

Yukiko I. Nakano  
Cristina Conati  
Thomas Bader *Editors*

# Eye Gaze in Intelligent User Interfaces

Gaze-based Analyses, Models and  
Applications

 Springer

# Eye Gaze in Intelligent User Interfaces

Yukiko I. Nakano • Cristina Conati • Thomas Bader  
Editors

# Eye Gaze in Intelligent User Interfaces

Gaze-based Analyses, Models and  
Applications



Springer

*Editors*

Yukiko I. Nakano  
Department of Computer and Information  
Science  
Seikei University  
Tokyo, Japan

Thomas Bader  
Research & Development  
AGT International  
Darmstadt, Germany

Cristina Conati  
Department of Computer Science  
University of British Columbia  
Vancouver, BC, Canada

ISBN 978-1-4471-4783-1

ISBN 978-1-4471-4784-8 (eBook)

DOI 10.1007/978-1-4471-4784-8

Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013930010

© Springer-Verlag London 2013

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

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

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

Printed on acid-free paper

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

# Preface

This book is an outgrowth of the 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at the 16th International Conference on Intelligent User Interfaces (IUI 2011), which was held at Palo Alto, California, USA on February 13, 2011. The first eye-gaze workshop was held at IUI 2010 in Hong Kong, and was organized by Dr. Elisabeth André and Dr. Joyce Y. Chai. Following the first workshop, this workshop has continued to explore this important topic and covers a wider range of topics, including eye-tracking technologies, analyses of human eye-gaze behaviors, multimodal interpretation, gaze-based interactive IUIs, and presenting gaze behaviors in humanoid interfaces. Moreover, the workshop aimed at creating a network of researchers with different backgrounds, such as human sensing, intelligent user interface, multimodal processing, and communication science, who are interested in exploring how attentional information can be applied to novel intelligent user interfaces.

The research areas and questions targeted in the workshop are as follows:

- Technologies for sensing human attentional behaviors in IUI
- Interpreting attentional behaviors as communicative signals in IUI
- Gaze model for generating eye-gaze behaviors by conversational humanoids
- Analysis of human attentional behaviors
- Evaluation of gaze-based IUI

From the workshop presentation, we carefully selected papers that significantly contribute to the theme of this book and asked the authors to extend their original work presented at the workshop. In addition, we have invited two papers so as to cover a wider range of topics for attention-aware interfaces: Chap. 5 by Marc-Antoine Nussli, Patrick Jermann, Mirweis Sangin, and Pierre Dillenbourg, and Chap. 10 by Jens Edlund, Samer Al Moubayed, and Jonas Beskow.

The collected papers are organized into three sections:

Part I: Gaze in Human Communication

Part II: Gaze-Based Cognitive and Communicative Status Estimation

Part III: Gaze Awareness in HCI

Part I focuses on analyzing human eye gaze behaviors to reveal the characteristics of human communication and cognition. Part II addresses the estimation and prediction of the cognitive state of the users using gaze information. Finally, Part III presents novel gaze-aware interfaces that integrate eye-trackers as a system component. This part provides information on the direction of future human-computer interaction and discusses issues to be addressed in designing gaze-aware interactive interfaces.

We would like to thank the program committee members of the IUI 2011 workshop: Elisabeth André (University of Augsburg, Germany), Nikolaus Bee (University of Augsburg, Germany), Justine Cassell (Carnegie Mellon University, USA), Joyce Chai (Michigan State University, USA), Andrew Duchowski (Clemson University, USA), Jürgen Geisler (Fraunhofer IOSB, Germany), Patrick Jerermann (Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland), Yoshinori Kuno (Saitama University, Japan), Kasia Muldner (Arizona State University, USA), Toyooki Nishida (Kyoto University, Japan), Catherine Pelachaud (TELECOM Paris Tech, France), Christopher Peters (Coventry University, UK), Shaolin Qu (Michigan State University, USA), Matthias Rötting (University of Berlin, Germany), and Candy Sidner (Worcester Polytechnic Institute, USA). These individuals donated their precious time and effort in reviewing the papers presented herein.

We also would like to thank SMI SensoMotoric Instruments GmbH for supporting the workshop and Springer London for their support and cooperation in publishing this collection.

Tokyo, Japan  
Vancouver, BC, Canada  
Darmstadt, Germany

Yukiko I. Nakano  
Cristina Conati  
Thomas Bader

# Contents

<b>1</b>	<b>Introduction</b> . . . . .	<b>1</b>
	Yukiko I. Nakano	
<b>Part I Gaze in Human Communication</b>		
<b>2</b>	<b>How Eye Gaze Feedback Changes Parent-Child Joint Attention in Shared Storybook Reading?</b> . . . . .	<b>9</b>
	Jia Guo and Gary Feng	
<b>3</b>	<b>Shared Gaze in Situated Referential Grounding: An Empirical Study</b>	<b>23</b>
	Changsong Liu, Rui Fang, and Joyce Y. Chai	
<b>4</b>	<b>Automated Analysis of Mutual Gaze in Human Conversational Pairs</b>	<b>41</b>
	Frank Broz, Hagen Lehmann, Chrystopher L. Nehaniv, and Kerstin Dautenhahn	
<b>Part II Gaze-Based Cognitive and Communicative Status Estimation</b>		
<b>5</b>	<b>REGARD: Remote Gaze-Aware Reference Detector</b> . . . . .	<b>63</b>
	Marc-Antoine Nüssli, Patrick Jermann, Mirweis Sangin, and Pierre Dillenbourg	
<b>6</b>	<b>Effectiveness of Gaze-Based Engagement Estimation in Conversational Agents</b> . . . . .	<b>85</b>
	Ryo Ishii, Ryota Ooko, Yukiko I. Nakano, and Tokoaki Nishida	
<b>7</b>	<b>A Computational Approach for Prediction of Problem-Solving Behavior Using Support Vector Machines and Eye-Tracking Data</b> . .	<b>111</b>
	Roman Bednarik, Shahram Eivazi, and Hana Vrzakova	
<b>Part III Gaze Awareness in HCI</b>		
<b>8</b>	<b>Gazing the Text for Fun and Profit</b> . . . . .	<b>137</b>
	Ralf Biedert, Georg Buscher, and Andreas Dengel	

**9 Natural Gaze Behavior as Input Modality for Human-Computer Interaction** . . . . . 161  
Thomas Bader and Jürgen Beyerer

**10 Co-present or Not?** . . . . . 185  
Jens Edlund, Samer Al Moubayed, and Jonas Beskow

**Index** . . . . . 205



# Contributors

**Samer Al Moubayed** KTH Speech, Music and Hearing, Stockholm, Sweden

**Thomas Bader** AGT Group (R&D) GmbH, Darmstadt, Germany

**Roman Bednarik** University of Eastern Finland, Joensuu, Finland

**Jonas Beskow** KTH Speech, Music and Hearing, Stockholm, Sweden

**Jürgen Beyerer** Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany

**Ralf Biedert** German Research Center for Artificial Intelligence, Kaiserslautern, Germany

**Frank Broz** Computer Science, University of Hertfordshire, Hatfield, UK

**Georg Buscher** Microsoft Bing, Redmond, WA, USA

**Joyce Y. Chai** Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

**Kerstin Dautenhahn** Computer Science, University of Hertfordshire, Hatfield, UK

**Andreas Dengel** German Research Center for Artificial Intelligence, Kaiserslautern, Germany

**Pierre Dillenbourg** CRAFT, EPFL, Lausanne, Switzerland

**Jens Edlund** KTH Speech, Music and Hearing, Stockholm, Sweden

**Shahram Eivazi** University of Eastern Finland, Joensuu, Finland

**Rui Fang** Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

**Gary Feng** Educational Testing Service, Princeton, NJ, USA

**Jia Guo** Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

**Ryo Ishii** Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan; Department of Computer and Information Science, Seikei University, Tokyo, Japan; NTT Communication Science Laboratories, NTT Corporation, Kanagawa, Japan

**Patrick Jermann** CRAFT, EPFL, Lausanne, Switzerland

**Hagen Lehmann** Computer Science, University of Hertfordshire, Hatfield, UK

**Changsong Liu** Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

**Yukiko I. Nakano** Department of Computer and Information Science, Seikei University, Tokyo, Japan

**Chrystopher L. Nehaniv** Computer Science, University of Hertfordshire, Hatfield, UK

**Tokoaki Nishida** Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan

**Marc-Antoine Nüssli** CRAFT, EPFL, Lausanne, Switzerland

**Ryota Ooko** Department of Computer and Information Science, Seikei University, Tokyo, Japan

**Mirweis Sangin** Sony, London, UK

**Hana Vrzakova** University of Eastern Finland, Joensuu, Finland

# Chapter 1

## Introduction

Yukiko I. Nakano

**Abstract** This chapter first discusses various areas of eye gaze research in order to show that gaze information plays an important role in various human cognitive activities. Based on the discussion, a three-step approach towards effective and natural gaze-aware human-computer interaction is proposed, which involves the analysis of human attentional behaviors, establishing computational models for interpreting eye gaze information, and building gaze-aware user interfaces. Finally, an overview of each chapter is provided to discuss how each of them contributes to one of these aspects.

### 1.1 Eye Gaze in Various Human Activities

Humans are looking at some sort of objects in order to obtain visual information, which is used to recognize objects and events to comprehend their situation. On the other hand, by observing the attentional behaviors of another person, we can estimate the object of concentration and interest of the person. Moreover, we may infer her/his internal cognitive state based on their eye gaze. Suppose that a person is operating a machine and the next step is pushing button A. If the person is looking at button B, we may infer that the person will push the wrong button and so provide a warning “No, that’s not it!”

Even though eye gaze is simple and subtle compared to gesture and speech, eye gaze provides rich information on human activities. Intriguingly, in many different languages, there are similar phrases suggesting the eloquence of eye gaze: “*One can say more with a look than with ten thousand words,*” and “*The eyes are the windows to the soul.*”

In face-to-face communication, conversational participants observe each other’s eye gaze, and this serves as a nonverbal communication signal (Kendon 1967). If the listener is looking at the speaker, this is a sign that the listener is engaged in the conversation. If the speaker looks at the listener and the listener also looks at the

---

Y.I. Nakano (✉)

Department of Computer and Information Science, Seikei University, Tokyo, Japan  
e-mail: [y.nakano@st.seikei.ac.jp](mailto:y.nakano@st.seikei.ac.jp)

speaker, they establish mutual gaze or eye contact, which contributes to maintaining the communication (Clark 1996). If the conversational participants talk about a shared object, then joint attention is indispensable to the process of identifying and using the referent as shared knowledge (Whittaker 2003). Eye gaze also plays an important role in floor management (Duncan 1974). When releasing a turn, the current speaker looks at the next speaker at the end of the turn. Then, the next speaker averts his/her gaze at the beginning of his/her turn and starts speaking. Therefore, gaze information is key in understanding face-to-face communication.

Not only in communication, but also in working on a task, we can see how well the person is performing the task by observing his/her attentional behaviors. Moreover, we can also infer the person's cognitive states by analyzing gaze. One of the most typical examples is reading activity. The gaze target indicates which word in a line the user is reading, and we can monitor the reading activity even if the user does not read aloud. In the object manipulation task, gaze information indicates on which object the user is concentrating, and this information is more precise and reliable than think-aloud reporting (Ericsson and Simon 1993). Even in a collaborative group work by multiple users, gaze information enables to estimate what is shared between the users (Brennan et al. 2008). Thus, gaze provides valuable information to understand task performing activities and to infer the cognitive states that are not expressed verbally.

## 1.2 Approach Towards Gaze-Aware Human-Computer Interaction

As discussed above, gaze information is indispensable in order to get better understanding of various human activities. However, attentional behaviors are very subtle compared to other expressions, such as gestures and speech, and are performed unconsciously in most cases. Therefore, if a system can recognize and interpret such delicate signals from the user, then the system can facilitate a completely new style of human-computer interaction where the system senses the user's cognitive states and responds to the user without any explicit command from the user. Due to the development of eye tracking technology, research on such gaze-aware user interfaces has just begun and has attracted a lot of attention from researchers in various fields.

Research on gaze-aware user interfaces consists of at least the following three phases, each of which has different issues that must be addressed.

### (1) *Analysis of human attentional behaviors*

As the first step towards gaze-aware user interfaces, the characteristics of human eye gaze behavior need to be revealed. In communication studies, researchers analyzed eye gaze by manually observing videos, now however eye trackers can automatically measure gaze behaviors. Using an eye tracker, gaze data can be automatically collected to an accuracy of milliseconds with an error of less than one degree. This enables far more precise analysis of gaze.

Moreover, gaze is deeply related to other modalities, such as speech, gesture, and object manipulation. There are many issues concerning multimodality, such as how gaze and speech correlate in referent identification, how gesture is used in coordination with gaze, and how gaze and object manipulations are correlated.

### *(2) Computational models for interpreting eye gaze information*

In order to implement gaze-awareness functionality, computational models for interpreting gaze information are necessary. A large amount of precise gaze data measured within 10 to 30 ms are applied to statistical models and machine learning techniques. Probabilistic models are also useful in combining information from other modalities. Note that the most important point is that the models and methods should be established based on the analysis of human attentional behaviors, which provide a reasonable basis for the models. After establishing a model, the accuracy and effectiveness of the model need to be evaluated.

### *(3) Building gaze-aware user interfaces*

At the final step, computational models are implemented and exploited in human-computer interaction. There are several ways of exploiting gaze-awareness, such as simply using as an input device, changing the system's behaviors according to the user's interest and attitude estimated from gaze information, and interpreting the user's intention in multimodal understanding. These functionalities can be used in various types of applications, such as multimodal input interfaces, educational systems, cooperative work environments, and human-robot communication. In developing such gaze-aware user interfaces, design and implementation issues, such as misrecognition and missing data due to the failure of measurement and the usability of calibration tools, must be considered.

## **1.3 Outline of the Book**

Based on the above discussion, this book consists of three parts, each of which addresses issues for each of the research phases discussed in the previous section. The overviews of each chapter are introduced below.

### *Part I: Gaze in Human Communication*

Part I presents research that focuses on analyzing human eye gaze behaviors in order to reveal the characteristics of human communication and cognition.

Guo and Feng (Chap. 2) investigated how gaze information affects story book reading between a parent and a pre-reading child. Children who received eye gaze feedback from the parent learned more keywords in the book. They also investigated whether informing the parents of children's real-time visual attention helps the parents regulate joint attention in shared book reading. They revealed that by showing where the parent is looking and where the child is looking, the number of occurrences of joint attention to print words increases, and in such environments, children learned more words.

Liu et al. (Chap. 3) conducted an experiment to solve a naming game. In their experiment, they found that the subjects had to spend more time and effort when the views were mismatched with the partner. In addition, the subjects' collaboration became significantly more efficient when the matcher was aware of the director's eye gaze during the interaction. These results suggest that the awareness of the partner's gaze is more helpful in mismatched views than in a matched view. The analysis in this chapter provides a good basis for modeling situated interaction between human users and vision-based interactive systems.

Broz et al. (Chap. 4) reported an automatic dyadic gaze analysis using two eye trackers. They implemented a mechanism that integrates gaze data from two eye trackers and automatically detects mutual gaze. They used this mechanism in collecting and analyzing mutual gaze in dyadic communication and characterized the communication using mathematical techniques. Although video analysis of gaze behaviors requires enormous time and effort, this chapter shows that eye-trackers are useful in reducing such costs.

### *Part II: Gaze-Based Cognitive and Communicative Status Estimation*

The research in Part II addresses the estimation and prediction of the internal state of the users using gaze information.

Nussli et al. (Chap. 5) proposed the Remote Gaze-Aware Reference Detector (REGARD) system. They measured the time lag between gaze and speech. There is a time lag between the time that the speaker looks at an object and the time that the reference word is uttered. There is another time lag between the time that the listener hears the word and the time that he/she looks at the referent. They applied the empirical results to a mechanism that automatically recognizes referential expressions in speech stream and identified the referred objects. The system provides multimodal understanding by combining speech and gaze in dialogues in which two individuals are working or discussing in a shared workspace.

Ishii et al. (Chap. 6) analyze various types of data obtained from an eye-tracker and propose a decision tree model for estimating conversational engagement. They also build a conversational agent that is aware of the user's conversational engagement. In their system evaluation, they show that the agent's simple feedback, which expresses only the awareness of engagement, has a significant impact on the subjects' verbal and nonverbal behaviors during the interaction with the agent, as well as on the subjective impression of the agent.

Bednarik et al. (Chap. 7) used a computational approach for predicting problem-solving behavior. They defined categories for cognitive status, such as cognition, evaluation, planning, and intention, based on the subjects' think-aloud utterances and then proposed a classifier for these cognition categories by applying a support vector machine (SVM) to eye-tracking data. They also proposed a model for predicting the performance level of a user in problem solving. Their model shows that the higher-order cognitive traits can be predicted from lower-order eye-tracking data and suggests the possibility of monitoring the user's cognitive status in real time.

### *Part III: Gaze Awareness in HCI*

Part III presents studies that propose novel gaze-based interfaces that integrate eye-trackers as a system component. These studies show the direction of future human-computer interaction and discuss issues to be addressed in designing gaze-aware interfaces.

Biedert et al. (Chap. 8) reported an evaluation experiment for their proposed eBook, namely, eyeBook, which provides background music, sound effects, and images at the moment that the user is looking at some specific sentences. They then extended their idea to a multimodal eBook reader, which integrates speech recognition, hand-writing recognition, and emotion recognition using EEG. These multimodal cues trigger specific scripts to provide feedback to the reader. This chapter shows how such gaze-aware user interfaces enable a richer reading experience.

Bader and Beyerer (Chap. 9) revealed different factors that influence natural gaze behaviors in object manipulation tasks. They focus on proactive attentional behavior, which is used to estimate the user's intention in human-computer interaction. They investigated whether such natural gaze behavior is useful as an additional input modality combined with gestures in a multi-display environment. They reported that gaze-based intention estimation is valuable in compensating for inaccurate hand gestures and results in less physical fatigue.

Edlund et al. (Chap. 10) discussed the production of eye-gaze expressions in humanoid interfaces. Not only recognizing human attentional behavior, but also studying how people perceive eye gaze expressions displayed by embodied interfaces is another important aspect in designing human-computer interaction. This chapter focuses on Mona Lisa gaze effect in viewing 2D images of gaze, and proposes projecting animated faces on head-shaped 3D surfaces, by which people can more reliably estimate gaze targets in the physical world.

## References

- Brennan SE et al (2008) Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106:1465–1477
- Clark HH (1996) *Using language*. Cambridge University Press, Cambridge
- Duncan S (1974) On the structure of speaker-auditor interaction during speaking turns. *Lang Soc* 3:161–180
- Ericsson KA, Simon HA (1993) *Protocol analysis: verbal reports as data* revised edition. MIT Press, Cambridge
- Kendon A (1967) Some functions of gaze direction in social interaction. *Acta Psychol* 26:22–63
- Whittaker S (2003) Theories and methods in mediated communication. In: Graesser A, Gernsbacher M, Goldman S (eds) *The handbook of discourse processes*. Erlbaum, Hillsdale, pp 243–286

**Part I**  
**Gaze in Human Communication**



# Chapter 2

## How Eye Gaze Feedback Changes Parent-Child Joint Attention in Shared Storybook Reading?

### An Eye-Tracking Intervention Study

Jia Guo and Gary Feng

**Abstract** Joint attention is critical for effective communication and learning during shared reading. There is a potential disassociation of attention when the adult reads texts while the child looks at pictures. We hypothesize the lack of joint attention limits children’s opportunity to learn print words. Traditional research paradigm does not measure joint attention in real-time. In the current study, three experiments were conducted to monitor parent-child joint attention in shared storybook reading. We simultaneously tracked eye movements of a parent and his/her child with two eye-trackers. We also provided real-time eye gaze feedback to the parent about where the child was looking at, and vice versa. Changes of dyads’ reading behaviors before and after the intervention were measured from both eye movements and video records. Baseline data showed little joint attention in the naturalistic parent-child shared reading. The real-time eye gaze feedback significantly increased parent-child joint attention and improved children’s learning.

### 2.1 Introduction

Joint attention, the capacity to coordinate attention with a social partner on a particular action or object, is essential for communication, visual search, problem solving, and many other collaborative activities (Brennan et al. 2008; Carletta et al. 2010; Nüssli et al. 2009; Richardson et al. 2007). Parent-child shared storybook reading is one of such activities. There is converging evidence that a key to learn print words is to engage children in a joint attention on texts during reading, i.e., children and adults must attend simultaneously to the target of learning and among themselves

---

J. Guo (✉)  
Department of Psychology and Neuroscience, Duke University, Durham, NC, USA  
e-mail: [jiaguo.elaine@gmail.com](mailto:jiaguo.elaine@gmail.com)

G. Feng  
Educational Testing Service, Princeton, NJ, USA  
e-mail: [gary.feng@gmail.com](mailto:gary.feng@gmail.com)

(Ezell and Justice 2000; Gong and Levy 2009; Justice et al. 2006, 2008). However, prior eye-tracking studies showed that pre-reading children focus almost exclusively on pictures while parents read from print texts (Evans and Saint-Aubin 2005; Feng and Guo 2012; Justice et al. 2005, 2008). Additionally, most existing joint attention regulation strategies such as finger pointing are adult-centered and limited, such that a parent regulates a child's attention without accurate information about the child's real-time attention state. We argue that the lack of joint attention and the limitation of traditional joint attention regulations impede children's learning of print words. And we conjecture that the ideal solution is to provide reading partners consistent, individualized, and real-time attention feedback.

The state of eye-tracking technologies allows us to show this feedback in real-time. We can show parents or children, a cursor on the computer screen that corresponds to the gaze location of the other person. The eye gaze feedback provides critical information that is missing in the traditional shared reading task.

First, the location of the eye gaze indicates the focus of attention at any given moment (Rayner 1998; Rayner et al. 2006). We expect that discovering children's real-time attention state may trigger adults' regulations of joint attention during shared book reading. The real-time eye gaze information is more instructional to the pre-reading children, who will see where and how grown-ups look when they read.

Second, having access to the other partner's eye movements may change the dynamics of shared reading as well as greatly reduce the time and energy that partners spend on reengaging joint attention. The success or failure of a pedagogical attempt is immediately seen on the screen. Adults can give children more prompt and precise feedback when they watch children's real-time eye movements.

Utilizing the advanced eye-tracking technique to measure and facilitate joint attention provides a new model of understanding the joint attentional interactions in shared book reading. As such, two aims are addressed in the current study. First, we seek to objectively measure the joint attention in shared storybook reading, by simultaneously tracking the eye gazes of both the parent and the child. While there are a handful of published studies looking at children's eye gaze during reading, none has investigated the correlation and contingency between eye movements of children and parents. Using two eye trackers, we tracked dyads' eye movements and measured their joint attention in real-time during shared book reading. The data and methodology will be useful to a wide array of researchers interested in joint attention and collaborative behaviors.

Second, we investigate whether two eye gaze based interventions enhance parent-child joint attention during reading. The interventions target the fact that partners in shared reading do not know where the other person is attending to. One intervention involves showing a moving cursor on the child's monitor that indicates where the parent is looking. The other intervention shows the parent where the child is looking. With this critical piece of information, it is hypothesized that the dyad can better regulate their joint attention, which will facilitate children's learning of print words.



A. The dyad has joint attention on texts.

B. The dyad does not have joint attention.



Note:  represents the child's real-time eye gaze,  represents the parent's real-time eye gaze.

