story/proofpoint-uncovers-internet-of-things-iot-cyberattack-2014-01-16?reflink=MW_news_stmp
7. Olson, P.: Why Google's Waze is trading user data with local governments. Forbes (2014). https://www.forbes.com/sites/parmyolson/2014/07/07/why-google-waze-helps-local-governments-track-its-users/#184a3f9339ba
8. Lambert, F.: Telsa updates data sharing policy to include collecting video in order to 'make self-driving a reality' (2017). https://electrek.co/2017/05/06/tesla-data-sharing-policy-collecting-video-self-driving/
9. Pingo, Z., Narayan, B.: When personal data becomes open data: an exploration of lifelogging, user privacy, and implications for privacy literacy. In: Morishima, A., Rauber, A., Liew, C.L. (eds.) ICADL 2016. LNCS, vol. 10075, pp. 3–9. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49304-6_1
10. Westin, A.F.: Privacy and Freedom. Athenaeum, New York (1967)
11. Mintel: Sixty percent of millennials willing to share personal info with brands (2014). http://www.mintel.com/press-centre/social-and-lifestyle/millennials-share-personal-info
12. Loudhouse: The future of digital trust: A European study on the nature of consumer trust and personal data (2014). https://www.orange.com/en/content/download/21358/412063/version/5/file/Orange+Future+of+Digital+Trust+Report.pdf
13. Fleming, J., Adkins, A.: Data security: Not a big concern for millennials (2014). http://news.gallup.com/businessjournal/192401/data-security-not-big-concern-millennials.aspx
14. Seneviratne, S., Kolamunna, K., Seneviratne, A.: Short: a measurement study of tracking in paid mobile applications. In: Proceedings of the 8th AMC Conference on Security and Privacy in Wireless and Mobile Networks, New York, USA (2015). https://doi.org/10.1145/2766498.2766523
15. Yu, Z., Macbeth, S., Modi, K., Pujol, J.M.: Tracking the trackers. In: Proceedings of the 25th International Conference on World Wide Web (WWW 2016). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 121–132 (2016)
16. Google: Powering ads and analytics innovation with machine learning (2017). https://adwords.googleblog.com/2017/05/powering-ads-and-analytics-innovations.html
17. Berry, M.J.A., Linoff, G.S.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd edn. Wiley Publishing, Indianapolis (2004)

18. Citron, D.K., Pasquale, F.A.: The scored society: Due process for automated predictions. WASH LAW REV, vol. 89, p. 1 (2014). https://ssrn.com/abstract=2376209. U of Maryland Legal Studies Research Paper No. 2014-8
19. O'Carroll, T., Franco, J.: Why build a Muslim registry when you can buy it? Amnesty Global Insights (2017). https://medium.com/amnesty-insights/data-brokers-data-analytics-muslim-registries-human-rights-73cd5232ed19
20. Equifax: Consumer Notice: Notice of Data Breach (2017). https://www.equifaxsecurity2017.com/consumer-notice/
21. Statista: Smart Home. Technical report (2017b). https://www.statista.com/study/27165/smart-homes-statista-dossier/
22. Statista: Facebook. Technical report (2017a). https://www.statista.com/study/9711/facebook-statista-dossier/
23. Statista: Wearable Devices. Technical report (2017c). https://www.statista.com/study/15607/wearable-technology-statista-dossier/

Author Index