Fig. 2.1 Examples of parent-child joint attention on one book page during the shared reading

## 2.2 General Method

### 2.2.1 Design and Analysis

Three experiments were conducted in the current study. Experiment 1 is the baseline in which we measured how much joint attention on print exists during the naturalistic shared book reading. We hypothesize that children rarely look at texts and therefore there is little joint attention on print between children and parents. This in turn implies limited print word learning during the reading sessions, as measured by children's gains in the sight word recognition. Experiment 1 serves as the control condition for Experiments 2 and 3 in both of which we investigated whether the real-time feedback of eye gaze enhances parent-child joint attention and children's word learning. We presented children how their parents read texts in Experiment 2 and showed parents their children's real-time eye movements in Experiment 3. We hypothesize that the new paradigm will help dyads regulate joint attention and help children learn print words.

Joint attention was defined in the study as when the partners look simultaneously at (or near) the same visual object on a page (see the examples in Fig. 2.1).

The distance between the screen coordinates of a parent's eye gaze and those of a child's eye gaze was calculated at every 20 milliseconds for all reading sessions. To measure the real-time joint attention, we compared the average distance of two partners' eye gaze locations with a cut-off value of 201.18 pixels which was determined for three reasons. First, when reading partners were asked to look at the same object on the screen in a pilot test, 80 percent of the calculated distance values were within 201.18 pixels. This result suggested when parents and children had joint attention on the screen, most of their eye gaze distance values were less than 201.18 pixels. Second, the visual angle which corresponds to the 201.18 pixels is about 10 degrees (Eyelink systems typically have 20 pixels/degree). The human fovea, where

we have clear vision, is about 2 degrees. So the visual angle of 10 degrees is not a too small window size for a definition of joint attention. Third, the 201.18 pixels are close to the size of two 5-letter-long print words in pixels (the average length of a 5-letter-long word is 100 pixels). Therefore, we believe this is a very reasonable window to define joint attention in reading.

We determined the joint attention exists if the distance is smaller than 201.18 pixels and does not exist if the distance is larger than or equal to 201.18 pixels. The percentage of time when the distance of two partners' eye gaze locations is smaller than 201.18 pixels represents how much joint attention the dyad has when reading together.

Video recordings of the parent-child reading interactions were transcribed and coded with InqScribe software. Adapting from the coding systems in previous studies (Chi et al. 2001; Ortiz et al. 2001; Whitehurst et al. 1988; Sulzby 1985), we have developed a coding system to analyze parent-child joint attention interactions. The inter-rater reliability analysis of the behavior coding was performed using 20 % of the sample.

We hypothesized that (a) there is limited parent-child joint attention to texts in the naturalistic shared storybook reading, and (b) the eye gaze feedback facilitates joint attentional regulation and improves the learning of print words.

### ***2.2.2 Participants and Materials***

Altogether we recruited ninety-two parent-child dyads for this study. Thirty-seven dyads participated in Experiment 1; they also serve as the comparison group for the subsequent intervention experiments. Experiment 2 involved twenty-seven parent-child dyads. Experiment 3 involved twenty-eight dyads. All children participants were 4–5 year old English speakers who had no history of hearing, vision, or cognitive impairments. Parent participants were required to be person who reads most frequently with children at home. Three age appropriate storybooks were presented for dyads to read in all three experiments. Children's sight word learning was measured before and after reading by asking children to name content words sampled from the storybooks.

### ***2.2.3 Apparatus***

Two contact-free eye trackers, a Tobii X50 and an Eyelink 1000 system, were used in the study. Both of the eye trackers are infrared-based remote eye tracking systems that make no contact with the participant. For each dyad, the parent was eye tracked by Tobii X50 and the child was eye tracked by Eyelink 1000. Two video recorders were used to record reading interactions among the dyad.

For each dyad, the parent and the child sat across a child-sized table at a 90 degree angle. One LCD monitor (1280 × 1024 pixels resolution) and Eyelink 1000

**Fig. 2.2** The apparatus and experiment set-up



were put approximately 60 cm away from the child; while another LCD monitor ( $1280 \times 1024$  pixels resolution) and Tobii X50 were put approximately 60 cm away from the parent (see Fig. 2.2 for details of the set-up). Stimuli were presented simultaneously on both monitors. Stimulus presentation and eye movement calibration and recording were done using the Double Tracker program developed in our lab. The data were then exported offline for statistical analyses.

### **2.3 Experiment 1: Joint Attention in the Naturalistic Shared Book Reading**

Experiment 1 serves as the baseline condition for the subsequent intervention experiments. For each reading dyad, both the parent and the child were eye tracked and video-taped in a naturalistic shared reading task. Before and after the reading trials, we measured children's sight word recognition to determine they have learned new words.

Specifically, each parent read three books to his/her child in four reading trials (order counter-balanced among participants). They read one storybook in the first and fourth trial, and the other two different storybooks in the second and third trial. In the fourth trial the parent was asked to teach three words that the child did not recognize in the pre-test. To tease apart the impact of the instructions from the moving cursor in Experiment 2, children were asked to follow the parent's eye gaze while listening to stories, even though they could not actually see the eye gaze on the screen. The average percentage of time children had joint attention with adults on texts in reading trial 1 (no word teaching) was compared to that in trial 4 (with word teaching). Children's sight word recognition was measured before and after the reading.

The results in Experiment 1 supported our hypothesis that parent-child dyads did not have much joint attention on print and children did not learn many keywords from the pre-test to the post-test. Specifically, in trial 1 when word teaching was not

required, the average percentage of time children had joint attention with parents on print was 2.91 %. In the reading trial 4, when adults were asked to teach children three words in the books, children significantly increased their joint attention on print to 6.41 %,  $t(36) = 2.48$ ,  $p = .018$ . Children's average pre-test raw score of the sight word recognition were 1.81 (out of 10 words), which was lower than the post-test raw score of 2.19 (out of 10 words),  $t(36) = 3.19$ ,  $p = .003$ . Children's average number of learned new words was .38 words, measured by the difference of the pre- and post-test raw scores of the sight word recognition.

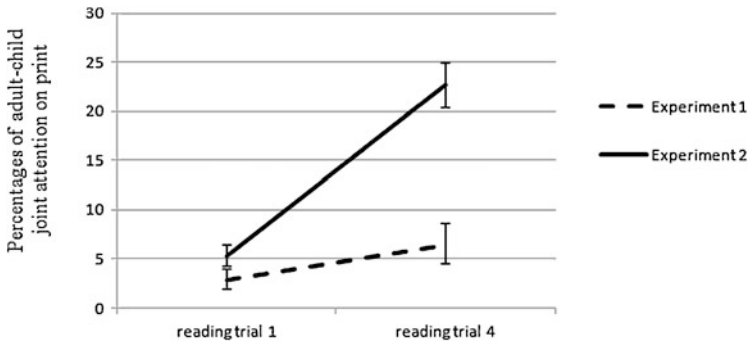
Taken together, children's joint attention with parents on print in the naturalistic shared book reading trial were small. Children's average word learning gain from the pre-test to the post-test was limited. These results were consistent with previous research findings (Evans and Saint-Aubin 2005; Evans et al. 2009; Feng and Guo 2012; Justice et al. 2005, 2008). When children read books with adults, they usually prefer pictures and avoid looking at texts. Since adults read from texts most of the time, children's ignorance of texts would lead to little adult-child joint attention on print during the naturalistic shared book reading. Therefore, children can hardly improve their sight word learning from the shared book reading activity.

The naturalistic nature of the data set and the analyses we have done in Experiment 1 point to the need for the experimental evaluation of the relationship between children's joint attention on texts and their word learning. A natural follow-up experiment would be to keep everything else (e.g., the reading materials, study set-up and procedures) the same, and use our newly developed eye tracking technology to let children see parents' real-time reading eye movements during shared book reading.

## **2.4 Experiment 2: Externalizing Adults' Visual Attention for Children in Shared Book Reading**

Our pilot studies showed that most preschool children are unaware of the facts that we take for granted such as adults pay attention to print during shared book reading and they read texts from left to right. In Experiment 2, we investigate whether the externalization of adults' reading processes helps children understand how adults read books as well as how this understanding helps children switch their own attention focus to print. Furthermore, we examine whether the increased joint attention on print promotes children's word learning.

Experiment 2 involved showing a moving cursor on the children's screen from reading trial 2 to 4; the moving cursor indicated the location of the parent's eye gaze in real-time. Parents looked at a normal, static display of the page for all four reading trials. We ensured children and parents understood the gaze indicator using an iSpy-like game. Even the youngest children had no problem understanding the correspondence. Children's sight word recognition was measured before and after the reading. In the fourth trial the parent was asked to teach three words that the child did not recognize in the pre-test. The average percentage of parent-child joint



**Fig. 2.3** Percentages of parent-child joint attention on print from reading trial 1 (no intervention) to trial 4 (with intervention) between Experiment 1 and 2

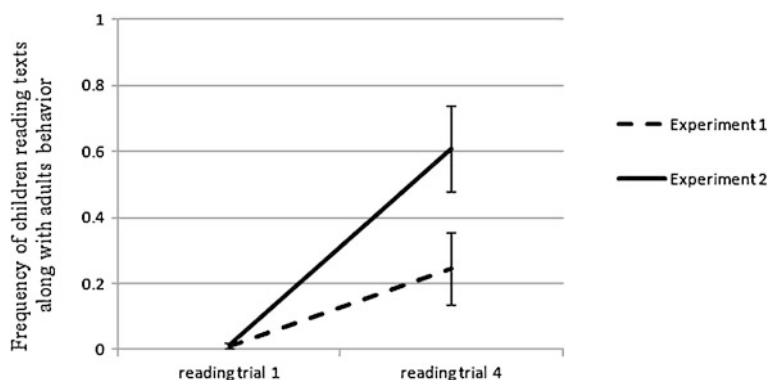
attention on texts and children's sight word learning outcomes in Experiment 2 were compared with those in Experiment 1.

The eye-tracking results showed that the intervention of eye gaze feedback for children significantly increased parent-child joint attention on texts. In trial 1 when no intervention was involved, the percentage of time children had joint attention with adults on texts was 5.35 %; in trial 4 when children read the same book while watching their parents' real-time eye gaze, children significantly increased the percentage of time they had joint attention on texts to 22.7 %,  $t(26) = -8.01$ ,  $p < .001$ .

To further compare the eye movement changes from the first to the fourth reading trial between Experiment 1 and 2, we did a repeated measures ANOVA using children's average percentage of joint attention on texts as the dependent variable, the reading trial as the within-subjects independent variable (the first vs. the fourth reading trial), and whether children received the eye gaze feedback as the between-subjects factor (Experiment 1 vs. Experiment 2). The results showed that the main effect of the within-subjects variable was significant,  $F(1, 62) = 70.8$ ,  $p < .001$ , suggesting that reading dyads on average significantly increased their joint attention on texts from the first to the fourth reading trial. The main effect of whether children received the eye gaze feedback was also significant,  $F(1, 62) = 23.51$ ,  $p < .001$ . So was the interaction effect,  $F(1, 62) = 31.24$ ,  $p < .001$ , indicating that the increase of percentage of joint attention on texts from the first to the fourth reading trial in Experiment 2 was significantly higher than that in Experiment 1 (see Fig. 2.3).

More print-directed joint attention resulted in more word learning: children learned on average 1.0 word, significantly higher than the 0.38 words children in Experiment 1 learned,  $t(62) = 2.37$ ,  $p = .02$ . This result indicates that children who received the eye gaze feedback learned more words from the pre-test to the post-test than children who did not receive this eye gaze information.

To examine how the eye gaze intervention changes parent-child reading interactions, we included thirty-six dyads in Experiment 1 (data from one dyad were excluded due to poor video quality) and twenty-seven dyads in Experiment 2 in the video coding and analysis. The inter-rater reliability was 0.79 ( $p < .01$ ), which



**Fig. 2.4** The frequency of the behavior of children’s reading texts along with parents from reading trial 1 (no intervention) to trial 4 (with intervention) between Experiment 1 and 2

could be claimed as good levels of agreement according to previous research (Landis and Koch 1977; Ortiz et al. 2001). We did repeated measures ANOVAs using the average frequency of each coded behavior (time per minute) as the dependent variable, the reading trial as the within-subjects independent variable (no intervention trial 1 vs. with intervention trial 4), and whether children received the eye gaze feedback as the between-subjects factor (Experiment 1 vs. Experiment 2). The results revealed children did respond more to parents’ reading strategies when seeing parents’ real-time eye scanning patterns. For example, children in both experiments increased the occurrences of the behavior of reading texts along with parents from the first to fourth reading trial ( $F_{within}(1, 61) = 23.09, p < .001$ ), but children who received the eye gaze feedback showed a significantly larger increase ( $F_{between}(1, 61) = 4.34, p = .041$ ;  $F_{interaction}(1, 61) = 4.31, p = .042$ , see Fig. 2.4).

Overall, the comparisons between Experiment 1 and 2 indicated that with the eye gaze direction more tightly tied to the focus of joint attention, children saw an external representation of reading processes unfolding in real-time. When children heard adults read texts while simultaneously looking at the corresponding words, they had the best opportunity to learn the correspondence between the sound, spelling, and meaning of the words. Furthermore, the real-time eye gaze feedback helped adults efficiently draw children’s attention to those target words and children also responded more to parents’ word teaching attempts. Children’s increased responses to parents’ pedagogical efforts, as well as the increased joint attention on print and the improved understanding of reading processes further promoted children’s print learning.

One limitation for Experiment 2 is that only children had the opportunities of knowing where adults look at on books but adults were still blind to children’s attention states. As a trade-off to this limitation, in Experiment 3 we show parents where children pay attention to during shared reading.



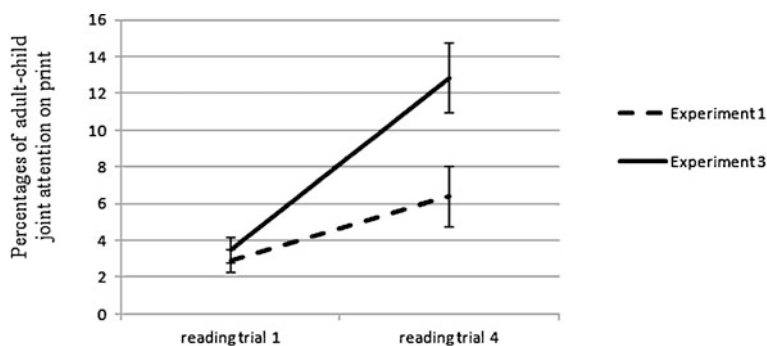
## 2.5 Experiment 3: Informing Parents of Children's Real-Time Visual Attention in Shared Book Reading

Parents control the reading activity in the traditional shared reading paradigm, but they have little knowledge about where their children are paying attention to and whether their reading strategies are effective. We argue that informing parents of children's real-time visual attention can help parents regulate joint attention in shared book reading.

In Experiment 3, we presented each parent a moving cursor on the screen from reading trial 2 to 4. The moving cursor indicated children's eye gaze location as children read. For all four reading trials children looked at a normal, static display of the page. Parents were encouraged to utilize the eye gaze information to regulate children's attention. To tease apart the impact of the instructions from the moving cursor in Experiment 2, children were asked to follow the parent's eye gaze while listening to stories, even though they could not actually see the eye gaze in Experiment 3. Dyads' eye movements and reading interactions during reading sessions were recorded and analyzed. Children's sight word recognition was measured before and after the reading. In the fourth trial the parent was asked to teach three words that the child did not recognize in the pre-test. The average percentage of parent-child joint attention on texts and children's sight word learning outcomes in Experiment 3 were compared with those in Experiment 1. We predict that seeing children's real-time visual attention makes parents adjust their reading strategies accordingly. The parents' changed reading behaviors would enhance the efficiency of joint attention regulation and therefore significantly increase the time children spend scanning texts in reading. Children can learn more words due to the increased print exposure.

Positive intervention effects were also found in Experiment 3. The percentage of joint attention on texts increased from 3.48 % in reading trial 1 (no intervention trial) to 12.87 % in reading trial 4 (with intervention trial),  $t(27) = -5.05$ ,  $p < .001$ . To further compare the eye movement changes from the first to the fourth reading trial between Experiment 1 and 3, we did a repeated measures ANOVA using children's average percentage of joint attention on texts as the dependent variable, the reading trial as the within-subjects independent variable (the first vs. the fourth reading trial), and whether parents received the eye gaze feedback as the between-subjects factor (Experiment 1 vs. Experiment 3). The results showed that the main effect of the within-subjects variable was significant,  $F(1, 63) = 31.72$ ,  $p < .001$ , suggesting that reading dyads on average significantly increased their joint attention on texts from the first to the fourth reading trial. The main effect of the eye gaze feedback was also significant,  $F(1, 63) = 5.35$ ,  $p = .024$ . So was the interaction effect,  $F(1, 63) = 6.62$ ,  $p = .012$ . The results indicated that the increase of percentage of joint attention on texts from the first to the fourth reading trial in Experiment 3 was significantly higher than that in Experiment 1 (see Fig. 2.5).

Parents became more effective facilitating children's word learning. Children learned 1.25 words, significantly higher than the word learning gain in Experiment 1 (0.38 words),  $t(63) = 3.83$ ,  $p < .001$ . This result further confirmed that children



**Fig. 2.5** Percentages of parent-child joint attention on print from reading trial 1 (no intervention) to trial 4 (with intervention) between Experiment 1 and 3

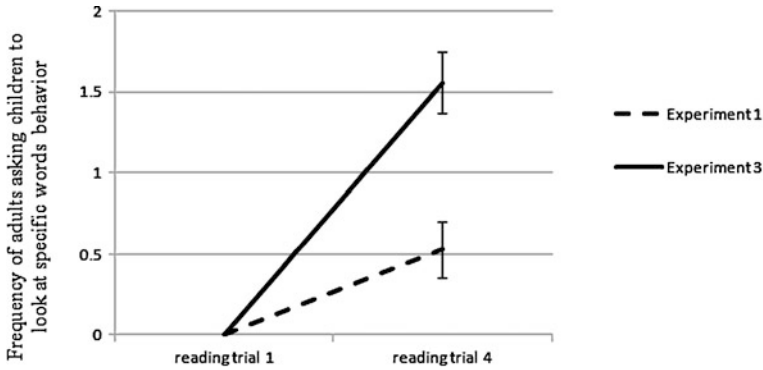
learned more words from the pre-test to the post-test when parents used the eye gaze feedback to effectively direct children’s attention to words.

The behavioral coding and analysis also supported our hypotheses. Thirty-six dyads in Experiment 1 (data from one dyad were excluded due to poor video quality) and twenty-eight dyads in Experiment 3 were included in the comparisons for the changes of the frequencies of the behaviors from reading trial 1 (no intervention) to reading trial 4 (with intervention) between Experiment 1 and Experiment 3. The inter-rater reliability was 0.78 ( $p < .01$ ). We did repeated measures ANOVAs using the average frequency of each coded behavior (time per minute) as the dependent variable, the reading trial as the within-subjects independent variable (no intervention trial 1 vs. with intervention trial 4), and whether parents received the eye gaze feedback as the between-subjects factor (Experiment 1 vs. Experiment 3).

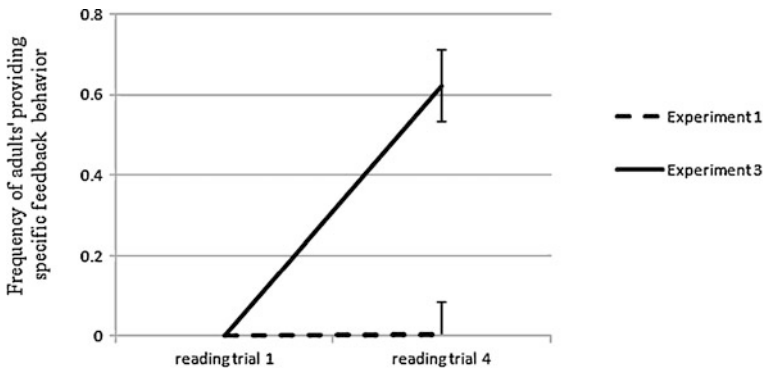
The results revealed when informed with children’s real-time eye gaze locations, parents increased the frequencies of regulating joint attention and teaching words. For example, parents in both experiments increased the occurrences of behavior of asking children to look at specific words (e.g., “Can you help me find the word ‘cat’ on the screen?”) from the first to fourth reading trial ( $F_{within}(1, 62) = 62.95$ ,  $p < .001$ ), but parents who received the eye gaze feedback showed a significantly larger increase ( $F_{between}(1, 62) = 15.32$ ,  $p < .001$ ;  $F_{interaction}(1, 62) = 15.32$ ,  $p < .001$ , see Fig. 2.6).

Compared to parents in Experiment 1, parents in Experiment 3 significantly increased the occurrences of the behavior of providing specific feedback (e.g., adults saying “Yes, you are looking at the right place.”; “No, you are not looking at the place I want you to look.”) from the first to fourth reading trial ( $F_{between}(1, 62) = 26.57$ ,  $p < .001$ ;  $F_{interaction}(1, 62) = 26.57$ ,  $p < .001$ ;  $F_{within}(1, 62) = 27.63$ ,  $p < .001$ , see Fig. 2.7).

Parents’ behavior changes induced children’s more frequent verbal responses. Although children in both experiments increased the occurrences of the behavior



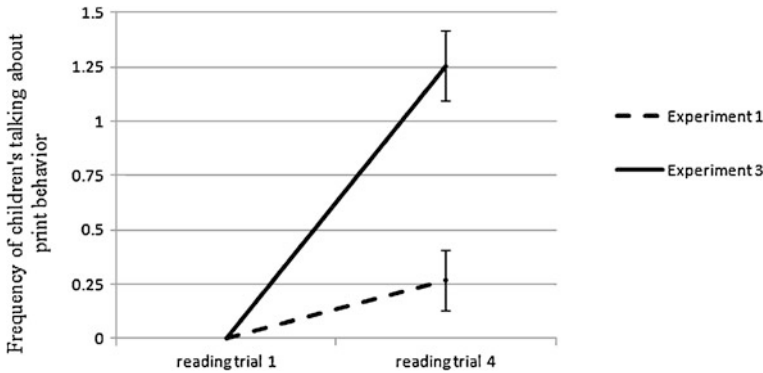
**Fig. 2.6** The frequency of adults asking children to look at specific words behavior from reading trial 1 (no intervention) to trial 4 (with intervention) between Experiment 1 and 3



**Fig. 2.7** The frequency of adults providing specific feedback behavior from reading trial 1 (no intervention) to trial 4 (with intervention) between Experiment 1 and 3

of asking or answering print-related questions from the first to fourth reading trial ( $F_{within}(1, 62) = 54.59, p < .001$ ), children in Experiment 3 showed a significantly larger increase ( $F_{between}(1, 62) = 22.71, p < .001$ ;  $F_{interaction}(1, 62) = 22.64, p < .001$ , see Fig. 2.8).

The above comparisons between Experiment 1 and Experiment 3 suggested when parents received the real-time eye gaze feedback, parents had better opportunities to observe their child’s attention state and fine-tune their strategies to increase child interest and participation. Children who experienced positive direction, coaching, and correction more easily attended to and internalized the knowledge parents attempted to teach them, and developed the interest and motivation to sustain their learning. These changes in turn provided more teachable moments for parents.



**Fig. 2.8** The frequency of children's talking about print behavior from reading trial 1 (no intervention) to trial 4 (with intervention) between Experiment 1 and 3

## 2.6 Conclusion

The current study measured parent-child joint visual attention in real-time, which allows us to go beyond prior research that focuses exclusively on the child in shared reading, and study shared reading as a joint attentional interaction involving dynamic transactions between partners and real-time cognitive strategies within individuals. Building on prior research, we found that pre-reading children had limited joint attention with their parents in the naturalistic shared reading paradigm. Parents had little information about where children were attending to, and children had even less idea about how adults actually read. This resulted in a poorly regulated joint attentional interaction when it comes to learning print words.

More importantly, our eye gaze interventions successfully remedied the joint attentional structure by leveraging the eye-tracking technology in the shared reading. By providing real-time feedback of the partner's visual attention, we demonstrated significant improvements in the amount of joint attention on print and changes in parental attentional regulation strategies during reading. More interestingly, children did not simply look at the moving cursor or print words, but actually read and processed the words. This was shown by increased word learning by children, along with children's changes of concept of reading processes. These results suggest that by providing a critical piece of information—namely, where the partner is looking—we can facilitate the regulation of joint attention and improve children's learning of print words. Our intervention targets limitations in joint attention regulation in the traditional shared reading practice, but it is not specific to reading. The data and methodology of this study would also be useful to a wide array of research topics on collaborative learning activities. To the extent learning involves joint attention (e.g., in math tutoring), the eye gaze feedback can be an effective aid for learning. Additionally, the scenario of this research is quite similar to online collaborative activities where partners focus on the common content on different screens (e.g., online gaming, collaborative search, etc.). Insights into such behaviors and mental processes may help design better multimedia software products and web applications.

## References

- Brennan SE, Chen X, Dickinson CA, Neider MB, Zelinsky GJ (2008) Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106:1465–1477
- Carletta J, Hill RL, Nicol C, Taylor T, de Ruitter JP, Bard EG (2010) Eyetracking for two-person tasks with manipulation of a virtual world. *Behav Res Methods* 42:254–265
- Chi MTH, Siler SA, Jeong H, Yamauchi T, Hausmann RG (2001) Learning from human tutoring. *Cogn Sci* 25:471–533
- Evans MA, Saint-Aubin J (2005) What children are looking at during shared storybook reading—evidence from eye movement monitoring. *Psychol Sci* 16:913–920
- Evans MA, Saint-Aubin J, Landry N (2009) Letter names and alphabet book reading by senior kindergarteners: an eye movement study. *Child Dev* 80:1824–1841
- Ezell HK, Justice LM (2000) Increasing the print focus of adult-child shared book reading through observational learning. *Am J Speech-Lang Pathol* 9:36–47
- Feng G, Guo J (2012) From pictures to words: young children’s eye movements during shared storybook reading. *J Educ Psychol* (submitted)
- Gong ZY, Levy BA (2009) Four year old children’s acquisition of print knowledge during electronic storybook reading. *Read Writ* 22:889–905
- Justice LM, Skibbe L, Canning A, Lankford C (2005) Pre-schoolers, print and storybooks: an observational study using eye movement analysis. *J Res Read* 28:229–243
- Justice LM, Skibbe L, Ezell HK (2006) Using print referencing to promote written language awareness. In: Ukrainetz TA (ed) *Contextualized language intervention: scaffolding preK-12 literacy achievement*. Thinking Publications University, Greenville, pp 389–428
- Justice LM, Pullen PC, Pence K (2008) Influence of verbal and nonverbal references to print on preschoolers’ visual attention to print during storybook reading. *Dev Psychol* 44:855–866
- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
- Nüssli M-A, Jermann P, Sangin M, Dillenbourg P (2009) Collaboration and abstract representations: towards predictive models based on raw speech and eye-tracking data. Paper presented at the conference on computer support for collaborative learning
- Ortiz C, Stowe RM, Arnold DH (2001) Parental influence on child interest in shared picture book reading. *Early Child Res Q* 16:263–281
- Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychol Bull* 124:372–422
- Rayner K, Chace KH, Slattery TJ, Ashby J (2006) Eye movements as reflections of comprehension processes in reading. *Sci Stud Read* 10:241–255
- Richardson DC, Dale R, Kirkham NZ (2007) The art of conversation is coordination—common ground and the coupling of eye movements during dialogue. *Psychol Sci* 18:407–413
- Sulzby E (1985) Children’s emergent reading of favorite storybooks: a developmental study. *Read Res Q* 20:458–481
- Whitehurst GJ, Fischel JE, Lonigan CJ, Valdezmenchaca MC, Debaryshe BD, Caulfield MB (1988) Verbal interaction in families of normal and expressive-language-delayed children. *Dev Psychol* 24:690–699

# Chapter 3

## Shared Gaze in Situated Referential Grounding: An Empirical Study

Changsong Liu, Rui Fang, and Joyce Y. Chai

**Abstract** In situated dialogue, although an artificial agent and its human partner are co-present in a shared environment, they have significantly mismatched capabilities in perceiving the environment. When a shared perceptual basis is broken, referential grounding between partners becomes more challenging. Our hypothesis is that in such a situation, non-verbal modalities such as eye gaze play an important role in coordinating the referential process. To validate this hypothesis, we developed a system to simulate mismatched visual perceptions of the shared environment between human partners. Using this system, we further designed experiments to examine how partners with mismatched visual perceptual capabilities collaborate to accomplish joint tasks. Our studies have shown that, partners with mismatched perceptions make more effort to collaborate. When one partner pays attention to the other partner's naturally occurred eye gaze during interaction, referential grounding becomes more efficient. This paper describes our empirical findings and discusses their potential implications.

### 3.1 Introduction

As a new generation of robots start to emerge into our daily life, techniques that enable situated human robot dialogue have become increasingly important (Bohus and Horvitz 2009). Human robot dialogue often involves objects and their identities in the environment. One critical problem is referential grounding, where the robot needs to correctly identify intended referents from the speaker's (human's) referring expressions (Clark and Brennan 1991).

---

C. Liu (✉) · R. Fang · J.Y. Chai

Department of Computer Science and Engineering, Michigan State University, East Lansing,  
MI 48824, USA

e-mail: [cliu@cse.msu.edu](mailto:cliu@cse.msu.edu)

R. Fang

e-mail: [fangrui@cse.msu.edu](mailto:fangrui@cse.msu.edu)

J.Y. Chai

e-mail: [jchai@cse.msu.edu](mailto:jchai@cse.msu.edu)

In situated dialogue, although an artificial agent and its partner (a human or a proxy controlled by a human) are co-present in a shared environment, they have significantly mismatched capabilities in perceiving the environment. The mismatched perceptions of the shared surroundings have a massive influence on the interaction between the human and the agent. For example, if the human refers to something in the environment which the robot cannot perceive correctly, referential grounding becomes more challenging. Therefore, language alone may be inefficient and other extralinguistic information will need to be pursued. In this paper, we investigate one type of non-verbal modalities—eye gaze during speech communication.

Eye gaze serves many functions in mediating interaction (Argyle and Cook 1976; Clark 1996), managing turn taking (Novick et al. 1996) and grounding (Nakano et al. 2003). Previous psycholinguistic findings have shown that eye gaze is tightly linked with language production and comprehension (Just and Carpenter 1975; Tanenhaus et al. 1995; Meyer et al. 1998; Griffin and Bock 2000). Eye gaze has also been shown efficient for providing early disambiguating cues in referential communication (Hanna and Brennan 2007), for intention recognition during object manipulation (Bader et al. 2009), and for attention prediction (Fang et al. 2009). Specifically, in human machine dialogue, recent work has incorporated eye gaze in resolving exophoric referring expressions (Prasov and Chai 2008, 2010).

Motivated by previous work, our hypothesis is that eye gaze plays an important role in referential grounding, especially between partners with mismatched perceptions of the shared environment. More specifically, we are interested in the following questions:

1. *How difficult is it for partners with mismatched perceptions of the shared environment to collaborate?*

When a shared perceptual basis of the environment is broken, partners may not be able to communicate as they normally do. We are interested in how the mismatched perceptions may impact collaboration, dialogue, and automated language processing.

2. *To what extent does the collaboration benefit from the shared gaze between partners? Is the shared gaze more helpful for partners with mismatched perceptions?*

Our hypothesis is that partners with mismatched perceptions could benefit more from the shared gaze. This is because on the one hand verbal communication could be more difficult in this situation, and on the other hand being aware of partner's eye gaze may allow many joint actions to be done non-verbally (Brennan et al. 2008).

To validate our hypothesis and address the above questions, we designed an experiment which required a *director* and a *matcher* to collaboratively play a naming game. The director has a complete view of the shared environment. By controlling what the matcher “sees” from the environment, we are able to simulate mismatched perceptions of the shared environment between the director and the matcher. Besides, by tracking either the director's or the matcher's eye gaze and showing it to the other, we are able to study the role of shared gaze in referential grounding.

Our results indicate that collaboration between partners with mismatched perceptions of the shared environment is inherently difficult. It takes extra efforts for

partners to overcome the perceptual discrepancies and to establish a shared basis for communication. We also found that under such circumstance human partners tend to rely on communicating strategies that are less preferred in normal situations. Nevertheless, when one partner's naturally occurred eye gaze was made available to the other during the interaction, their collaboration became significantly more efficient. These results imply that monitoring and sharing naturally occurred eye gaze provides an important mechanism to facilitate language-based interactions between partners with mismatched perceptions such as in human robot communication.

## 3.2 Related Work

### 3.2.1 Collaborative Model of Referring

Conversation is a joint activity between its participants (Clark 1996). In conversation, participants coordinate their mental states based on their mutual understanding of their intention, goals, and current tasks (Clark 1992). An important notion, also a key to the success of communication is *grounding*, a process to establish mutual understanding between conversation participants. Specifically for the referential process in conversation, Clark and Wilkes-Gibbs developed *the collaborative model of referring* to explain the referential behaviors of participants (Clark and Wilkes-Gibbs 1986). This work states that grounding references is a collaborative process following *the principle of least collaborative effort* (Clark and Wilkes-Gibbs 1986):

... speakers and addressees try to minimize collaborative effort, the work both speakers and addressees do from the initiation of the referential process to its completion.

The collaborative model indicates that speakers tend to use different types of noun phrases other than elementary noun phrases during communication. The addressee attends to what has been said almost at the same time that the utterance is produced by the speaker. The speaker often adapts his language production in the middle of the planning based on the feedback from the addressee. Similarly, addressees make efforts to accept or reject references using alternative descriptions and indicative gestures (e.g., pointing, looking, or touching) (Clark and Brennan 1991). Different types of evidence can be used to indicate grounding of references such as back-channel responses, relevant next turn, and continued attention (e.g., indicated by eye gaze) (Clark and Brennan 1991). When an initial noun phrase is not acceptable, it must be refashioned. The collaborative model also identified three types of refashioning: repair, self-expansion, and replacement. Through these mechanisms, the speaker and the addressee strive to minimize the collaborative effort in grounding the reference (Clark and Wilkes-Gibbs 1986).

The collaborative model and the concept of grounding have motivated previous work on spoken dialogue systems (Traum 1994) and embodied conversational agents (Cassell et al. 2000). However, the implications of the collaborative model is not clear in situated dialogue where conversation partners have significantly mismatched capabilities in perceiving the environment. It is not clear in this setting how



participants strive to collaborate and minimize the collaborative effort in grounding references. Understanding these new implications is critical to enable the collaborative referential process between a human and an artificial agent such as a robot. The work presented in this paper is our first step towards understanding how participants with mismatched capabilities use language and indicative gestures such as eye gaze to coordinate the collaborative referring process.

### *3.2.2 Eye Gaze in Human Machine Communication*

Eye gaze serves many functions in human communication. It is the most basic form of what the addressee is attending to (Clark 1996). In face-to-face conversation, eye gaze can signal partners' intention to give or keep the turn (Novick et al. 1996). Addressees often signal their attentiveness with eye gaze (Argyle and Cook 1976). Recent work has also shown that speakers' gaze can provide early disambiguating cues for the addressee to interpret referring expressions (Hanna and Brennan 2007). Eye gaze can also facilitate grounding by establishing joint attention indicating mutual acceptance (Garrod and Pickering 2004; Pickering and Garrod 2004).

At the utterance level, psycholinguistic studies have shown that eye gaze is tightly linked to human language processing (Tanenhaus et al. 1995). Almost immediately after hearing a word, the eyes move to the corresponding real-world referent (Allopena et al. 1998). Directly before speaking a word, the eyes move to the mentioned object (Griffin and Bock 2000). Objects are fixated in the same order in which they are spoken (Bock et al. 2004). Not only is eye gaze highly reliable, it is also an implicit, subconscious reflex of speech. The user does not need to make a conscious decision; the eye automatically moves toward the relevant object, without the user even being aware.

Based on these empirical findings, gaze modeling has been incorporated in human computer interaction in several aspects. For example, the role of eye gaze in coordinating conversation is implemented in embodied conversational agents (Nakano et al. 2003). Previous work has also incorporated human eye gaze to help interpret user input (Kaur et al. 2003; Cooke and Russell 2008; Byron et al. 2005). In our group's own work, we have developed approaches to incorporate eye gaze for language interpretation in human computer dialogue. Our findings indicate that gaze fixation intensity serves as an integral role in attention prediction (Prasov et al. 2007). Incorporation of eye gaze can significantly improve automated language processing at multiple levels from recognition of spoken hypotheses (Qu and Chai 2007) and reference resolution (Prasov and Chai 2008, 2010) to automated vocabulary acquisition (Liu et al. 2007; Qu and Chai 2008). Motivated by previous work, our hypothesis in this paper is that eye gaze plays an important role in referential communication between partners with mismatched perceptions of the shared environment.

In human robot interaction, previous research has focused on the control of robot gaze behaviors to facilitate interaction (Sidner et al. 2004, 2005; Breazeal et al.

2004; Miyauchi et al. 2004; Staudte 2006; Yoshikawa et al. 2006; Mutlu et al. 2009a, 2009b), for example, to provide feedback of understanding (Breazeal et al. 2004), to signal engagement (Sidner et al. 2005), to demonstrate the role of participation (Mutlu et al. 2009a), and to indicate intention (Mutlu et al. 2009b). However, most of previous work on robot’s gaze controls is not concerned with language processing and interpretation of referents. More related work can be found in Staudte (2006), Kruijff and Staudte (2007), Staudte and Crocker (2009), where the gaze of the robot is controlled during the referential process based on simple strategies. A pilot study has shown statistical significant interactions between congruence (robot gaze to the intended referent) and the believability of the robot (Kruijff and Staudte 2007). A recent work has shown that the robot’s eye gaze can inform its human partners about its intended referents and help humans comprehend the robot’s speech (Staudte and Crocker 2009).

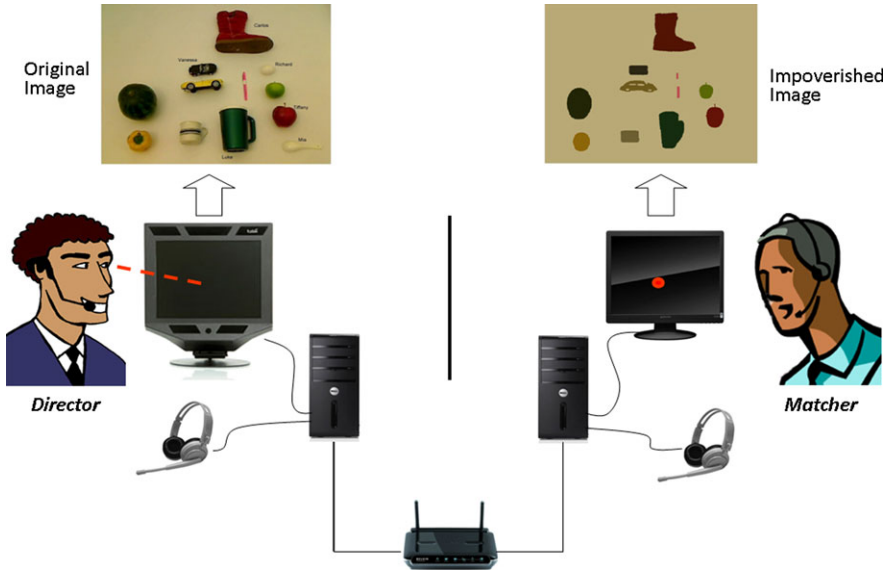
These previous works have provided empirical evidence on the important role of gaze modeling in human machine or human robot dialogue. To extend these previous works, here we specifically evaluate the role of shared gaze in facilitating collaborations between partners with mismatched visual perceptual capabilities. The empirical results will contribute to the design of collaborative robots that can effectively mediate this discrepancy through manipulations of shared gaze.

### 3.3 Method

The architecture of our experimental system is shown in Fig. 3.1. Two partners (a director and a matcher) collaborate on an object naming task. The director’s goal is to communicate the “secret” names of some objects to the matcher, so that the matcher knows which object has what secret name. They both face the same scene that is composed by some daily-life objects (office supplies, fruits, etc.). However, what they actually see can be different: while the director always sees the original image taken from the scene, the matcher sees either the original image or an impoverished version of the image, depending on the experimental condition. During their interaction, either the director’s or the matcher’s naturally occurred eye gaze is captured by a display-mounted eye tracker and can be made available to the other (shown as a gaze cursor on the other’s screen) in real time.

#### 3.3.1 Mismatched Views

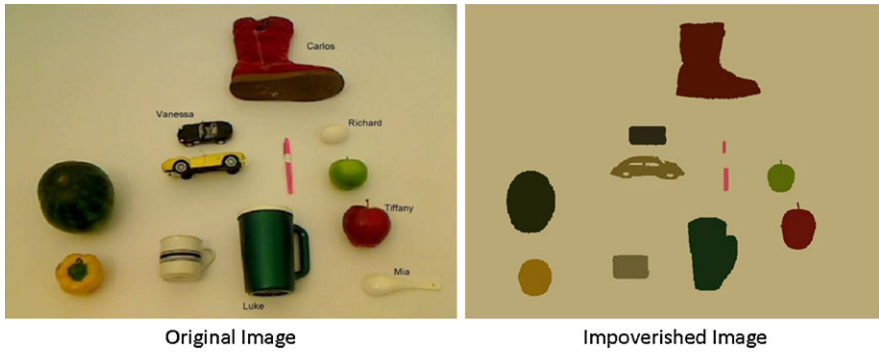
In previous studies (Passonneau et al. 2009; Levin and Passonneau 2006), an ablated Wizard of Oz paradigm was applied to investigate human strategies to deal with ASR errors. In these ablated Wizard of Oz studies, a human wizard’s capacities were incrementally restricted to simulate the capability of a real system, thus the system could learn better error-handling strategies from the wizard. Inspired by



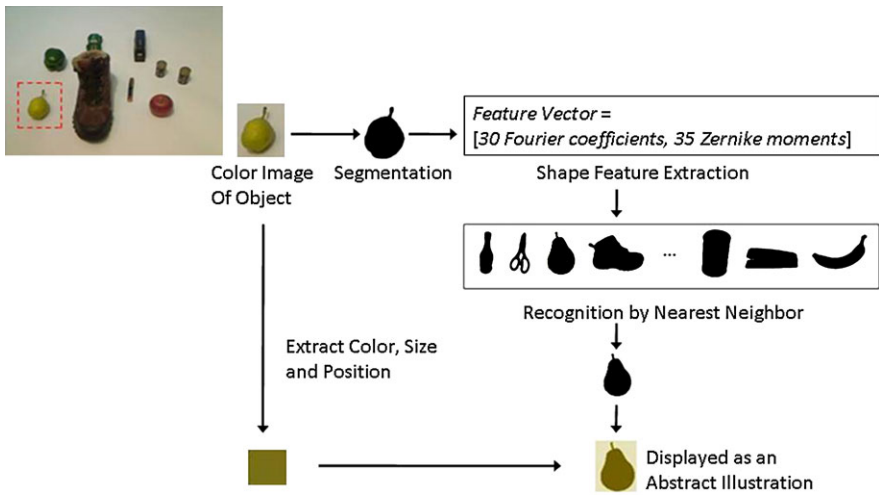
**Fig. 3.1** The architecture of our experimental system. Two partners in the same room are separated by a divider. The director is seated in front of a display-mounted Tobii 1750 eye tracker (Tobii Technology) and the matcher in front of a regular computer. Two computers are connected and synchronized via an Ethernet hub. The director’s eye gaze position is captured and can be displayed as a gaze cursor (a 32 by 32 pixels pink dot) superimposed over the matcher’s display. A bi-directional microphone-speaker system is used as the speech channel for two partners to verbally communicate with each other

those studies, we designed our experiment to investigate the collaborative strategies and the potential role of eye gaze in a referential communication task. To simulate mismatched perceptions, we ablated the matcher’s capability by only showing him/her an impoverished version of the original image. An example of the original and the impoverished images is illustrated in Fig. 3.2.

To faithfully simulate the perceptual capability of an artificial agent, we applied standard computer vision algorithms to process the original image and generate the impoverished representation of the same scene. This procedure is illustrated in Fig. 3.3. To create the impoverished image, we first used the OTSU algorithm (Otsu 1975) to separate foreground objects from the background. Then each segmented object was fed into a feature extraction routine that computed a set of region-based and contour-based shape features of the object (Zhang and Lu 2002). The feature vector of the object was then compared with all the “known” objects in a knowledge base. The object was recognized as the class of its nearest neighbor in the knowledge base. After this *segmentation*  $\rightarrow$  *feature extraction*  $\rightarrow$  *recognition* pipeline, the final outcome was then displayed as an abstract illustration in the impoverished image. For instance, if an object was recognized as a pear, an abstract illustration of pear was displayed in the impoverished image at the position of the original object. The color of the illustration was set to the average color of the pixels of the



**Fig. 3.2** An example of the “mismatched views” in our experiment. The *left image* is the original image which is always shown to the director. The *right image* is the impoverished version of the *left image*, which is shown to the matcher depending on the experimental condition



**Fig. 3.3** The procedure of generating the impoverished image from the original image

original object, and the height and width were set according to the bounding box of the original object.

### 3.3.2 Experiment Design

As mentioned in Sect. 3.1, we are mainly interested in two specific questions:

1. How difficult is it for partners with mismatched representations of the shared environment to collaborate?

## 2. To what extent does the collaboration benefit from the shared gaze between partners?

To address those questions, we designed a 2 by 2 factorial experiment to investigate the effects of two factors on the task performance. We denote the two factors as *view* and *gaze*, each of which has two levels:

- *View*: whether the director and the matcher have matched views or mismatched views. “*v+*” means their views are matched, i.e. both of them see the original image; “*v-*” means the views are mismatched, i.e. the director sees the original image, but the matcher sees the impoverished image.
- *Gaze*: whether shared gaze is available between partners. “*g+*” means one partner’s eye gaze is captured and rendered as real-time gaze cursors on the other partner’s screen; “*g-*” means eye gaze is not available to the other partner.

Based on the above two factors, we have a total of four experimental conditions:

- *v + g-*: matched views without shared gaze.
- *v - g-*: mismatched views without shared gaze.
- *v + g+*: matched views with shared gaze.
- *v - g+*: mismatched views with shared gaze.

Besides the two main factors, two other nuisance factors may also affect the task performance: the randomly composed scenes and the participants. Therefore, a Latin square design (Montgomery 2008) was used to block the possible effects of these two nuisance factors.

### 3.3.3 Participants and Procedure

Sixteen (eight pairs) undergraduate/graduate students from Michigan State University were recruited to participate in our experiments. In each pair of participants, one played the role of the director and the other played the role of the matcher throughout the entire experiment. Each pair of participants went through two experiments. In the first experiment, the matcher’s eye gaze was captured and shared. In the second experiment, the director’s gaze was captured and shared. In both experiments, each pair of participants went through four trials with different conditions as described above. In each trial, the director needed to communicate to the matcher the secret names of six randomly selected objects from a total of twelve objects in an image. When the matcher believed that he/she acquired the name of an object, he/she needed to record the name by clicking on the object and repeating its name, e.g. to click on the pear in his image and say “this is Alexis”. A trial was finished when all the secret names had been recorded by the matcher. We put no restrictions on what the two partners could say to each other. The only requirement was to finish each trial as quickly as possible. Each experiment lasted approximately 40 minutes.

**Table 3.1** Time (seconds) spent to finish each trial

Pair of participants	Image			
	1	2	3	4
1	$(v + g+) = 63.9$	$(v + g-) = 75.3$	$(v - g-) = 83.2$	$(v - g+) = 69.8$
2	$(v - g+) = 62.5$	$(v - g-) = 103.6$	$(v + g+) = 66.1$	$(v + g-) = 70.5$
3	$(v - g-) = 82.1$	$(v - g+) = 53.6$	$(v + g-) = 53.3$	$(v + g+) = 49.9$
4	$(v + g-) = 49.1$	$(v + g+) = 44.8$	$(v - g+) = 71.4$	$(v - g-) = 128.2$
5	$(v + g+) = 52.3$	$(v + g-) = 51.0$	$(v - g-) = 123.8$	$(v - g+) = 61.8$
6	$(v - g+) = 122.7$	$(v - g-) = 245.4$	$(v + g+) = 80.7$	$(v + g-) = 90.3$
7	$(v - g-) = 218.8$	$(v - g+) = 140.5$	$(v + g-) = 55.5$	$(v + g+) = 74.5$
8	$(v + g-) = 80.4$	$(v + g+) = 55.5$	$(v - g+) = 65.0$	$(v - g-) = 151.7$

### 3.4 Result and Discussion

During the experiment, several kinds of information were logged, including all the speech communications between the director and the matcher, the starting and ending time of each trial, the screen positions where the matcher clicked to record the names, etc. Two specific kinds of information were extracted from the logged data: the time the participants spent to finish each trial and the total number of utterances they issued in each trial. These two kinds of information were used as the response variables in our statistical analyses.

Our results have not found statistical significance on the effect of sharing the matcher's gaze. This could be partly due to the experiment design. During the experiments, the matchers were asked to use the mouse to click on the communicated objects and record the names. Using the mouse could possibly interfere with the matchers' gaze behaviors. Therefore, the finding of the role of the matcher's gaze is not conclusive, which requires further investigation. For the rest of this section, we will only report the results from experiments where the director's gaze is captured and shared by the matcher.

#### 3.4.1 Hypothesis Test

Table 3.1 shows the time (in seconds) spent to finish each trial and Table 3.2 shows the number of utterances of each trial. Based on our experiment design, we employed a 2 by 2 factorial ANOVA with replicated Latin square to analyze our data (Montgomery 2008). The analysis results are shown in Table 3.3 and Table 3.4. The ANOVA results show that the effects of both the *view* and the *gaze* are significant. It indicates that on one hand it is really difficult for the partners to collaborate and they had to spend more time and efforts when the views were mismatched, and on the other hand their collaboration became significantly more efficient when the director's eye gaze was available to the matcher during the interaction.

**Table 3.2** Number of utterances in each trial

Pair of participants	Image			
	1	2	3	4
1	$(v + g+) = 20$	$(v + g-) = 22$	$(v - g-) = 22$	$(v - g+) = 22$
2	$(v - g+) = 27$	$(v - g-) = 45$	$(v + g+) = 22$	$(v + g-) = 35$
3	$(v - g-) = 40$	$(v - g+) = 23$	$(v + g-) = 20$	$(v + g+) = 25$
4	$(v + g-) = 22$	$(v + g+) = 23$	$(v - g+) = 37$	$(v - g-) = 65$
5	$(v + g+) = 16$	$(v + g-) = 14$	$(v - g-) = 26$	$(v - g+) = 19$
6	$(v - g+) = 41$	$(v - g-) = 91$	$(v + g+) = 34$	$(v + g-) = 37$
7	$(v - g-) = 72$	$(v - g+) = 53$	$(v + g-) = 16$	$(v + g+) = 25$
8	$(v + g-) = 33$	$(v + g+) = 28$	$(v - g+) = 33$	$(v - g-) = 65$

**Table 3.3** ANOVA results for time

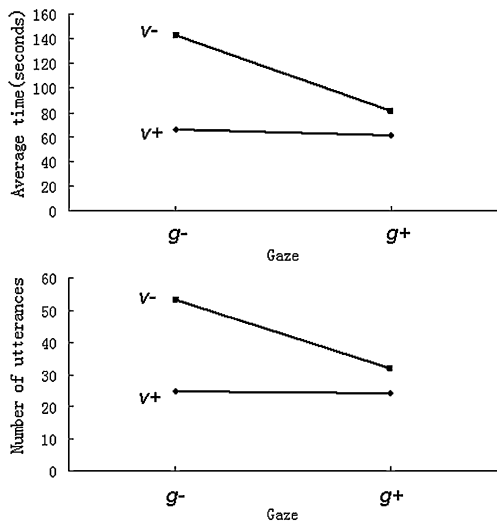
Source of variation	Sum of squares	Degrees of freedom	Mean square	$F_0$	$p$ -value
<i>View</i>	18576.3	1	18576.3	24.4	<0.0001
<i>Gaze</i>	8685.7	1	8685.7	11.4	0.0034
<i>Interaction</i>	6378.8	1	6378.8	8.4	0.0097
<i>Row</i>	10803.5	6			
<i>Column</i>	2010.0	3			
<i>Replication</i>	9200.5	1			
<i>Error</i>	13708.1	18	761.6		
Total	69362.9	31			

The interaction effect between these two factors is also significant, which indicates that the shared gaze could be more helpful under mismatched views. The interaction plot shown in Fig. 3.4 reveals more details. As it shows, when the partners had matched views, the effect of gaze was not significant (paired t-test between  $v + g-$  and  $v + g+$ ,  $t(7) = 1.07$ ,  $p > .3$  for time and  $t(7) = .32$ ,  $p > .75$  for number of utterances). Since our task was relatively easy when the two partners had matched views, thus they could finish one trial efficiently either with or without the help of gaze. As expected, gaze became very helpful under the condition of mismatched views. In this situation, making the director's gaze available to the matcher resulted on average 61.2 seconds shorter time and 21 less utterances to finish a trial. The improvement of collaboration efficiency is significant ( $v - g-$  versus  $v - g+$ ,  $t(7) = 4.97$ ,  $p = .002$  for time and  $t(7) = 3.91$ ,  $p = .006$  for number of utterances). The results have confirmed our initial hypothesis that collaboration can benefit from the shared gaze, especially when the partners have mismatched perceptions of the shared environment.

**Table 3.4** ANOVA results for number of utterances

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F_0$	$p$ -value
<i>View</i>	2610.1	1	2610.1	26.0	<0.0001
<i>Gaze</i>	979.1	1	979.1	9.7	0.0059
<i>Interaction</i>	850.8	1	850.8	8.5	0.0093
<i>Row</i>	2713.4	6			
<i>Column</i>	619.9	3			
<i>Replication</i>	552.8	1			
<i>Error</i>	1808	18	100.4		
<b>Total</b>	<b>10134</b>	<b>31</b>			

**Fig. 3.4** Interaction plot between *gaze* and *view*



We further examined the accuracy of our participants in the referential communication tasks.<sup>1</sup> All the pairs of participants made no mistake (100 % accuracy) when the views were matched, and they achieved an average accuracy of 87.5 % when the views were mismatched. The impact of the mismatched views on accuracy is marginally significant (paired t-test between the average accuracy of *v+* and *v-* within each pair of participants,  $t(7) = 2.0$ ,  $p = .085$ ). Four out of eight pairs of participants achieved 100 % accuracy even under mismatched views. For those other four pairs who made mistakes under mismatched views, shared gaze helped to

<sup>1</sup>The accuracy is measured by the ratio between the number of objects that were correctly named by the matcher and the total number of objects to be named.



improve the accuracy from 58.3 % to 91.8 %. This effect is significant ( $t(3) = 4.97$ ,  $p < .02$ ).

### 3.4.2 *Mismatched Perceptions*

Under the mismatched views, it is more difficult for partners to collaborate in the naming task that is otherwise easy to do when the views are matched. The participants in our experiment spent 76.4 seconds longer and issued 28 more utterances on average to finish a trial, when gaze from the director was not made available. Even with the help of gaze, it still took an average of 20.0 seconds longer and 8 more utterances under mismatched views. Since we applied standard computer vision algorithms to generate the mismatched views, it implies that computer vision errors can significantly impact the interaction between humans and artificial agents. In general, there are two types of errors:

- (1) Recognition error: an object or part of an object is correctly separated from the background and other objects, but it is not correctly recognized due to the insufficiency of machine recognition or partial segmentation.
- (2) Segmentation error: three cases are considered as this type of error.
  - Missing object: an object cannot be separated from the background, thus it is unseen to the agent.
  - Merging error: two or more objects cannot be separated from each other, e.g. due to overlapping, and together they are treated as a single object.
  - Splitting error: an object is split into separated parts, each of which is treated as a different object.

Due to these computer vision errors, a shared perceptual basis of the environment is missing, and communication between partners becomes more challenging. To overcome this difficult situation, artificial agents will need to integrate linguistic information, situated environment and non-verbal modalities such as eye gaze. Besides, error-tolerant approaches such as partial constraint satisfaction or inexact graph matching (e.g. Chai et al. 2004; Liu et al. 2012) will need to be pursued.

### 3.4.3 *Use of Spatial Language*

Despite the difficulties of collaboration under mismatched views, the participants in our experiments still performed reasonably well with an overall 87.5 % accuracy rate. An interesting question here is how they strive to collaborate under mismatched views. What strategies did they use to cope with the visual perception errors? By examining our data, we found that all of our participants overwhelmingly relied on using spatial language to ground referential communications. Here is one example from our data:<sup>2</sup>

---

<sup>2</sup>In our transcripts, “D” stands for director and “M” for matcher.

D: ok, the left, ah, upper there is a coffee cup named Ryan  
 M: this is Ryan  
 D: yes  
 D: top row left, um, actually in the middle of the top row is a small key named David  
 M: the middle of the top row?  
 M: where is the position to compare to some scissors  
 D: it is immediately to the left of the scissors  
 M: and what is the name again  
 D: David  
 M: this is David  
 D: the scissors next to it are named Lauren  
 ... ..

Spatial language, although less preferred in normal situations, turned out to be the spontaneously chosen strategy of all our participants in the experiments. Once the participants had gone through some practice trials and figured out that those object-properties based descriptions would not work well under the mismatched views, they switched to relying on spatial language. As demonstrated in the above example, spatial language, combined with the object-property descriptions, is a good strategy to establish common ground and facilitate referential communication. There are three types of spatial descriptions that were commonly used in our data:

- Relatum-based: the intended object (i.e. referent) is referred by its spatial relation with respect to another known object (i.e. a relatum), e.g. “*it is immediately to the left of the scissors*”.
- Group-based: the referent is identified by its relative position within a local group of objects, e.g. “*in the middle of the top row is a small key named David*”.
- Environment-based: the referent is identified by its position with respect to the global environment, e.g. “*the left, ah, upper there is a coffee cup named Ryan*”.

By properly issuing one of the three types of spatial descriptions based on the situation in hand, it can uniquely identify those objects with perceptual errors, which otherwise would be impossible or ambiguous to be referred using object-property based descriptions. Therefore, under the situation of mismatched views, spatial language becomes the most reliable and effortless channel to establish common ground. This possibly is the reason that spatial language became the best choice of collaborating strategies in our experiments.

For developing situated dialogue agents, spatial language brings in both great opportunities and challenges. Advanced spatial sensing technologies such as Microsoft’s Kinect can provide very accurate spatial representation of the environment. Since every object uniquely occupies a space in the environment, spatial language, if issued properly, can always be used to distinguish an object from all other objects (Kriz et al. 2007). However, spatial language understanding is also a challenging problem by itself. The use of spatial expressions presupposes underlying conceptual reference systems, or the so-called *frame-of-reference* (Levinson 2003). While

a significant amount of research has been focused on frames of reference and perspective taking in spatial cognition (Schober 1993; Carlson-Radvansky and Logan 1997; Tversky and Lee 1998; Bryant and Tversky 1999) and more recently in human robot interaction (Trafton et al. 2005; Moratz and Tenbrink 2006; Moratz 2006), automated recognition of frames of reference still remains a challenging problem (Tenbrink et al. 2007; Liu et al. 2010). Potential solutions to this problem perhaps should be based on designing sophisticated dialogue management and integrating non-verbal modalities such as eye gaze.

### 3.5 Conclusion

In situated human robot dialogue, robots and human partners have different capabilities in perceiving the shared environment. Because of the lack of shared visual basis of the environment, referential grounding is more difficult. To help design robots to better engage in referential grounding, we conducted experiments to examine how human partners with mismatched visual perceptual capabilities collaborate with each other. In particular, our hypothesis is that when shared visual basis is missing, eye gaze from partners can be especially important in the referential grounding process. Our empirical results validated this hypothesis. When the naturally occurred eye gaze of the director (i.e. the partner with a higher perceptual capability) was displayed to the matcher (i.e. the partner with a lower perceptual capability) in real time, the efficiency and accuracy of accomplishing the collaborative task were significantly improved. In addition, our results have further demonstrated the critical need of spatial language processing in mediating shared basis. All these findings have important implications for developing artificial agents that interact with humans in the real world.

**Acknowledgements** This work was supported by Award #1050004 and Award #0957039 from National Science Foundation and Award #N00014-11-1-0410 from Office of Naval Research.

### References

- Allopena PD, Magnuson JS, Tanenhaus MK (1998) Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J Mem Lang* 38:419–439
- Argyle M, Cook M (1976) *Gaze and mutual gaze*. Cambridge University Press, Cambridge
- Bader T, Vogelgesang M, Klaus E (2009) Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In: *Proceedings of the 2009 international conference on multimodal interfaces*. ACM, New York, pp 199–206
- Bock K, Irwin D, Davidson D (2004) Putting first things first. In: Henderson JM, Ferreira F (eds) *The interface of language, vision, and action: eye movements and the visual world*. Psychology Press, New York, pp 249–278
- Bohus D, Horvitz E (2009) Dialog in the open world: platform and applications. In: *Proceedings of the 2009 international conference on multimodal interfaces*. ACM, New York, pp 31–38

- Breazeal C, Hoffman G, Lockerd A (2004) Teaching and working with robots as a collaboration. In: Proceedings of third international joint conference on autonomous agents and multi agent systems (AAMAS'04), pp 1028–1035
- Brennan SE, Chen X, Dickinson CA, Neider MB, Zelinsky GJ (2008) Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106(3):1465–1477
- Bryant DJ, Tversky B (1999) Mental representations of perspective and spatial relations from diagram and models. *J Exp Psychol Learn Mem Cogn* 25(1):137–156
- Byron D, Mampilly T, Sharma V, Xu T (2005) Utilizing visual attention for cross-modal coreference interpretation. In: Proceedings of fifth international and interdisciplinary conference on modeling and using context (CONTEXT-05). Springer, Berlin, pp 83–96
- Carlson-Radvansky LA, Logan GD (1997) The influence of reference frame selection on spatial template construction. *J Mem Lang* 37:411–437
- Cassell J, Bickmore T, Campbell L, Vilhjalmsson H (2000) Human conversation as a system framework: designing embodied conversational agents. In: Cassell J, Sullivan J, Prevost S, Churchill E (eds) *Embodied conversational agents*. MIT Press, Cambridge
- Chai JY, Hong P, Zhou MX, Prasov Z (2004) Optimization in multimodal interpretation. In: Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL'04). Association for Computational Linguistics, Stroudsburg, article 1
- Clark HH (1992) *Arenas of language use*. University of Chicago Press, Chicago
- Clark HH (1996) *Using language*. Cambridge University Press, Cambridge
- Clark HH, Brennan SE (1991) Grounding in communication. *Perspect Soc Shared Cogn* 13:127–149
- Clark HH, Wilkes-Gibbs D (1986) Referring as a collaborative process. *Cognition* 22:1–39
- Cooke NJ, Russell M (2008) Gaze-contingent ASR for spontaneous, conversational speech: an evaluation. In: International conference in acoustics, speech and signal processing
- Fang R, Chai JY, Ferreira F (2009) Between linguistic attention and gaze fixations in multimodal conversational interfaces. In: Proceedings of the 2009 international conference on multimodal interfaces. ACM, New York, pp 143–150
- Garrod S, Pickering M (2004) Why is conversation so easy? *Trends Cogn Sci* 8:8–11
- Griffin ZM, Bock K (2000) What the eyes say about speaking. *Psychol Sci* 11(4):274–279
- Hanna JE, Brennan SE (2007) Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *J Mem Lang* 57(4):596–615
- Just MA, Carpenter PA (1975) Eye fixations and cognitive processes. *Cogn Psychol* 8:441–480
- Kaur M, Tremaine M, Huang N, Wilder J, Gacovski Z, Flippo F, Mantravadi CS (2003) Where is "it"? Event synchronization in gaze-speech input systems. In: Proceedings of fifth international conference on multimodal interfaces, pp 151–157
- Kriz S, Trafton JG, McCurry JM (2007) The role of spatial information in referential communication: speaker and addressee preferences for disambiguating objects. In: Proceedings of the 29th annual Cognitive Science Society
- Kruijff GJM, Staudte M (2007) Producing believable robot gaze when comprehending visually situated dialogue. In: *Language and robots: proceedings from the symposium (LangRo'2007)*
- Levin E, Passonneau R (2006) A WOz variant with contrastive conditions. In: Proceedings of the dialog-on-dialog workshop (Interspeech)
- Levinson SC (2003) *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press, Cambridge
- Liu Y, Chai JY, Jin R (2007) Automated vocabulary acquisition and interpretation in multimodal conversational systems. In: Proceedings of the 45th annual meeting of the Association of Computational Linguistics (ACL)
- Liu C, Walker J, Chai JY (2010) Ambiguities in spatial language understanding in situated human robot dialogue. In: Proceedings of the AAAI fall symposium on dialog with robots
- Liu C, Fang R, Chai J (2012) Towards mediating shared perceptual basis in situated dialogue. In: Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue. Association for Computational Linguistics, Seoul, pp 140–149

- Meyer AS, Sleiderink AM, Levelt WJM (1998) Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66(2):B25–B33
- Miyauchi D, Sakurai A, Makamura A, Kuno Y (2004) Active eye contact for human-robot communication. In: *Proceedings of CHI 2004*, pp 1099–1104
- Montgomery DC (2008) *Design and analysis of experiments*. Wiley, New York
- Moratz R (2006) Intuitive linguistic joint object reference in human-robot interaction: human spatial reference systems and function-based categorisation for symbol grounding. In: *Proceedings of the twenty-first national conference on artificial intelligence (AAAI)*
- Moratz R, Tenbrink T (2006) Spatial reference in linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations. *Spat Cogn Comput* 6(1):63–106
- Mutlu B, Shiwa T, Kanda T, Ishiguro H, Hagita N (2009a) Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In: *Proceedings of HRI*, pp 61–68
- Mutlu B, Yamaoka F, Kanda T, Ishiguro H, Hagita N (2009b) Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In: *Proceedings of HRI*, pp 69–76
- Nakano YI, Reinstein G, Stocky T, Cassell J (2003) Towards a model of face-to-face grounding. In: *Proceedings of the 41st annual meeting of the Association for Computational Linguistics (ACL'03)*, vol 1. Association for Computational Linguistics, Stroudsburg, pp 553–561
- Novick DG, Hansen B, Ward K (1996) Coordinating turn-taking with gaze. In: *Proceedings of the fourth international conference on spoken language, ICSLP 96*, vol 3, pp 1888–1891
- Otsu N (1975) A threshold selection method from gray-level histograms. *Automatica* 11:285–296
- Passonneau RJ, Epstein SL, Gordon JB (2009) Help me understand you: addressing the speech recognition bottleneck. In: *AAAI spring symposium: agents that learn from human teachers*. AAAI, Menlo Park, pp 119–126
- Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. *Behav Brain Sci* 27(2):169–189
- Prasov Z, Chai JY (2008) What's in a gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces. In: *Proceedings of the 13th international conference on intelligent user interfaces*. ACM, New York, pp 20–29
- Prasov Z, Chai JY (2010) Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In: *Conference on empirical methods in natural language processing (EMNLP)*, pp 471–481
- Prasov Z, Chai JY, Jeong H (2007) Eye gaze in attention prediction in multimodal human machine conversation. In: *Proceedings of the AAAI 2007 spring symposium on interaction challenges for artificial assistants*
- Qu S, Chai JY (2007) An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In: *Proceedings of the conference of the North America chapter of the Association of Computational Linguistics (NAACL)*, pp 284–291
- Qu S, Chai JY (2008) Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In: *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*
- Schober MF (1993) Spatial perspective-taking in conversation. *Cognition* 47:1–24
- Sidner C, Kidd CD, Lee C, Lesh N (2004) Where to look: a study of human-robot engagement. In: *Proceedings of the 9th international conference on intelligent user interfaces*, pp 78–84
- Sidner C, Lee C, Kidd CD, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. *Artif Intell* 166(1–2):140–164
- Staudte M (2006) *Grounding robot gaze production in a cross-modal category system*. Master's thesis, Saarland University, Saarbrücken, Germany
- Staudte M, Crocker MW (2009) Visual attention in spoken human-robot interaction. In: *Proceedings of HRI09*, pp 77–84
- Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217):1632–1634

- Tenbrink T, Maiseyenko V, Moratz R (2007) Spatial reference in simulated human-robot interaction involving intrinsically oriented objects. In: Proceedings of the symposium on spatial reasoning and communication at AISB'07 artificial and ambient intelligence
- Trafton JG, Cassimatis NL, Bugajska MD, Brock DP, Mintz FE, Schultz AC (2005) Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Trans Syst Man Cybern, Part A, Syst Hum* 35(4):460–470
- Traum D (1994) A computational theory of grounding in natural language conversation. PhD thesis, University of Rochester
- Tversky B, Lee PU (1998) How space structures language. In: Freksa C, Habel C, Wender K (eds) *Spatial cognition. Lecture notes in computer science*, vol 1404. Springer, Berlin, pp 157–176
- Yoshikawa Y, Shinozawa K, Ishiguro H, Hagita N, Miyamoto T (2006) Responsive robot gaze to interaction partner. In: *Proceedings on robotics: science and systems II*
- Zhang D, Lu G (2002) An integrated approach to shape based image retrieval. In: *Proceedings of the 5th Asian conference on computer vision (ACCV)*, pp 652–657

# Chapter 4

## Automated Analysis of Mutual Gaze in Human Conversational Pairs

Frank Broz, Hagen Lehmann, Chrystopher L. Nehaniv,  
and Kerstin Dautenhahn

**Abstract** Mutual gaze arises from the interaction of the gaze behavior of two individuals. It is an important part of all face-to-face social interactions, including verbal exchanges. In order for humanoid robots to interact more naturally with people, they need internal models that allow them to produce realistic social gaze behavior. The approach taken in this work is to collect data from human conversational pairs with the goal of learning a controller for robot gaze directly from human data. In a small initial data collection experiment, mutual gaze between pairs of people is detected and recorded in real time during conversational interaction. A Markov model representation of human gaze data is produced in order to demonstrate how this data could be used to create a controller. We also discuss how an algebraic analysis of the state transition structure of such models may reveal interesting properties of human gaze interaction. Results are also presented from a second, larger experiment in which mutual gaze is detected offline using recorded video data for greater accuracy. Trends in behavior linking gaze and speech in this data set are also discussed.

### 4.1 Introduction

The information revealed by the movement of the eyes during social interactions is essential for the coordination of the complex social behaviors characteristic for the human species. Compared to other primates humans have very visible eyes (Kobayashi and Kohshima 1997, 2001). A possible explanation for this phenomenon is the evolution of a new function of the human eye in close range social interactions

---

F. Broz (✉) · H. Lehmann · C.L. Nehaniv · K. Dautenhahn  
Computer Science, University of Hertfordshire, Hatfield, UK  
e-mail: [f.broz@herts.ac.uk](mailto:f.broz@herts.ac.uk)

H. Lehmann  
e-mail: [h.lehmann@herts.ac.uk](mailto:h.lehmann@herts.ac.uk)

C.L. Nehaniv  
e-mail: [c.l.nehaniv@herts.ac.uk](mailto:c.l.nehaniv@herts.ac.uk)

K. Dautenhahn  
e-mail: [k.dautenhahn@herts.ac.uk](mailto:k.dautenhahn@herts.ac.uk)

as an additional source of information about the intention of the other (Tomasello et al. 2007). In many studies it has been shown that apes and monkeys have only very limited abilities to follow a human experimenters eye movement to locate a hidden reward (Call and Tomasello 2003). Human infants on the other hand are able to reliably follow eye movements from around 18 months of age (Moore and Corkum 1998).

The importance of eye gaze especially during cooperative, mutualistic social interactions shows in the trouble humans with autism have in understanding the intentions of others which could be inferred from information contained in the eye region of the face (Baron-Cohen et al. 1995, 1997; Ristic and Kingstone 2005). Gazing and the ability to follow the eye gaze of others enables us to communicate non-verbally and improves our capacity to live in large social groups. It serves as a basic form of information transmission between individuals which understand each other as intentional agents. Additionally, human eyes signal relevant emotional states (Baron-Cohen et al. 1997, 2001) enabling us to interact empathically. More mutual gaze for example is positive for engagement (Cook and Smith 1975), while too much can be threatening or stressful (Mazur et al. 1980). As a consequence humans need eye gaze information to feel comfortable and to function adequately while interacting with others.

For most social interactions it is essential to coordinate one's behavior with one or more social partners. It is therefore not only necessary to transmit information, but also to jointly regulate the eye contact in an continues ongoing process with one another. This process is called mutual gaze (Argyle 1988). Mutual gaze involves always more than one individual. Being able to interact with others in this fashion is of great social importance from an early developmental stage and seems to be the basis of and precursor to more complex task-oriented gaze behaviors such as visual joint attention (Farroni 2003). Recent research in neuroscience suggest that episodes of mutual gaze may "prime" the brain for joint attention (Saito et al. 2010). Mutual gaze is also important for face-to-face communication. It is a component of turn-taking "proto-conversations" between infants and caretakers that set the stage for language learning (Trevarthen and Aitken 2001) and is known to play a role in regulating conversational turn-taking in adults (Kleinke 1986; Novick et al. 1996).

Since gaze behavior and mutual gaze is such an inherent part of our perception of others and of they way we interact with our social environment it seems imperative for the development of artificial systems fulfilling a role in this environment to understand the mechanisms of human gaze. This is especially true in cases where the system that a person interacts with has a humanoid form that includes eyes, as is the case with many interactive virtual agents or robots. Having the capability of producing readable gaze behavior may lead humans to expect these agents to exhibit natural and/or meaningful gaze. If these expectations are not met, the quality of interaction with the agent may be reduced or the agent may even be rejected as interaction partner.

There have recently been a number of studies on people's responses to mutual gaze with robots in conversational interaction tasks. But the models used to produce the robot's gaze behavior are typically either not based on human gaze behavior or



not reactive to the human partner's gaze actions. In work by Yoshikawa and colleagues, the robot responds to human gaze, but its gaze controller is not based on human data and does not take any action to regulate the duration or frequency of mutual gaze (Yoshikawa et al. 2006). In a story-telling robot study by Mutlu, Forlizzi and Hodgins, a robot produces human-directed gaze behavior based on a model with realistic timings that is not responsive to its audience's gaze (Mutlu et al. 2006). Yu and colleagues performed a temporal analysis of human gaze and speech behavior from a human-robot interaction word teaching task with a robot that autonomously performed a simple form of joint attention (Yu et al. 2010). While this study provides insight into patterns of human gaze at a robot, the simplicity of the robot's controller makes it unlikely that humans found the gaze interaction to be natural or its dynamics to be similar to gaze between two humans.

We hypothesize that correctly modeling the social aspect of gaze is important to achieving natural interactions between humans and agents that give gaze cues, and there is some experimental evidence to support this. In a study of interaction with a virtual agent, simple approaches to achieve high levels of mutual gaze through constant attentiveness by the agent led to negative reactions from the people the agent interacted with, demonstrating the need for a more realistic model (Wang and Gratch 2010). In a study comparing human tutoring behavior towards a human child and a childlike virtual robot, Vollmer and colleagues used a gaze controller based on low-level salience rather than the face-oriented nature of human social gaze (Vollmer et al. 2009). In their discussion of their results, they suggest that the robot's gaze policy may have affected tutoring behavior, causing people to interact differently with the robot because its gaze was noticeably dissimilar to a child's. Gaze behavior is part of conversational interaction, and the robot's gaze policy will have an impact on both the human's gaze behavior and the impressions they form about the agent they are interacting with. Robotic systems designed to learn language through interaction by exploiting the structure of child-directed speech (e.g., work by Saunders et al. 2010) could especially benefit from a gaze model that supports social engagement.

In order to support natural and effective gaze interaction between artificial agents and humans, it is worthwhile to first look at gaze behavior in human-human pairs. We propose that by using eye tracking software to record dyadic interactions in a setting with as few constraints as possible and computer programs to analyze the collected data in an automated manner we can achieve a much higher resolution in the examination of behavioral data than with usual methods like the manual coding of video recordings. The interpretation of the results from these analyses should give us the possibility to gain insight into how to build better gaze policies for agents that interact with people.

There has been some previous research into using automated collection of human-human gaze data to produce agent gaze. Raidt and colleagues conducted a study into face-to-face real time communication and gaze direction (Raidt et al. 2007). However, people interacted through a pair of video displays, which, while appropriate to their computer-agent model, unnaturally constrains people's options for movement (as opposed to co-located face-to-face conversation). Also, the speech

task involved was one of repetition and memorization rather than natural conversation. Given these constraints, it is unclear whether the data collected is representative of human conversational gaze behavior.

In this chapter we describe a set of studies in which we used an experimental setup that allowed us to monitor the dyadic eye gaze directions of social partners with a high degree of precision. To ensure the ecological validity of the data, the participants of our study were encouraged to engage in a free casual conversation. We will illustrate our automated approach to analyze this kind of behavioral data and give a first insight on how to generate a gaze controller for robotic platforms based on naturalistic gaze behavior.

Our goal is to underline the importance of a precise understanding of human eye gaze during conversations for the development of more comfortable human-robot interaction. We will show how the gaze and speech data can be represented by Markov models that express interaction dynamics and how algebraic analysis these models may reveal characteristics of the behavioral data.

## 4.2 Experiment 1

For our initial experiment, a real-time system for detecting mutual gaze between conversation partners was designed. This system was used to collect data from the conversational interaction of a small number of human-human pairs. The capabilities and limitations of this system for detecting mutual gaze will be discussed. Data collected from this experiment was used to produce a Markov model of the pairs' interaction behavior. The algebraic analysis of this model suggests the use of new mathematical approaches to assess the complexity of these types of social interactions.

### 4.2.1 System

The automated detection of mutual gaze requires a number of signal-processing tasks to be carried out in real time and their separate data output streams to be combined for further processing. Note that if the goal of this work were solely to study mutual gaze in humans rather than to provide input for a robot control system, there would be no requirement for real-time operation. The video could be collected and then analyzed later offline. The system is a mixture of off-the-shelf programs and custom-written software combining and processing their output. The interprocess communication was implemented using YARP (Metta et al. 2006).

ASL MobileEye gaze tracking systems were used to collect the gaze direction data (see Fig. 4.1 for an example) (Applied Science Laboratories 2009). The output of the scene camera of each system was put into real-time face-tracking software based on the faceAPI library (Seeing Machines, Inc. 2011). Each participant also



**Fig. 4.1** ASL MobileEye gaze tracking systems were worn by both conversation partners during the experiment

wore a microphone which was used to record a simple sound level (speech content was not stored during this experiment). Timestamped data of gaze direction (in  $x, y$  image pixel coordinates), location of the partner's facial features (in pixel coordinates), and microphone sound level were logged for each participant at a rate of 30 Hz. In order to synchronize time across machines to maintain timestamp accuracy, a Network Time Protocol (NTP) server/client setup was used. NTP is typically able to maintain clock accuracy among machines to within a millisecond or less over a local area network (Mills 1994).

### **4.2.2 Setup and Procedure**

Experiment participants were recruited in pairs from the university campus. A requirement for participation was that the members of each pair know one another. This restriction was used because strangers have been shown to exhibit less mutual gaze than people who are familiar with one another and because the conversational task could be awkward for participants to perform with a stranger. The pairs were seated approximately six feet apart with a desk between them. An example image of the experiment setup can be seen in Fig. 4.2. Ten pairs of people participated in the study.

Pairs were informed that they would engage in an unconstrained conversation for ten minutes while multimodal data was recorded. The participants were asked to avoid discussing upsetting or emotionally charged topics and given a list of suggestions should they need one, which included: hobbies, a recent vacation, restaurants, television shows, or movies. After filling out a consent form and writing down their demographic information, each participant was led through the procedure to calibrate the gaze tracking system by the experimenter before the trial began. Because



**Fig. 4.2** Two participants engaged in an conversation during Experiment 1

the gaze trackers used calculate gaze direction as a 2 dimensional image coordinate rather than as a 3 dimensional vector, it is important that the systems be calibrated for an image plane that is at the correct distance (the distance from the wearable camera to the partner's face) in order to obtain the most accurate possible results. This was achieved by using a board with numbered dots for calibration which was held at the distance of the partner's face while each participant was seated at the table as they would be throughout the experiment. During the experiment, the experimenter stayed out of sight of the participants behind a divider in order to monitor the computers running the tracking and data collection software.

### ***4.2.3 Data Analysis***

Of the ten pairs, five experienced errors during data collection that resulted in their data being discarded from the study. The nature of these errors were: loss of gaze tracker calibration due to the glasses with the camera mount slipping or being moved by the participant, failure of the face tracker to acquire and track the face of a participant, and failure of the firewire connection that was used to transmit the video data to the computers for analysis. These failures reflect the difficulty of deploying a real-time system for mutual gaze tracking due to the complexity of the necessary hardware and software components. The five remaining pairs of participants for whom complete face and gaze tracking data were available were used for data analysis. They ranged in age from 23 to 69. Of the pairs, two were male-male, two were male-female, and one was female-female.

The face location is reported by the face tracker in terms of facial features, specifically the eyes and mouth. The location and dimensions of these features are used to compute the bounding box for the face. The bounding box was computed based on the location of the mouth and eyes and width of the eye. This heuristic method

(rather than using the full contour defining the outline of the face) is easy to compute and typically results in a box that covers the width of the face and the vertical area of the face from the hairline to the chin. The location of the eyes are reported as a diamond-shaped outer counter. From these points, the width of each eye is computed. The larger of these two widths is used to define the face bounding box as: one eyewidth past the outside corner of each eye, three eyewidths above, and 1.5 eyewidths below the bottom of the mouth. The facial feature contours and the bounding boxes computed from them can be seen in Fig. 4.6.

The data was classified into high-level behavioral states depending on where both participants were looking and who was speaking at each timestep. In order to generalize across multiple interactions between different partners, each member of a pair was assigned an identifier based on their gaze behavior over the entire duration of experimental data selected for analysis. The pair member that looked the most at her/his partner's face during their interaction will be referred to as the "high" gaze participant and the partner with the lower level of face-directed gaze will be referred to as "low" (in all pairs observed, one participant looked at their partner noticeably more than the other). The gaze states and their descriptions are given below. Note that the states are mutually exclusive.

- Mutual—Mutual gaze, as defined as both participants' looking at one another's face area
- At Low—The high gaze level partner looks at the face of the low gaze level partner while they look elsewhere
- At High—The low gaze level partner looks at the face of the high gaze level partner while they look elsewhere
- Away—Both partners look somewhere other than their partner's face
- Unknown—Gaze state could not be classified due to missing gaze direction or face location data

It should be noted that the "Unknown" state may be caused by loss of gaze or face tracking for one of the participants at a timestep. There are a number of reasons this may occur: intermittent gaze tracking failure caused by gaze directed outside the field of view of the system, intermittent face tracking failure caused by rapid movement of the scene image (such as when a participant nods their head vigorously), or the conversation partner's face being absent from the scene image due to a participant's head direction. This state measures a combination of system error and participant behavior that cannot reliably be distinguished between using the current approach. Tracking participants' head orientations and their location in the shared 3D space would improve the system's ability to determine whether missing face location data was due to tracking failures or to turning the head away from the conversation partner.

The data was analyzed according to speaker role as well as gaze behavior. Which participant was speaking at a particular timestep was determined by computing the sum over a one-second wide sliding window for the sound level recorded from each participant's microphone and assigning the participant with the higher sum as the speaker. This was intended to smooth over brief pauses while speaking and detection errors. While the sound recording levels for the microphones were adjusted for

each speaker at the start of an experiment, the microphones still sometimes failed to detect quiet speech. These results most likely have classified some parts of both speakers' conversational turns as times when neither are speaking. In the second experiment, we revised our method of data collection for speech in order to improve its accuracy and stored full audio for later analysis. The high level states used for analysis were created by combining the gaze states described above with additional state information about which participant in the pair was speaking as follows:

- Neither Speaking—Neither participant is speaking for the last second
- High Speaking—High gaze level participant is speaking more over the last second
- Low Speaking—low gaze level participant is speaking more over the last second

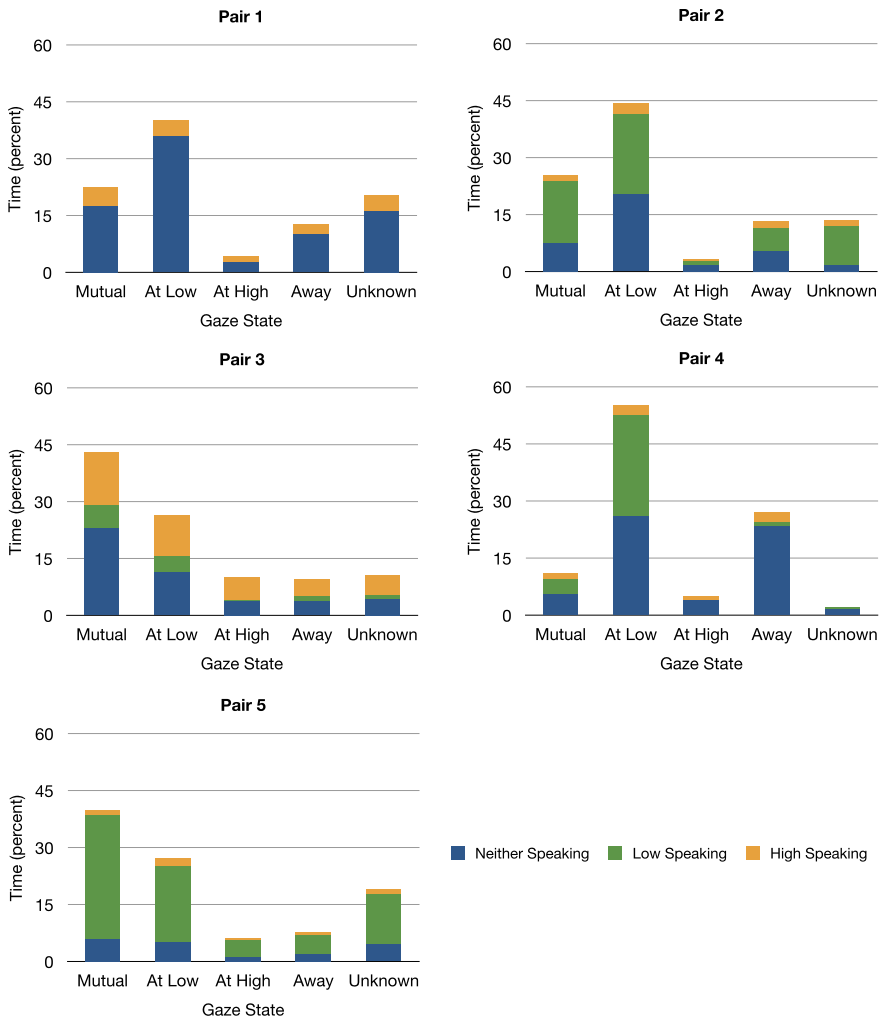
There are fifteen behavioral states in all.

#### **4.2.4 Results**

For each pair, the contiguous two minute period of conversation with the lowest amount of missing data was selected for analysis. The overall amount of time spent in each state by each pair is shown in Fig. 4.3. It can be seen that the amount of time spent in each gaze state varies a great deal among the pairs. This is because their behavior was likely determined by who was speaking as well as individual differences based on personality and characteristics of their interpersonal relationship. In the second experiment, we collect data from a larger set of participants so that we get a more clear picture of what constitutes average gaze and speech behavior in this conversational situation.

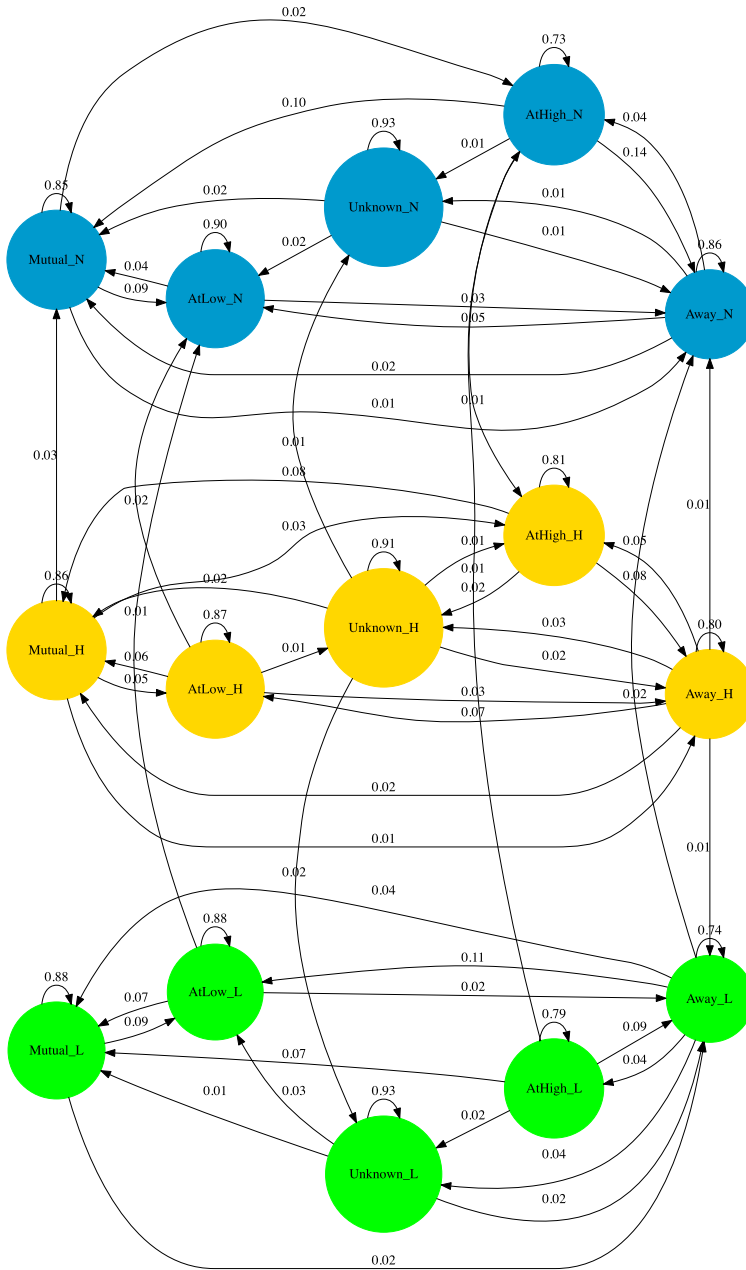
#### **4.2.5 Markov Model**

As a method of analysis and as a first step towards using this data to implement a gaze controller for a robot, we created a Markov model of the interaction using data from all five pairs. A Markov model (or Markov chain) is a graphical probabilistic model that describes the state transitions of a system or process (Meyn and Tweedie 1993). The same behavioral gaze state classifications for each timestep of data that were described in Sect. 4.2.3 and analyzed in terms of percentages in Sect. 4.2.4 were used as the discrete states of the Markov model. This model is shown in Fig. 4.4. Each gaze state of the interaction is a node in the model. The chance of reaching any other state from a given state at the next timestep is given by the probabilities on the outgoing edges from that state. The probability of staying in the same state at the next timestep is the probability of the state's edge that points back to itself. These self-transitions cause the time spent in each state to follow a geometric distribution, which agrees well with the form of the data observed. In order to improve the readability of the model and emphasize its major dynamics, transitions of less than 0.01 probability are not shown.



**Fig. 4.3** The percentage of time spent in each gaze state for each speech state for all of the conversational pairs in Experiment 1

It can be seen in Fig. 4.4 that the gaze states in which the same member of the pair is the speaker are highly connected. This reflects the fact that the gaze behavior varies at a faster timescale than the conversational turn. The model’s connections show that there may be different dynamics in the gaze behavior depending on who is speaking. It would be difficult to draw generalizable conclusions from this small data set, but this type of modeling provides us with a tool to examine the way that gaze behavior changes over time during an interaction.



**Fig. 4.4** Markov model of the gaze state transitions for all the conversational pairs in Experiment 1. A node's color and its ending label letter indicates the speaker role for its gaze state: neither speaking ("N", blue), high gaze partner speaking ("H", yellow), or low gaze partner speaking ("L", green)



### 4.2.6 Algebraic Analysis

It is possible to explore the interactions for hidden structure algebraically. Krohn-Rhodes Theory (or algebraic automata theory) established already in 1965 how to decompose any deterministic finite-state automaton into a series-parallel product of irreducible components (Krohn and Rhodes 1965), founding a field that has grown in mathematical sophistication since then. One of its founders, John Rhodes, suggested early on to apply the theory to the analysis of interaction, e.g. to analyze games, marriages or other interpersonal relationships (Rhodes 2009). This has not yet been carried out to date, but the methods apply equally to analysis of non-verbal interactions or other types of human-human interaction. Only in the last few years, however, have computational tools to carry out such a decomposition become available (Egri-Nagy and Nehaniv 2005, 2008, 2011). Markov models (such as the ones reflecting the dyadic gaze interactions) and non-deterministic automata in general can be converted to deterministic models using a standard power set construction, making it possible to use these decomposition methods to explore their structure.

Using this method, our preliminary analysis shows that pair 4's interaction is more complex than that of other dyads: the number of series levels needed to decompose the automaton corresponding to their interaction (using the holonomy method) is nearly twice that required for the other dyads, and also unlike the other pairs contains a non-trivial group.

The behavior of pair 4 is clearly distinct from the other pairs (as can be seen in Fig. 4.3) in that the overall amount of mutual gaze during the interaction is far lower, though we cannot yet characterize what relationship (if any) there is between this distinction in behavior and the observed differences in complexity. Pair 4 was one of the two male-female pairs we observed, and the most notable difference between them and the other groups was that they both indicated that they knew each other only "a little" on the questionnaire, while in all other pairs at least one participant answered that they knew the other "fairly well" or "very well". There is far to little data to determine whether this may play a role in the behavior differences observed, but it is an area for further investigation. We are currently exploring what aspects of interaction are reflected by this algebraic complexity.

## 4.3 Experiment 2

The first experiment demonstrated the power of using automated methods to detect mutual gaze, allowing real-time detection and providing data that had a high temporal resolution to capture the quickly changing dynamics of gaze interaction. However, the system employed had limitations both in tracking performance and in the ability to accurately detect speech. Rather than creating an entirely real-time system as in the earlier experiment, in the second experiment the face tracking was performed on the video stream offline after the completion of the experiment. This was done in order to ensure the best possible accuracy from the face tracking library, as

it would then not be restricted by the computational constraints of real-time performance. Additionally, high quality directional microphones and more sophisticated analysis of the audio signal were employed.

The size of the initial experiment and the short time period of analysis also made it unreliable to generalize the results from individual interactions. A larger number of participants were studied and longer conversations were analyzed in this second experiment in order to begin to identify common characteristics of conversational gaze and speech interaction.

### **4.3.1 System**

ASL MobileEye gaze tracking systems were used to collect the gaze direction data as in the initial experiment (Applied Science Laboratories 2009). Video of the scene and eye-directed cameras, as well as gaze direction data (in  $x$ ,  $y$  image pixel coordinates) indexed by its corresponding video frame, was logged by the gaze tracking systems' proprietary software. The output of the scene camera of each system (which is forward-mounted on the glasses worn for each system to capture the area the wearer is facing) was later input into face-tracking software based on the faceAPI library (Seeing Machines, Inc. 2011). The face-tracking software output facial feature coordinates in image pixels that were later used to compute a face bounding box, indexed by the video frame number. This allowed the conversation partner's face to be located in the video images from each participant's scene camera and compared to their gaze direction so that face-directed gaze could be detected. Two directional microphones were also arranged in such a way that each only recorded the speech of one of the participants. These microphones were used to record the audio track for each gaze tracking system's video.

### **4.3.2 Setup and Procedure**

Experiment participants were recruited in pairs from the university campus. The only requirement for participation in this study was that a participant be comfortable having a fifteen minute conversation in English with their experiment partner. The pairs were seated approximately six feet apart with a desk between them. After giving the instructions the experimenter disappeared behind a blind and was not present during the entire session.

The participants were informed that they would engage in an unconstrained conversation for fifteen minutes while multimodal data was recorded. The participants were also informed that they could discuss topics of their own choice and given a list of suggestions should they need one, which included: hobbies, a recent vacation, restaurants, television shows, or movies. Participants filled out a consent form and an additional form collecting their demographic information and level of familiarity



**Fig. 4.5** The setup for the second experiment was similar to that of the first, with the addition of the use of directional microphones to capture each participant's speech

with their partner. Each participant was led through the procedure to calibrate the gaze tracking system by the experimenter before the trial began. The setup for the second experiment can be seen in Fig. 4.5.

In order to allow the video streams from the two gaze tracking systems to be correctly aligned for analysis, the experimenter clapped his hands over the table between the participants (with the handclap visible in both systems' scene cameras) at the beginning and end of the experiment trial. During the course of the conversation, the experimenter stayed behind a divider out of sight of both participants so as not to be a source of distraction during the conversation. At the end of the session, participants were asked to complete a short paper questionnaire, the Ten Item Personality Inventory (TIPI) (Gosling et al. 2003), in order to evaluate their personality dimensions (this data is not used in the analysis described in this chapter). Thirty-seven pairs of people participated in this experiment.

### 4.3.3 Data Analysis

For the analysis, the data from 34 of the 37 pairs were used. Two pairs had to be excluded due to technical difficulties in the calibration process leading to the early termination of the experiment. One pair was excluded from analysis because of poor performance by the face tracking software due to large amounts of rapid head movement during conversation. Compared to its real-time performance, the offline performance of the face tracking library used was more robust to problems caused by rapid head movement by the participant wearing the camera. This is because more computationally intensive methods could be used without the real-time operation requirement. It is possible that real-time tracking would be made easier in the presence of head movement by the use of cameras with a higher frame rate (because there would be smaller differences in the field of view from frame to frame), but this system was restricted to the 30 Hz framerate of the gaze tracking system used.

The gaze data for a participant and the face tracking data for their conversation partner can be easily associated for each frame of scene camera video of the experiment conversation. This produces an individual data file of aligned gaze direction

and facial feature location data, expressed in pixel coordinates and indexed by frame number.

The data was analyzed according to speaker role as well as gaze behavior, as in the first experiment. Due to the limitations of the low quality microphones and crude approach to determining the speaker in the first experiment, an alternative method was used in the second. Each speaker's audio was recorded onto their video stream for a more thorough later analysis. The timesteps at which each participant was speaking were determined using Praat, a software tool commonly used for audio speech analysis (Boersma and Weenink 2011). A Praat script<sup>1</sup> was used to identify periods of silence between a speaker's utterances based on pitch and sound level information. The time of each frame of video data was compared to script output in order to determine whether it fell into a time period of speech or silence. This information was added to the gaze and face location for that frame in the speaker's data log. Which participant was classified as speaking at a particular timestep was determined by computing the max sum over a one-second wide sliding window for the speech signal for each speaker, as it was in the first experiment.

Because we are interested in measuring mutual gaze, the data for both individuals in a conversational pair must be combined so it can be determined where each was looking during the interaction. It is critical to correctly align the face and gaze data for each participant with their partner's so that data from the video frames recorded closest together in time are combined for analysis. This alignment was achieved by manually locating the frames in which handclaps occurred at the start and end of conversation in each partners' scene camera video files. Aligned frames from a conversation with their gaze and face tracking data overlaid are shown in Fig. 4.6.

The data was classified into the same high level behavioral gaze and speech states used in Experiment 1. The features making up these states are summarized in Table 4.1.

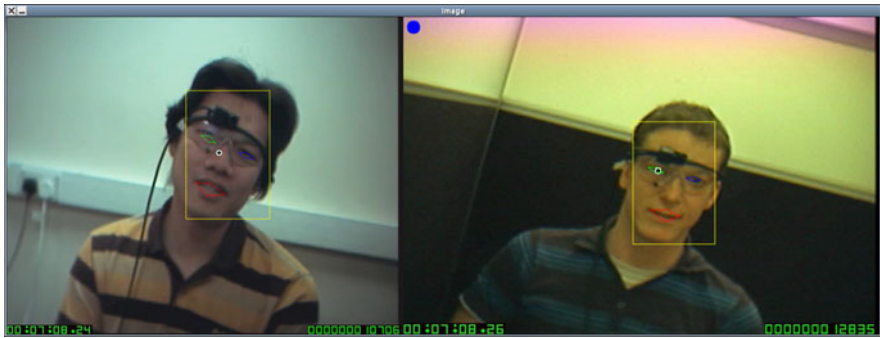
#### 4.3.4 Results

For each pair, the contiguous twelve minute period with the smallest amount of missing data was selected for analysis. The average percentage of time spent in each state for all of the experiment pairs analyzed are shown in Fig. 4.7.

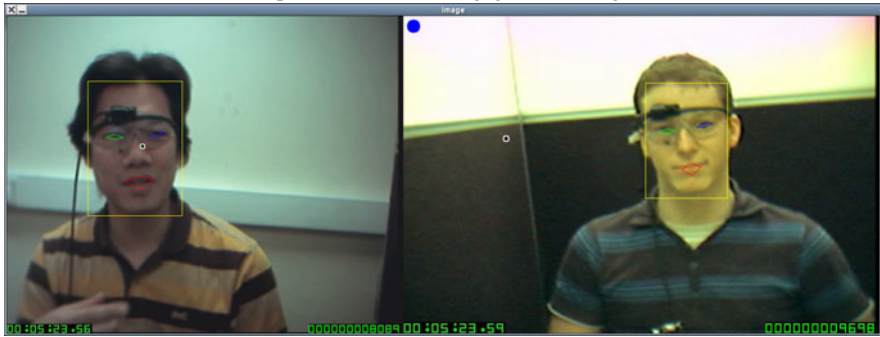
It can be seen in Fig. 4.7 that on average there is a large difference in the percentage of time that the high gaze and low gaze participants look at their partner. On average, during an interaction one participant will exhibit more face-directed gaze. Because mutual gaze can only occur when face-directed gaze is reciprocated, we hypothesize that the low gaze participant is the conversation partner that actually controls the amount of mutual gaze that occurs during a conversation. The low gaze participant in a pair can keep the amount of mutual gaze to a level that they

---

<sup>1</sup>This script was developed by Frank Kuegler and can be found at [http://www.ling.uni-potsdam.de/~kuegler/docs/praatut/mark\\_pauses.script](http://www.ling.uni-potsdam.de/~kuegler/docs/praatut/mark_pauses.script).



The speaker and listener engage in mutual gaze.



The listener gazes at the speaker's face while the speaker looks away.



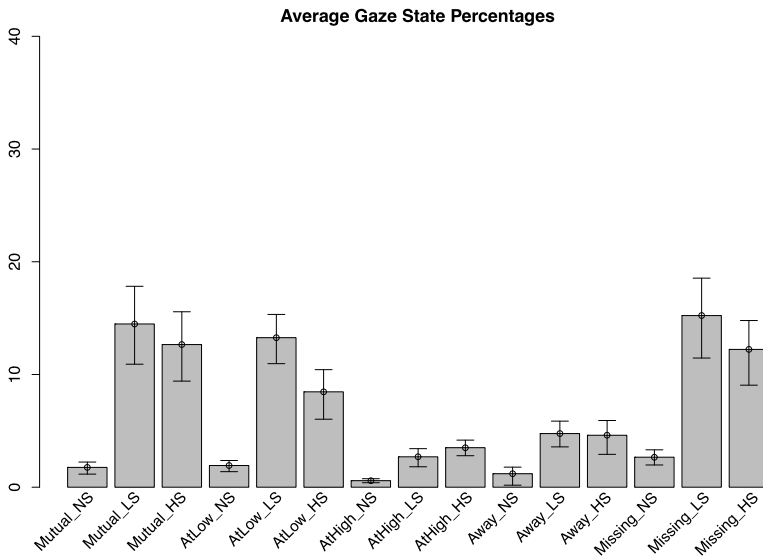
The speaker looks away from the listener's face, resulting in missing face location data.

**Fig. 4.6** Aligned video frames from the conversation between a pair of participants in Experiment 2. The facial features detected and face bounding box are shown. The gaze location is displayed as a *white circle with a black circle inside it*. The *blue dot* in the upper left of the *right image* indicates that the other partner is speaking

are comfortable with by less frequently reciprocating the face-directed gaze of their partner. We expect that the way in which individual gaze behavior produces different amounts of mutual gaze during different conversations is likely to be influenced both by conversational context and by the interaction of the individual traits of the conversation partners. Identification of these factors and the exploration of their im-

**Table 4.1** Definition of the features of the behavioral states for the experiment

Partner ID	
High	Performed the most face-directed gaze
Low	Gazed less at their partner’s face
Gaze state	
Mutual	Looking at each other’s face
At Low	High looks at Low’s face, Low looks away
At High	Low looks at High’s face, High looks away
Away	Both partners look away
Unknown	Missing gaze or face data
Speech state	
High Speaking (HS)	High gaze participant is speaking
Low Speaking (LS)	Low gaze participant is speaking
Neither Speaking (NS)	Neither participant is speaking



**Fig. 4.7** The average percentage of time spent in each gaze state for all conversational pairs in Experiment 2, shown with 95 % bootstrap confidence intervals

fact could give insight into how robot controllers that can establish comfortable mutual gaze with a variety of individuals in different situations should be designed.

An important issue in designing robots that interact with people is their perceived social dominance. Depending on the application and conversation partner, an agent

that behaves in a socially dominant manner may or may not be desirable. It can be seen in the data set that while the amount of mutual gaze was similar whether the high gaze or low gaze partner was speaking, the high gaze participant exhibited more unreciprocated face-directed gaze at the low gaze participant when they were listening to them (AtLow\_LS) than when they were speaking themselves (AtLow\_HS). Based on studies linking gaze while speaking versus listening to social dominance as in Dovidio and Ellyson's visual dominance ratio (Dovidio and Ellyson 1982), this suggests that the high gaze partners of a pair may on average be less socially dominant than the low gaze partners. The fact that the pair members' behavior did not follow the same pattern on average demonstrates that factors such as conversational role or the relationship between the conversation partners has an impact on each member's gaze behavior in relation to the other. Considering these differences may be important for designing controllers that exhibit natural-seeming mutual gaze.

## 4.4 Conclusions

In this chapter, a system for the automated detection of mutual gaze was described, and results were presented from two experiments measuring natural conversational interactions between human pairs. The real time system used in the first experiment is designed not purely for analysis, but to demonstrate that mutual gaze can be detected for use as input to a controller for a humanoid robot in the future. Preliminary approaches to estimating the current speaker role based on the sound level of speech were also employed in order to associate gaze state with conversational state and investigate the relationship between gaze behavior and conversational turns.

As a demonstration of how we intend to use this human-human gaze data to produce a robotic gaze controller, we created a Markov model from the data collected and discussed how it captures the gaze behavior dynamics of the human conversational pairs. These models allow the tracking of gaze and speech behavior. Related Markov models that allow for action selection, such as Markov decision processes (MDPs) or partially observable Markov decision processes (POMDPs), could be used by a robot to choose its conversational gaze behavior so as to optimize certain desirable characteristics of an interaction. Additionally, we present preliminary results from an algebraic analysis of the structure of the Markov model obtained from the data and discuss how this type of analysis may be used to computationally investigate qualities of the gaze interaction.

In response to the limitations of the original system's real-time operation, a revised approach based on offline analysis and employing improved audio processing was designed. The system was used to collect data in a larger second experiment focused on accurately measuring characteristics of human-human conversational gaze behavior. Results showed differences between the gaze behavior of the conversation partners making up a pair that may relate to who was the dominant partner in the conversation. Our results demonstrate that different conversational pairs may engage in very different amounts of mutual gaze. Understanding the characteristics of



the participants or the conversation itself that may lead to these differences and the relationship between the amount of mutual gaze and the quality of interaction for a particular conversational pair are major future directions for this work.

There are still limitations to the technical approach used that could be addressed in order to improve the accuracy of the system, primarily that the ability to detect mutual gaze is limited by the accuracy and robustness of the gaze tracking and face tracking systems employed. The limited field of view of the gaze tracking systems (which leads to missing data) could be overcome by using a more costly custom camera setup. In terms of speech data, the approach currently employed is very simple. While determining which partner is currently speaking automatically is relatively straightforward, there is other information about the conversational state that is more difficult to obtain through automated means. For example, it is not straightforward to determine which speaker's "turn" it is in the conversation, because a person might speak to indicate attention (backchanneling) or they might fall silent during the middle of a turn. The semantic content of speech is also difficult to determine without manual coding, and our work does not begin to address possible relationship between gaze and types of speech acts. Still, this work demonstrates the feasibility of the automated detection of mutual gaze and shows how mutual gaze behavior differs between different pairs of people. These differences highlight why it is necessary to design robot controllers that can engage in mutual gaze by detecting and adapting to the gaze behavior of their human partner.

**Acknowledgements** This research was conducted within the EU Integrated Project ITALK (Integration and Transfer of Action and Language in Robots) funded by the European Commission under contract number FP7-214668.

## References

- Applied Science Laboratories (2009) Mobile eye gaze tracking system. <http://asleyetracking.com/>
- Argyle M (1988) *Bodily communication*, 2nd edn. Routledge, London
- Baron-Cohen S, Campbell R, Karmiloff-Smith A, Grant J, Walker J (1995) Are children with autism blind to the mentalistic significance of the eyes? *Br J Dev Psychol* 13:379–398
- Baron-Cohen S, Wheelwright S, Jolliffe T (1997) Is there a "language of the eyes"? Evidence from normal adults, and adults with autism or Asperger syndrome. *Vis Cogn* 4:311–331
- Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I (2001) The 'reading the mind in the eyes' test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry* 42:241–252
- Boersma P, Weenink D (2011) Praat: doing phonetics by computer [computer program], version 5.3.03. Retrieved from: <http://www.praat.org/>
- Call J, Tomasello M (2003) Social cognition. In: Maestriperi D (ed) *Primate psychology*. Harvard University Press, Cambridge, pp 234–253
- Cook M, Smith JM (1975) The role of gaze in impression formation. *Br J Soc Clin Psychol* 14(1):19–25
- Dovidio JF, Ellyson SL (1982) Decoding visual dominance: attributions of power based on relative percentages of looking while speaking and looking while listening. *Soc Psychol Q* 45(2):106–113. <http://www.jstor.org/stable/3033933>



- Egri-Nagy A, Nehaniv CL (2005) Algebraic hierarchical decomposition of finite state automata: comparison of implementations for Krohn-Rhodes theory. In: *Implementation and application of automata (CIAA) 2004, revised selected papers*. Lecture notes in computer science, vol 3317. Springer, Berlin, pp 315–316
- Egri-Nagy A, Nehaniv CL (2008) Hierarchical coordinate systems for understanding complexity and its evolution, with applications to genetic regulatory networks. *Artif Life* (special issue on evolution of complexity) 14(3):299–312
- Egri-Nagy A, Nehaniv CL (2011) SgpDec: hierarchical composition and decomposition of permutation groups and transformation semigroups. <http://sgpdec.sourceforge.net/>
- Farroni T (2003) Infants perceiving and acting on the eyes: tests of an evolutionary hypothesis. *J Exp Child Psychol* 85(3):199–212. doi:10.1016/S0022-0965(03)00022-5
- Gosling SD, Rentfrow PJ, Swann WB Jr (2003) A very brief measure of the big five personality domains. *J Res Pers* 37:504–528
- Kleinke C (1986) Gaze and eye contact: a research review. *Psychol Bull* 100(1):78–100. doi:10.1037/0033-2909.100.1.78
- Kobayashi H, Kohshima S (1997) Unique morphology of the human eye. *Nature* 387:767–768
- Kobayashi H, Kohshima S (2001) Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. *J Hum Evol* 40:419–435
- Krohn K, Rhodes J (1965) Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Trans Am Math Soc* 116:450–464
- Mazur A, Rosa E, Faupel M, Heller J, Leen R, Thurman B (1980) Physiological aspects of communication via mutual gaze. *Am J Sociol* 86(1):50–74
- Metta G, Fitzpatrick P, Natale L (2006) YARP: yet another robot platform. *Int J Adv Robot Syst* (special issue on software development and integration in robotics) 3(1)
- Meyn SP, Tweedie RL (1993) *Markov chains and stochastic stability*. Springer, London
- Mills DL (1994) Improved algorithms for synchronizing computer network clocks. *Comput Commun Rev* 24:317–327. doi:10.1145/190809.190343
- Moore C, Corkum V (1998) Infant gaze following based on eye direction. *Br J Dev Psychol* 16(4):495–503
- Mutlu B, Forlizzi J, Hodgins J (2006) A storytelling robot: modeling and evaluation of human-like gaze behavior. In: *Humanoids*, pp 518–523. doi:10.1109/ICHR.2006.321322
- Novick D, Hansen B, Ward K (1996) Coordinating turn-taking with gaze. In: *Proceedings of the fourth international conference on spoken language, ICSLP 96*, vol 3, pp 1888–1891
- Raidt S, Bailly G, Elisei F (2007) Analyzing and modeling gaze during face-to-face interaction. In: *7th international conference on intelligent virtual agents, IVA'2007*, Paris, France, 17–19 September 2007, pp 100–101
- Rhodes J (2009) Applications of automata theory and algebra via the mathematical theory of complexity to finite-state physics, biology, philosophy, and games. World Scientific, Singapore
- Ristic J, Kingstone A (2005) Taking control of reflexive social attention. *Cognition* 94(3):B55–B65
- Saito DN, Tanabe HC, Izuma K, Hayashi MJ, Morito Y, Komeda H, Uchiyama H, Kosaka H, Okazawa H, Fujibayashi Y, Sadato N (2010) Stay tuned: inter-individual neural synchronization during mutual gaze and joint attention. *Front Integr Neurosci* 4:127
- Saunders J, Nehaniv CL, Lyon C (2010) Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors. In: *2nd international symposium on new frontiers in HRI*. AISB
- Seeing Machines, Inc. (2011) *faceAPI*. <http://seeingmachines.com/>
- Tomasello M, Hare B, Lehmann H, Call J (2007) Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *J Hum Evol* 52:314–320
- Trevarthen C, Aitken KJ (2001) Infant intersubjectivity: research, theory, and clinical applications. *J Child Psychol Psychiatry Allied Discipl* 42(1):3–48. doi:10.1017/S0021963001006552
- Vollmer AL, Lohan KS, Fischer K, Nagai Y, Pitsch K, Fritsch J, Rohlfing KJ, Wrede B (2009) People modify their tutoring behavior in robot-directed interaction for action learning. In: *Proceedings of the 2009 IEEE 8th international conference on development and learn-*

- ing, DEVLRN'09. IEEE Comput. Soc., Washington, pp 1–6. doi:[10.1109/DEVLRN.2009.5175516](https://doi.org/10.1109/DEVLRN.2009.5175516)
- Wang N, Gratch J (2010) Don't just stare at me! In: Proceedings of the 28th international conference on human factors in computing systems, CHI'10. ACM, New York, pp 1241–1250
- Yoshikawa Y, Shinozawa K, Ishiguro H, Hagita N, Miyamoto T (2006) The effects of responsive eye movement and blinking behavior in a communication robot. In: IROS, pp 4564–4569. doi:[10.1109/IROS.2006.282160](https://doi.org/10.1109/IROS.2006.282160)
- Yu C, Scheutz M, Schermerhorn P (2010) Investigating multimodal real-time patterns of joint attention in an HRI word learning task. In: 5th ACM/IEEE international conference on human-robot interaction, HRI'10. ACM, New York, pp 309–316. doi:[10.1145/1734454.1734561](https://doi.org/10.1145/1734454.1734561)

**Part II**  
**Gaze-Based Cognitive and Communicative**  
**Status Estimation**

# Chapter 5

## REGARD: Remote Gaze-Aware Reference Detector

Marc-Antoine Nüssli, Patrick Jermann, Mirweis Sangin,  
and Pierre Dillenbourg

**Abstract** Previous studies have shown that people tend to look at a visual referent just before saying the corresponding word, and similarly, listeners look at the referent right after hearing the name of the object. We first replicated these results in an ecologically valid situation in which collaborators are engaged in an unconstrained dialogue. Secondly, building upon these findings, we developed a model, called REGARD, which monitors speech and gaze during collaboration in order to automatically detect associations between words and objects of the shared workspace. The results are very promising showing that the model is actually able to detect correctly most of the references made by the collaborators. Perspectives of applications are briefly discussed.

### 5.1 Introduction

When two people discuss together about some shared visual content, gazes and speech become coupled. Indeed, we tend to look at the things we are talking about in a methodical way. More precisely, it has been shown that, during speech production and speech comprehension, eye-movements are closely related to verbal reference (Griffin and Bock 2000; Griffin 2001; Griffin and Oppenheimer 2006; Zelinsky and Murphy 2000; Meyer et al. 1998; Allopenna et al. 1998; Richardson and Dale 2005). For example, the pronunciation of a word referring to an object present in the visual field of the speaker is generally preceded, around one second

---

M.-A. Nüssli (✉) · P. Jermann · P. Dillenbourg  
CRAFT, EPFL, Lausanne, Switzerland  
e-mail: [marc-antoine.nuessli@epfl.ch](mailto:marc-antoine.nuessli@epfl.ch)

P. Jermann  
e-mail: [patrick.jermann@epfl.ch](mailto:patrick.jermann@epfl.ch)

P. Dillenbourg  
e-mail: [pierre.dillenbourg@epfl.ch](mailto:pierre.dillenbourg@epfl.ch)

M. Sangin  
Sony, London, UK  
e-mail: [mirweis@gmail.com](mailto:mirweis@gmail.com)

before, by a fixation on the corresponding object. Similarly, but to a lesser extent, a person hearing the name of a visually reachable object will often fixate the corresponding object a short time after having heard this name. In this work, we first replicated previous results in an ecologically valid situation. This is an important step if we consider that they were generally found in highly controlled settings. Second, based on these results, we developed an algorithm, called REGARD, that aims to identify automatically verbal references present in the dialogue of two collaborators working on a shared workspace. The principle of this algorithm is that, if a word, which is a verbal reference to a visual object, is pronounced several times, the corresponding speaker and listener's gazes before and after these utterances should be essentially distributed on the referenced object. Hence, by monitoring the gaze data of the speaker (respectively the listener) before (respectively after) each pronounced word and by computing how well they match over multiple pronunciations of the same word, we should be able to decide whether the word is a reference or not and if it is, to which object it is directed. We developed a model from this principle and we tested its effectiveness to detect automatically from a raw transcribed dialogue the words that are verbal references and to which object they refer to.

## 5.2 Background

The model developed in this contribution is largely based on the tight inter-coupling between gaze and speech. Hence, in the following section, we review the main results concerning this interplay.

### 5.2.1 Gaze and Speech

Several previous studies show that eye-movements may be related to collaborative activities. Indeed, it appears that gaze is largely influenced by speech which is at the heart of collaboration. Gaze appears to be used to monitor the environment while speaking or listening. More specifically, while speaking, there exists an *eye-voice span* which is a time delay between gazing at some specific object and uttering the name of that same object and conversely, while listening there is a *voice-eye span* which is the time between hearing the name of an object and the first gaze on that object.

For example, in a simple picture description task (Meyer et al. 1998), subjects started to gaze at the object to be named 700 ms before starting their utterance. Also they started to look at the second object to be named 300 ms before their utterance. In a more realistic situation, it has been shown in a simple sentence formulation experiment that speakers will tend to look at the things they are referring to just before pronouncing the corresponding words, in average 900 ms before the word onset (Griffin and Bock 2000). In this same study, Griffin and Bock showed that

this effect was not significantly affected by variations of causal structure (active or passive) of the formulated sentence. On the opposite the order of mention was the main factor explaining what was fixated when. These results suggest that the visual scene is used as a kind of cognitive support during sentence formulation. Other studies (Griffin 2001; Griffin and Oppenheimer 2006) show that the durations of the fixations on the referents is affected by the difficulty of retrieving the corresponding word and also it is increased when speakers are forced to use an inaccurate label. This latter result tends to show that these gazes on referents preceding speech are not simple aids to find the word but rather a cognitive support for sentence construction.

In a similar way, Allopenna et al. (1998) have shown that in a simple task in which people have to find a referenced object among four objects, they gaze at the referent object some time after hearing the corresponding noun. More precisely, the probability of fixating a referent becomes larger than the probability of fixating any other object 500 ms after the onset of the word and becomes close to one 800 ms after the onset of the word. This is of course a very simple situation not comparable to dialogue but this gives a first insight on the time required for gaze to follow perception of speech. We can however expect longer lags in more realistic, more complex situations in which the number of objects is larger, as this could require some time to find the object of interest.

The existence of those systematic spans between voice and gaze leads to a lagged coupling of the gazes of two interlocutors, as both the speaker and the listener monitor the objects that are referenced. Using cross-recurrence analysis,<sup>1</sup> Richardson and Dale (2005) showed that there exists a coupling between speaker's gazes and listener's gazes. More specifically, the listener tends to look at the same objects as the speaker with a delay of 2 seconds. Moreover, the level of this coupling appears to be related with the level of understanding of the listener, thus suggesting that this monitoring mechanism is important for the comprehension process. We can also note that this 2-seconds lag is coherent with the spans found in the studies cited above. Indeed, this lag should correspond to the sum of the speaker's span and of the listener's span, which is almost the case. Actually, the sum of the individual spans is slightly lower than 2 seconds which could be explained by a difference of complexity of the visual field.

### 5.2.2 Motivations

Building upon these findings, we investigated the *eye-voice* as well as the *voice-eye spans* from the eye movement and verbal interactions data collected during a computer-supported collaborative learning experiment. First, we replicated these results in our specific situation which contrasts a lot with the strictly controlled experimental settings used in the studies presented above. Indeed, our participants

---

<sup>1</sup>Cross-recurrence is a general measure that quantifies the similarity or the coupling between two dynamical systems.

were involved in a problem solving dialogue and a far more complex task, namely concept-mapping, in terms of social and cognitive processes. Hence, while we expect less systematic results, we believe that replicating these results in more complex and ecologically valid settings may provide insightful results and perspectives, if we keep the richness and complexity of the underlying processes in mind. Second, by using the data from the same experiment, we developed and tested a model that takes advantage of these phenomena in order to automatically detect which words are verbal references to objects and also which objects are referenced by these words. This model takes as input the transcribed speech of the collaborators, the list of objects present in the visual field and the synchronized streams of eye-movements of both partners. From these data, it categorizes the words of the speech into two categories, the common words, i.e. words which are not references to any of the object, and the reference words and for each reference word, it indicates the corresponding object. Such a model is a promising step for the design of multi-modal intelligent user interfaces, as it takes advantage of the natural relationships that exist between speech and eye-movements. This contrasts with classical speech or gaze-sensitive user interfaces in which the user has to behave in a specific predefined way (for example, by pronouncing specific keywords or by looking at specific places intentionally). On a more practical side, the proposed model could be used to annotate in real-time the visual field of collaborators in order to improve their interaction or to produce a visual output of their verbal interaction. Finally, this model has a strong potential for applications in artificial intelligence and robotics. Indeed, it could be used within a larger system that aims to identify objects, understand semantic relationships and communicate with humans.

## 5.3 Method

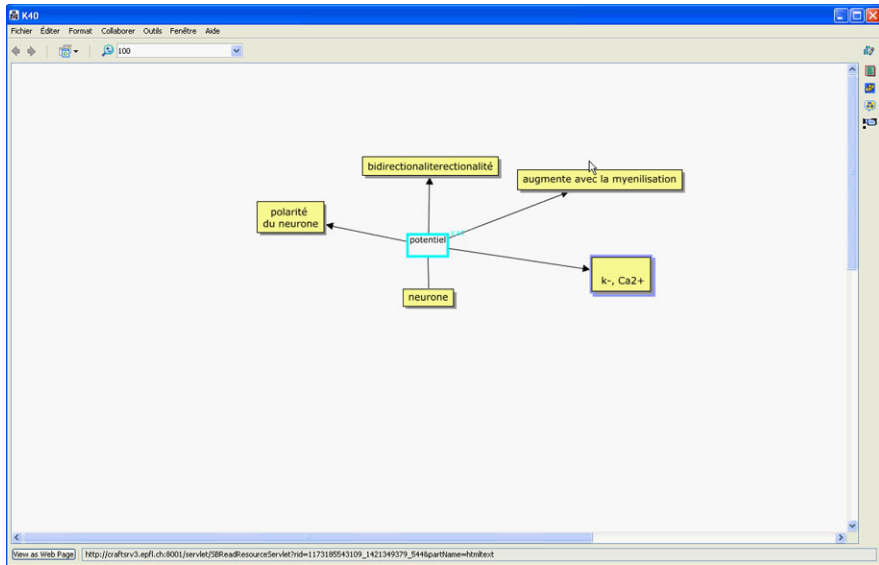
The experiment was conducted primarily to study the effect of a specific collaborative tool on the collaboration between two students learning together a complex topic. However, we will not describe it here in details as it is not the focus of this work but interested readers may refer to Sangin et al. (2011, 2008) and Sangin (2009) for more information about the experiment and the main results.

### 5.3.1 Task Description

The task consisted in building collaboratively a concept-map<sup>2</sup> to synthesize a text about “Neural transmission” read by both of the subjects individually beforehand

---

<sup>2</sup>Concept-maps are diagrams consisting of boxes representing concepts and labeled links representing relations between concepts.



**Fig. 5.1** Screenshot of the experiment. We can see the beginning of a concept-map built by the collaborators. Note that the labels are in French because the subjects were French-speaking

(see Fig. 5.1). The goal was to understand the concepts presented in the text by sharing their knowledge while constructing the concept-map. This consisted in drawing boxes (yellow boxes in Fig. 5.1) representing the main concepts of interest (neuron, axon, action potential, etc.) as well as connections between concept-boxes with linking phrases (“is a”, “contains”, “is produced by”, ...) to relate the concepts (see the arrows and the central white box in Fig. 5.1). Both subjects could modify the concept-map and the changes were immediately visible to their peer. They could speak to each other which was necessary to complete the task correctly. Dialogue excerpts can be found in the [Appendix](#).

### 5.3.2 Participants

Sixty-four French-speaker first semester university students (18 women and 46 men) were recruited and remunerated to participate to the study. Learners with a high degree of knowledge about the instructional material (i.e. the neural transmission) were filtered through a prior knowledge test and were excluded from the sample. Peers within same pairs did not know each other before the experiment.



### 5.3.3 Procedure

The two subjects were installed in front of two identical computer setups running the instructional material. The instructional material consisted of an explanatory text about the phenomenon of neural transmission, based on textbook materials and developed with the help of two experts of the domain. During the collaborative phase, we used an online concept-map building software, CMapTools.<sup>3</sup> The experimental session lasted about 90 min and consisted of six phases including two main learning phases: an individual explanatory reading phase and a remote collaborative concept-map building phase. These phases were:

1. Prior knowledge verification test. This was used to detect and remove potential experts of the domain.
2. Individual text reading. Subjects had 12 minutes to read and learn individually a text about neural transmission.
3. First learning test (pretest). A multiple-choice questionnaire to measure what they learn during the individual reading phase.
4. Collaborative concept-map instructions. Instructions about the collaborative phase with a short video tutorial on how to use the concept-map tool.
5. Collaborative concept-map building. Subjects had 20 minutes to build together a concept-map about the text they read. They could speak to each other through a headset.
6. Second learning test (posttest). It was similar to the pretest but with questions in a different order. It was used to assess what they learn during the collaborative phase.

### 5.3.4 Data Collection and Analysis

Gaze data of both participants were collected using two Tobii1750 eye-trackers. Speech was recorded by using a video-conferencing tool and was transcribed afterwards. In addition, all actions on the concept-map were also logged by the collaborative concept-mapping tool. These were logged within the same file and with the same time base for both subjects. A post-synchronization was performed first to match the time of both gaze streams with the time of the concept-map by finding correspondences between specific concept-map events and input events logged by the eye-tracker software (see Nüssli 2011, Chap. 4). A second post-synchronization was also accomplished to match the time of the audio recordings and the time of the gaze data. Finally, a spatial synchronization was accomplished as the two users could scroll independently in the shared workspace, which made their gaze coordinates not directly comparable.

---

<sup>3</sup>IHMC: <http://cmap.ihmc.us/>.

In order to analyze the precise timing between verbal references and gazes on the referenced objects, we had to identify within the whole speech corpus every time a reference to an object is made. Actually, the specific situation of this experiment greatly facilitates this process as the objects referred to are the concepts (yellow boxes in Fig. 5.1) and the linking phrases (white box in Fig. 5.1) drawn and labeled by the collaborators. Thus, we could consider, as an approximation, that when one of the collaborators pronounces the label of one of those concept-map objects, she is actually referring to this latter object. Hence, we simply matched each word of the transcript with the labels of all the objects present when the word was uttered. We used a fuzzy matching in order to prevent mis-detection due to typos or to very small differences between uttered words and objects labels.

Secondly, we also needed to have the precise onset for all detected verbal references. Indeed, the original speech transcription was not segmented at every uttered words but rather for utterances consisting of between 5 and 30 words. Hence, it was necessary to segment more precisely for those verbal reference words. We performed this step in two different ways. First, we manually segmented the verbal references of two selected dyads (11.2 % of the overall corpus). This consisted in selecting for each of the references the right portion of the corresponding audio file by using a transcription software (we used Transcriber<sup>4</sup>). This process is relatively easy and could not lead to important mistakes. Hence, a single person did this audio segmentation task alone. This resulted in 49 references for the first pair and 107 references for the second pair. Hence, we obtained a set of 156 verbal references for which we knew precisely the timing and thus, allowing us to perform precise analyses. In a second step, we developed and used an automatic transcript-speech alignment software using the Sphinx speech recognition library.<sup>5</sup> Since the speech data was in French, we had to use a language model, as well as an acoustic model, for French developed by LIUM (Deléglise et al. 2005). This engine was used to segment the original speech transcription into individual words in order to get the onsets for each individual word. This technique was applied on the dialogue of the 18 pairs for which we had workable data (i.e. sufficient quality speech data and eye-gaze data) and we detected a total of 431 verbal references with a mean of 23.9 references per pair. While this automatic solution allowed us to get much more data, it has the drawback that the resulting data are more noisy. Indeed, first, not all words are detected by the speech recognition engine. Second and more importantly, some words are incorrectly detected and consequently, get misaligned, thus resulting in incorrect onsets. Finally, even for correctly detected words, the alignment is sometimes not very precise and can lead to difference of several hundredth of milliseconds with the actual onset. To sum up, the results of this initial data processing are two sets of words which are verbal references with their precise onsets as well as their corresponding object in the concept map. The first set contains 156 references

---

<sup>4</sup><http://transag.sourceforge.net/>.

<sup>5</sup>CMU-Sphinx (<http://cmusphinx.sourceforge.net/html/cmuspinx.php>) is an open-source general speech recognition engine developed at Carnegie Mellon University.

and results from a manual segmentation of the words. It offers data of high quality with only correct and precise word onsets. The second set favor quantity over quality and contains 4 times more references (431) but potentially with more noise.

## 5.4 Eye/Voice Spans Analyses

As outlined above, our first contribution concerns the replication of the results found in the literature about the link between speech and gaze. This is the foundation upon which the REGARD model (see Sect. 5.5) is based.

### 5.4.1 Data Analysis

We analyzed for each verbal reference when, relatively to the reference onset, do fixations on the referred object occur. More precisely, our goal was to estimate the time, relative to the word onset, for which there is most chances that the speaker, respectively the listener, look at the referenced object. To do so, we computed the ratio of fixations falling on the referred object for different values of eye-voice span. We define the set  $V$ , of size  $N_V$ , of all verbal references, with  $v_i$  being an object referenced at time  $t_i$ . We also define the object fixated by the speaker at time  $t$  as  $G_S(t)$ . Then, for a given eye-voice span  $\delta_S$  (called speaker's span hereafter), the corresponding ratio of speaker's fixations on referenced object  $R_S(\delta_S)$  is computed according to the following formula:

$$R_S(\delta_S) = \frac{\sum_{v_i \in V} \mathbf{1}_{G_S(t_i - \delta_S) = v_i}}{N_V} \quad (5.1)$$

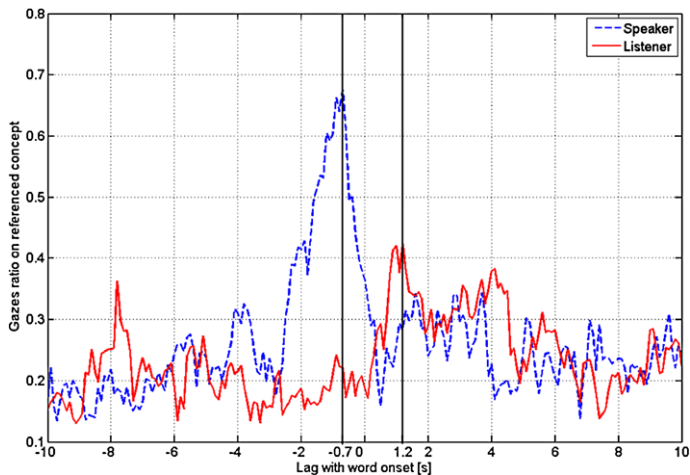
where  $\mathbf{1}_p$  is indicator function returning 1 when predicate  $p$  is true and 0 otherwise. Similarly, we define the object fixated by the listener at time  $t$  as  $G_L(t)$ . And we can define, for a given voice-eye span  $\delta_L$ , the corresponding ratio of listener's fixations on referenced object  $R_L(\delta_L)$  (called listener's span hereafter) with formula:

$$R_L(\delta_L) = \frac{\sum_{v_i \in V} \mathbf{1}_{G_L(t_i + \delta_L) = v_i}}{N_V} \quad (5.2)$$

We used these formulas to produce two curves. The first represents the proportion of speaker fixations on the referenced object according to the time relative to the verbal reference onset and the second represents the same thing but for the listener's fixations.

### 5.4.2 Results

We computed the ratio of matching fixations for the speaker and the listener separately for various spans between  $-10$  s and  $+10$  s using Eqs. (5.1) and (5.2). We

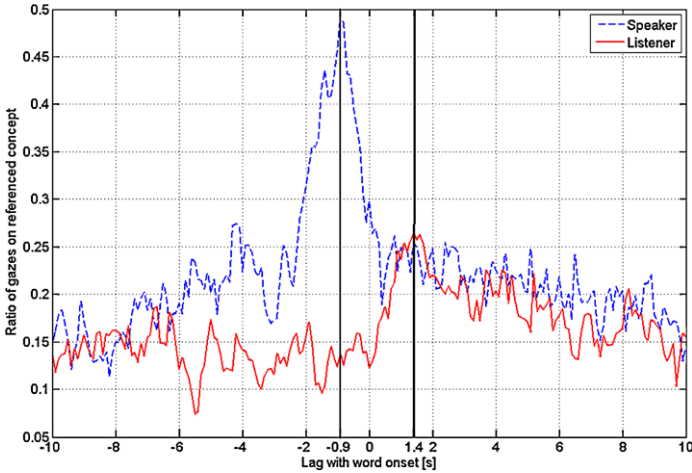


**Fig. 5.2** Average gaze ratio on referenced object for different time-spans from the verbal reference onset for manually segmented verbal references

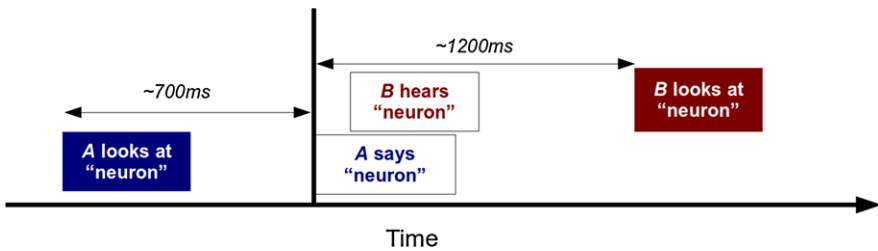
performed this computation both for the manually segmented data and the automatically segmented data (see Sect. 5.3.4). With regards to the hand-segmented verbal references produced by the two selected dyads, we plotted the gaze ratios of the speakers and the listeners for different time-spans in Fig. 5.2. The origin of the graph has a span of 0, which corresponds to the onset of the verbal references. Consequently, the negative values on the X-axis correspond to positive (respectively negative) values of the speaker’s (respectively listener’s) span; the positive values of the X-axis correspond to the positive (respectively negative) values of the listener’s (respectively speaker’s) span. The dotted line (ratio of speaker’s fixations on reference) peaks at  $-700$  ms with an associated average gaze ratio of 0.675. In other words, there is 67.5 % chance of getting a speaker’s fixation on the to-be-referenced object 700 ms before she formulates the verbal reference. With regards to the listeners’ gaze ratios on reference (continuous line), the peak is not as well defined as for the speaker. The highest average gaze ratio on the referenced object peaks at 0.425 at 1200 ms. Accordingly, there is 42.5 % chance that the listener looks at the referenced object 1200 ms after the speaker started to pronounce its name. In both cases, the peaks are not very sharp which indicates that these spans are not very systematic but may vary from one situation to another.

We replicated the same procedure for the automatically segmented verbal references and the results are shown in Fig. 5.3. The corresponding optimal speaker-span is 900 ms with a corresponding gaze ratio of 0.49. With regards to the listener, the peak is even more diffuse and reaches 0.27 of average gaze ratio for a span of 1400 ms. Figure 5.4 illustrates schematically these findings by showing the time-course of the gaze and speech when one person pronounces a verbal reference to the concept-box “neuron”.

These results give the probability that the speaker or the listener looks at the referred object at a specific time lag from the word onset but this does not tell us the



**Fig. 5.3** Average gaze ratio on referenced object for different time-spans from the verbal reference onset for automatically detected verbal references



**Fig. 5.4** Schematic time-course of gaze and speech during a verbal reference illustrating our main results

probability that the subject looks at the referred object for any lag in some range, for example between 0 s and 2 s from the word onset. Hence, we did additional computations with the automatically segmented data to obtain the probability of getting at least one fixation of the speaker on a referred object during a four second and a two second time-window before the onset of the verbal reference (namely between  $-4$  s and 0 and  $-2$  s and 0). There is a probability of 0.74 that the speaker fixates the referred-object within four seconds before producing the verbal reference and a probability of 0.69 within two seconds. For instance, if during an utterance a speaker refers to the concept “neuron”, there is a 74 % chance that she looked at the associated concept-box labeled “neuron” within the four seconds preceding the verbal reference, and a 69 % chance that she looked at the box within the two last seconds before producing the verbal reference. For the listener, there is a probability of 0.66 that the listener fixated on the referred-object four seconds after the verbal reference and a probability of 0.55 that the listener looked at the object within two seconds after the reference.

### 5.4.3 Discussion

Overall, both the manually and automatically segmented datasets provide results that are fairly supportive of the robustness of the speaker's span (eye-voice span) and listener's span (voice-eye span) mechanisms, even in realistic and complex collaborative situations. Indeed, we found that the ratio of speaker's fixations on referenced object peaked at 700 ms (900 ms with the automatically segmented data) before the onset of the reference. This is relatively consistent with the findings of Griffin and Bock (2000) who found an eye-voice-span between 800 ms and 1000 ms. Similarly, for the listener, we found that the ratio of fixations on the referenced object peaked at 1200 ms (1400 ms for the automatically segmented data) after the onset of the reference. Again, this is relatively consistent with the estimation of Allopenna et al. (1998) who found a voice-eye span between 500 ms and 1000 ms. The longer span that we found could be explained by the fact that we have a complex visual stimulus for which it can take more time to localize the referenced object, while Allopenna et al. (1998) used a very simple stimulus with only few objects to be looked at. The more diffuse nature of the listener's peak can be explained by the fact that the listener's span (or eye-voice span) is less systematic than the speaker's span and may depend on various factors, such as the familiarity of the subject with the stimulus or the number of objects present in the stimulus. Indeed, this voice-eye span may actually reflect, at least partly, the seek time required to find the referred object among all possible objects and this is certainly affected by variety of factors. In particular, the number of objects in the stimulus could play an important role in our case because it varied over time as subjects built the concept-map. More specifically, at the beginning of the task, they were generally no more than 3 objects, while it could reach up to 20 objects in the last minutes of the experiment. This could lead to longer spans for references made at the end of task.

The ratios of fixations on references for the automatically segmented data are much lower than for the manually segmented references. This is mainly explained by the fact that the automatically segmented data contains a number of false references and possibly also words which have been badly aligned with the audio file because of errors from the speech recognition engine. This could be certainly improved by further developing the alignment module and having better rules for the detection of references in speech. However, the analysis performed with the automatically segmented data shows that these results are generalizable and are not specific to the two manually segmented pairs.

Besides their empirical value, these results have an interesting application potential. Indeed, this relation between verbal references and eye-movements seems relatively robust, at least for the speaker (70 % chances of looking at the reference within the 2 seconds preceding the verbal reference) but also in a lesser extent, for the listener (55 % chances of looking at the reference within the 2 seconds following the verbal reference). These results, coupled with the recent technological advances, open exciting perspectives in terms of design and implementation of eye and speech-sensitive attentive technologies. In the next section, we present a prototype of such an application, called REGARD, that aims to use these phenomena to automatically detect verbal references in real-time.

## 5.5 REGARD Model

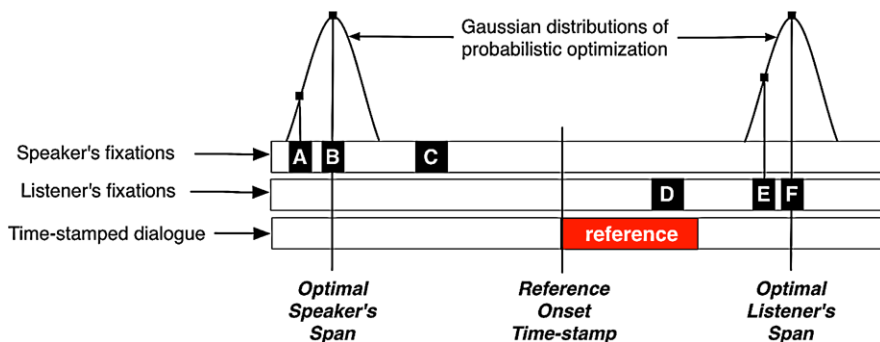
We developed an algorithm, called REGARD<sup>6</sup> (Remote Gaze Aware Reference Detector) that identifies automatically verbal references in the dialogue of two collaborators working on a shared workspace. The principle of this algorithm relies on the existence and the robustness of the eye-voice interrelation that we presented above. The idea is that, if a word, which is a reference to a visual object, is pronounced several times, the corresponding speaker and listener's gazes before and after these utterances should be essentially distributed on the referred object. Hence, by monitoring the gaze data of the speaker (and of the listener) before (respectively after) each pronounced word and by computing how well they match over multiple pronunciations of the same word, we should be able to decide whether the word is a reference or not and if it is, to which object it is directed. More generally, this model can be seen as a sort of classifier. It receives a set of words, along with the corresponding interlocutors' gazes and the list of objects in the workspace and it classifies these words either into common-words which are not reference to any object, or into reference-words which are references to objects of the workspace. Furthermore, it also provides for each of the words classified as references, the corresponding object of the workspace.

### 5.5.1 Model Design

In practice, the algorithm consists in creating for each pronounced word, a gaze density vector over all the possible objects and in aggregating at each pronunciation of a given word the speaker's and listener's gazes inside the corresponding vector. More specifically, we aggregate the fixations with weights which depend on the time of the fixation relatively to the word onset. The rationale is that the closer is a speaker's (or listener's) fixation from the optimal eye-voice (respectively voice-eye) span, the more chances it has to be on the referenced object. Hence, we give more weight to fixations which are close to the optimal span and less weight to those that are further. More precisely, we use a Gaussian function centered on the optimal spans to compute the weight of a fixation (see Fig. 5.5). After each aggregation of gazes in the gaze density vector associated to a given word, the algorithm performs some computations with the vector values to check whether there seems to be a verbal reference. More specifically, it first compares the total amount of gaze data that have been aggregated in the vector to a *minimum gaze threshold* to test whether these accumulated data may be considered as meaningful and thus if a decision can be taken for this word. Secondly, if enough data are present, it tests whether one cell of the vector contains a high amount of data compared to all other cells by comparing the sum of the maximum cell, i.e. the cell having the highest value, and

---

<sup>6</sup>“ REGARD” is also the French word corresponding to “gaze”.



**Fig. 5.5** Gaussian weighting functions used to aggregate the speaker and listener's gazes at each word utterance. Fixations *B* and *F* fully contribute to the model, whereas fixations *A* and *E* partially contribute to the model. Fixations *D* and *C* do not contribute to the model

the difference between the two maximum cells to a *matching threshold*. Indeed, if the word was a reference to an object, then most the accumulated gazes should be in the cell of the corresponding object and all other cells should have low values. Hence, if one cell contains a high amount of gazes relatively to all other cells then the algorithm marks this word as being a reference to the corresponding object. Otherwise, it tags the word as not being a reference to an object.

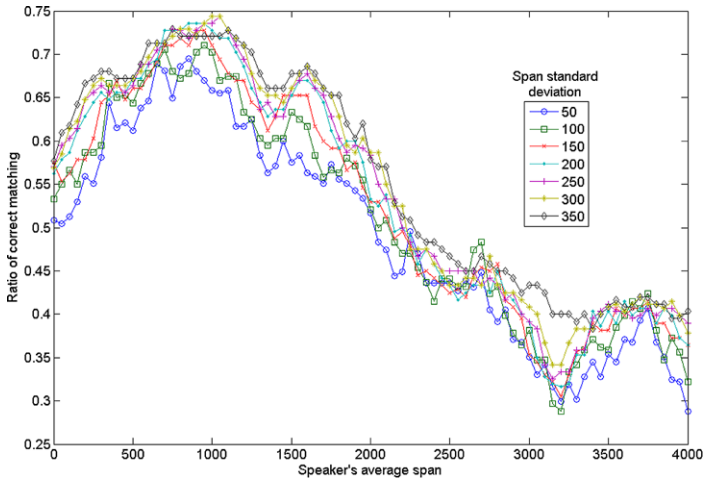
### 5.5.2 Parameters Optimization

Thanks to the data processing described in Sect. 5.3.4, we had a set of words for which we knew which ones were references and also which objects they refer to. This constitutes a ground-truth against which our algorithm could be tested. To build this dataset, we took only the 18 pairs that had high quality gaze data. Hence, we ended up with 301 references among the 587 manually or automatically segmented (see Sect. 5.3.4) and 737 words which were not references to any object, called common-words. Our first step was to use these ground-truth data to optimize the different parameters that appear in the algorithm.

First, we optimized the mean and the standard deviation of the Gaussian used as weighting functions. Second, we tried to identify the optimal values for the *minimum gaze threshold* used to assess whether a gaze density vector contains enough accumulated gazes and the *matching threshold* which assess whether the relative amount of gazes in a cell is sufficient to consider the word as a verbal reference.

The Gaussian parameters were optimized by using only the 301 reference-words. The procedure consisted in applying this algorithm, without the two decision steps, on these words and in computing for how many of these words, the highest cell in the gaze density vector corresponded to the associated object. This was done for several values of mean and standard deviation so that we found the optimal values for both of these parameters (see Fig. 5.6). The peak value suggests that the optimal





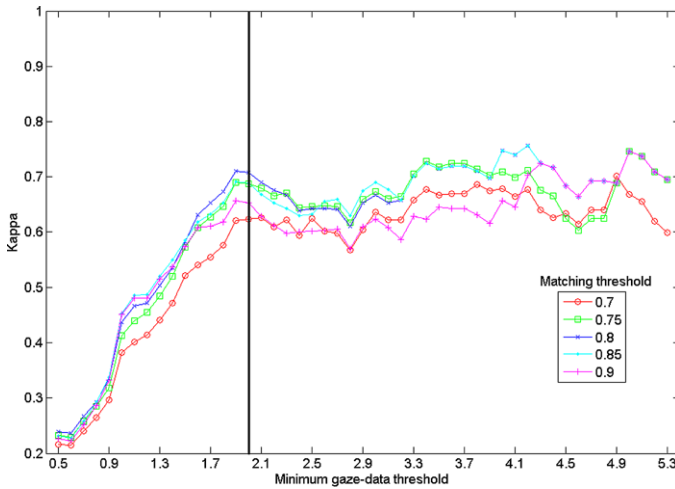
**Fig. 5.6** The speakers' average span and plotted against the model's detection score. The various curves represent different values for the standard deviation

values for the speaker's span mean is 800 ms and the optimal speaker's span standard deviation, the optimal value is 100 ms. A similar procedure was undertaken to detect the optimal listener's span mean and standard deviation which resulted in values of 1600 ms for the mean and 350 ms for the standard deviation.

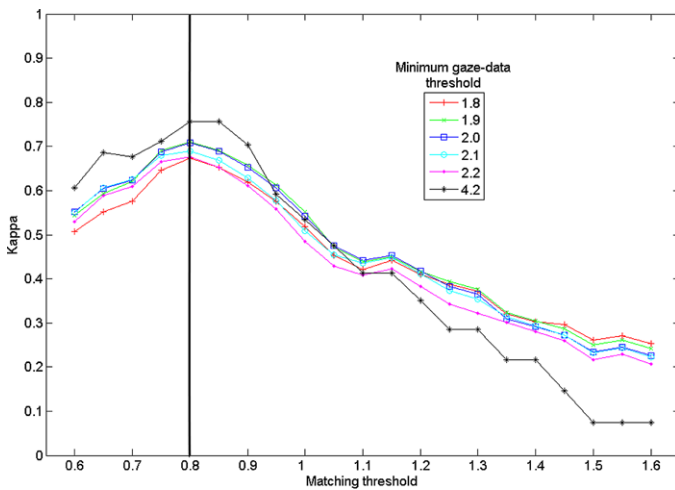
For the optimization of the two decision thresholds, namely the *minimum gaze threshold* and the *matching threshold*, the whole dataset was used and the kappa statistics measuring the agreement between the algorithm output and the reality was used as a measure of fitness. We optimized both parameters at the same time as there could potentially be some interaction between the two. Similarly than for the Gaussian parameters, we simply computed the algorithm performance for various couples of decision threshold values. We obtained the graphs shown in Figs. 5.7 and 5.8. First, we noted that there was very few interaction between the two parameters (the curves are almost parallels). From this, we deduced the optimal *minimum gaze threshold* above which the kappa score does not increase anymore (see Fig. 5.7) which corresponds to a value of 2. We also extracted the optimal *matching threshold* that maximizes the kappa score (see Fig. 5.8) which is 0.8.

### 5.5.3 Classification Results

In order to assess the performance of the model with optimal parameters, we ran the optimized algorithms on the whole dataset and measured how well results fit with reality. Note that while we could have used a classical validation technique, such as cross-validation, this is not a real concern in our specific situation because our model is highly specific and has very few parameters compared to the number of



**Fig. 5.7** REGARD classification performance plotted against the *minimum gaze threshold*. The **bold vertical line** shows the optimal value that has been chosen (2.0). Indeed, beyond this value, the performance doesn't increase anymore, suggesting that it doesn't help to have more aggregated gaze data to take the classification decision. The various *curves* are for different values of the other decision parameters (*minimum gaze threshold*) so that we can see a potential interaction between the two parameters



**Fig. 5.8** REGARD classification performance plotted against the *matching threshold*. The **bold vertical line** shows the optimal value that has been chosen (0.8). As we can see, it is simply the value which yields the best performance. A higher value would cause more reference-words to be detected as common-words, while a lower value would have the opposite effect (more common-words detected as reference-words)

**Table 5.1** REGARD model’s classification and matching performance details for the 182 words that passed the *minimum gaze threshold*. Incorrect cases (gray cells) have been grayed according to the seriousness of the error

	Classified as common-words	Classified as reference-words		Totals
		Correctly matched	Incorrectly matched	
Actual common-words	112	9		121
Actual reference-words	14	43	4	61
Totals	126	56		182

data (6 parameters for more than 1000 data samples). Hence, it is highly improbable that the optimal model has simply over-fitted the data used to optimize it.

Among the 1038 words composing the dataset, 182 passed the *minimum gaze threshold* allowing the algorithm to classify them as common or reference words. Table 5.1 shows the confusion matrix between the actual classification and the output of the algorithm for these 182 words. It indicates first whether the words were correctly classified as common or reference word and secondly if the identified reference-words were correctly associated to their object on the map.

The algorithm makes several types of errors which can be classified according to their seriousness level. The actual seriousness of an error certainly depends on the application for which the model is used but we can still make a general classification in order to get a rough picture of the actual performance of the model. The first type of error is reference words classified as common words (light gray cell), i.e. fail to detect that a word is a reference to an object, which is not a very serious type of error. Secondly, it can classify correctly a reference-word but associate it to the wrong object (medium gray cell) which is more serious error. Finally, it can incorrectly classify a common-word as being a reference-word (dark gray cell) which is also a serious type of error. These two latter types of errors are the most serious as they correspond to incorrect reference detection, i.e. the model “believes” that a word refers to an object while it refers to another object or it doesn’t refer to any object. However, they are quite few with only 13 cases over 182 words which represent 7 % of the words. Overall, the performance of the algorithm is quite good with a kappa score of 0.71. We should also note that some errors are due to the fact that the word onset timestamps, found by aligning automatically the speech transcript with the audio file, were not always correct and also that the reference-words are not always correct as they were also detected automatically by comparing words and concept’s labels on the map.

In Table 5.2, we detail the words incorrectly classified and/or incorrectly matched by the model in order to shed light on the different kinds of mistakes the model makes. We discuss hereafter these various mistakes.

- Reference-words incorrectly classified (false negatives): These words are those the model failed to classify. A closer look to the words points out to some ambigu-

**Table 5.2** Lists of the words incorrectly classified and/or incorrectly associated by the REGARD model. Note that the participants were French speakers. Hence, the words used are in French

Reference-words incorrectly classified	Reference-words matched with the wrong object	Common-words incorrectly classified
neurone	excitation	donc
met	phase	cellule
mettre	seuil	elle
pompe	canaux	pour
entre		cellule
intérieur		comme
potentiel		pense
polarisation		donc
		peut

ities. Indeed, “met”, “mettre” (put), and “entre” (enter) are seemingly common-words that could also have been used by co-learners in their map (e.g. “entre” could be identified as an actual reference-word because it can be considered as a part of the concept-box label “entrée de K+” or a common word, as in saying “enter”). The automatic technique used to define the testing dataset (see Sect. 5.3.4) may explain part of the model’s mistakes. More generally, this list seems to contain the most basic words of the domain knowledge. Therefore, it is possible that their use in the dialogue goes far beyond the simple referencing process explaining the model’s mistakes.

- Reference-words matched with the wrong object: These types of mistakes are probably those that we most would like to avoid as they result in an incorrect link between a word and an object. Fortunately, these errors are very rare with only 4 cases among the 182 words (2 %) for which the model took a decision. The most likely explanation for these mistakes is that these words have been used extensively while referring to another concept-box. This is supported by the fact the words under concern are likely to be part of an explanation involving other concepts (for example “phase” may be used to speak of the “rising phase” or “falling phase”).
- Common-words incorrectly classified: On a semantic level, we can see that among the nine false positives, two are potential references (i.e. “cellule”) and seven are actual common words. As “cellule” means cell and can be considered as a reference to a “brain cell” (consequently a synonym of “neurone”), we may argue that these false positives may be actual true positives on the socio-cognitive level. Collaborators may have referred to a concept-box they named “neurone” with the more general reference, “cell”. It can be argued that these types of mistake are mainly due to the limitation of the automatically defined dataset serving as reference to assess the model’s performance. The automatic detection of actual reference-words in the set of objects’ labels (see Sect. 5.3.4) is not able to identify synonyms.

As we have seen, it is likely that several of the errors done by the model could actually be explained by imperfections of the automatic method used to detect references in the speech corpus. Hence, these are not real mistakes but rather problems in the ground truth dataset. Thus, the actual performance of the system could be a little higher than the numbers reported above.

### 5.5.4 Discussion

We developed a multi-modal computational model that simultaneously monitors two (or potentially more) users' gaze patterns and verbal interactions in order to match the verbal references to the visual objects of reference. For sake of conceptualization and testing, we developed a version of the model which assumes the availability of certain data. More specifically, in addition to the gaze data, the REGARD model in its current iteration relies on two other datasets which may be not available in any situation. First, it requires to have a complete knowledge of all the objects that are visually accessible by the collaborators. In our case, these objects were simply the boxes and links drawn by the users on the map. They were easily accessible to us as these latter are logged by the concept-map software. However, in other situations, these objects and their geometry may be not directly accessible while in other cases, such as with natural images, the objects may even be difficult to define. The second kind of data that the current version of REGARD requires is the speech transcript. Indeed, we relied on the fact that we knew exactly the words that were uttered which allowed us to detect when a given word is uttered again and again. This data is clearly hard to obtain in a general situation as it has been accomplished by a manual process.

However, these two main limitations on the current implementation of REGARD, namely the need for a complete list of objects and for a transcribed version of the speech may be overcome in future versions. First, concerning the knowledge of all objects, this could be avoided by designing a slightly more complex version that would be based on workspace areas instead of objects (see Cherubini et al. 2008 for an example of a similar approach). Indeed, we could imagine that instead of maintaining a gaze vector associating aggregated gazes to the stimulus objects, we could a more complex structure that associate to every word gazes to region of the workspace. Then the algorithm would have to determine whether the accumulated gazes are mainly clustered in one region or rather are distributed uniformly across the workspace and if it appears that they are clustered, it could be decided that the corresponding word is a reference for that zone. Such a solution would however require that the workspace landscape doesn't change over time.

The second issue, namely the fact that the model requires transcribed speech, is more difficult to overcome. The best solution would be to have a working speech recognition module that would make the transcription automatic. However, while this technology is constantly improving, speech recognition has currently not yet reached a sufficient level of accuracy to perform such a task correctly in any situation. A possible workaround could be not to do full speech transcription but rather,

to detect similarity of sound between different part of the speech. The idea is to be able to detect several occurrence of a word at different moment in time. Hence, instead of associating written word with region or object of the stimulus, the algorithm would be able to associate sounds.

The results of the classification suggest that the REGARD model performs reasonably well at associating objects on the screen to specific verbal references. Given the complexity of the collaborative situation and its ecological validity, REGARD reaches a more than satisfactory level of performance. It is also noteworthy that the performance of the model depends on the fine-tuning of the parameters and thresholds. Even though the first verbal reference would be sufficient to establish, with reasonable confidence, matches between a set of fixations and a verbal reference, a greater number of occurrences (e.g. waiting for at least three co-occurrences of a verbal reference and associated fixation on a specific object) would lead to better confidence and precision. In other words, the performance of the model greatly depends on the quality and quantity of data it collects.

Furthermore, the performance of the model also depends on the complexity of the data (i.e. knowledge domain). The experiment done in this work is characterized by a highly complex vocabulary that may imply a high level of ambiguity such as the use of multiple words to refer to the same concept (e.g. “neuron”, “cell”, “brain cell”). Furthermore, the dynamic aspect of the concept-map building task induces an extra level of complexity that may have impaired the REGARD model’s performance. Most often, co-learners talk about the knowledge domain concepts first and build the associated concept-map objects afterwards. Consequently, we can expect significantly fewer references to objects that are already present on the shared interface. The performance of REGARD should be significantly higher in the cases of more static collaborative activities such as discussing graphical contents, where the visual scene is less prompt to constantly evolve.

Finally, one of the main criticism that can be done concerns the type of stimulus that was used. Indeed, in our specific situation, the verbal references are simply the labels used for the objects. While this was of a great help to perform the analyses presented here as it allowed us to detect automatically the verbal references, it makes a limitation on the scope of the results because we have a special situation in which the references are written on the objects. Hence, the process of referring to an object may be mixed with a process of reading the label of the object. It is possible that the effects exposed in this work are partly explained by this reading aspect. However, considering the previous works in this area (Griffin and Bock 2000; Griffin 2001) which have been done with pure image stimuli, it is very likely that these phenomena are mainly due to the referencing process.

## 5.6 Conclusion and Future Works

This work brings two contributions on both the theoretical and the practical sides. First, on the theoretical aspect, we have replicated important results about the time

intercourse of gaze and speech in a complex ecologically valid collaborative situation. This is an important finding as it shows the robustness and the systematic aspect of such phenomena. Moreover, it also shows that these effects are not artifacts due to the specific highly-controlled settings used in previous works. Secondly, on a more practical side, we have developed an algorithm that is able to detect verbal references, as well as their referenced object, from the dialogue and the gazes of two individuals working or discussing about some shared workspace. This model is a great step towards the design of intelligent multi-modal interface that would take advantage of other modalities in a natural way, i.e. without forcing the user to adapt to the computer.

The potential applications that would benefit from such a model are various. For example, this could be used to automatically annotate images viewed by collaborators. This could provide interesting outputs or feedbacks for the collaborators or for other subsequent automatic post-processing. Alternatively, REGARD could part of a more complex system that could use the discovered references in order to infer semantics from the dialogue of the partners. Indeed, the system could use the relations between words and objects it found coupled with a more general model of semantics in order to interpret the meanings of what the collaborators say so that it could take appropriate actions. Finally, REGARD could also serve as a grounding detector, by identifying which terms are used as verbal references in a coherent way by both the speaker and the listener. Such information can be useful to provide intelligent feedback or user-interface adaptation from the collaborators.

These results open several interesting avenues for future work. First, on the theoretical side, it would be worth to replicate these results in different, while still ecologically valid, settings. Specifically, it would be interesting to make similar analyses with a stimulus with none-labeled object, i.e. that have implicit names instead of being labeled. This would allow us to avoid effects due to reading the labels. Another interesting point would also to vary the type of referents. For example, we can have a situation with usual easy-to-label objects such as cars or animals, compared with a situation with more abstracts objects that may resemble to several possible known objects (such as Tangram). This could inform us on the possibility of detecting common-ground with REGARD. Concerning the algorithm itself, the main directions to follow consist in making it more general by overcoming the two main limitations described above, namely the need of the list of all objects and the need of transcribed speech. In this respect, the best potential we can see come from the proposed solutions, namely to deal with fuzzy regions instead of objects and to compare sounds instead of comparing written words.

**Acknowledgements** This work was funded by the Swiss National Science Foundation (grant #K-12K1-117909).

## Appendix

In order to allow the reader to get a general idea of the type of dialogue that occurred during the task, Tables 5.3 and 5.4 show two translated excerpts from two different

**Table 5.3** Translated excerpt from the verbal interaction of a dyad (originally in French)

Subject	Duration [s]	Utterance
A	3.3	Hence we can add under “axon”, “myelin” for instance.
A	2.2	With the “Ranvier’s nodes”.
A	27.6	We put “myelin” like this. Wait I’ll do it.
B	2.6	Yes, then we put « have » everywhere.
A	2.8	It’s an idea. I added “Jumps” for the. . .
B	1.1	“Jumps”? Why?
A	8.8	Yeah because these are « jumps » . . . huh. . . hum! . . . I don’t know. . .
B	11.8	Ha but. . . yeah. . . between “electric potential” and “axon” I’d put “electric potential IN the axons”.
A	2.6	Yeah! It’s not bad.
B	7.0	Moreover the “membranar potential”, yes between “electric potential” and “membranar potential” we should change. Remove “have”.
A	7.6	We can put “more precisely”.

**Table 5.4** Translated excerpt from the verbal interaction of a dyad (originally in French)

Subject	Duration [s]	Utterance
A	7.6	Afterwards, there is the current that passes, it is. . . it goes. . . I forgot where exactly. Anox. . . something like that.
A	10.9	Wait! Anox? Anobsine. I forgot. Onoxine something like this. It goes there.
B	4.2	Huh! Ok then just add what you’re thinking about.
A	7.9	I can’t really manage to. . .
A	17.7	We’ll put it with “passing current”. It is better.
B	15.3	It is becoming pretty nice.
A	3.2	At the end of the synapse.
B	8.3	Huh! Yes it was. . . end of the synapse and so forth. And you add what you think of, concerning the term. I don’t remember the term.
A	3.3	There are the chemical reactions.
B	4.1	But I don’t remember the term, I didn’t retain all the terms.
A	0.6	Ok no problem.

dyads. The references to objects of the map have been put in quotation marks. The first excerpt (see Table 5.3) shows a dialogue centered around the objects that are drawn on the map while the second excerpt (see Table 5.4) is more conceptual with few explicit references to the objects of the concept-map.



## References

- Allopenna PD, Magnuson JS, Tanenhaus MK (1998) Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J Mem Lang* 38(4):419–439. doi:[10.1006/jmla.1997.2558](https://doi.org/10.1006/jmla.1997.2558)
- Cherubini M, Nüssli M-A, Dillenbourg P (2008) Deixis and gaze in collaborative work at a distance: a computational model to detect misunderstandings. In: Proceedings of the 2008 symposium on eye tracking research and applications (ETRA'08). ACM, New York, pp 173–180
- Deléglise P, Estève Y, Meignier S, Merlin T (2005) The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news. In: Interspeech 2005
- Griffin ZM (2001) Gaze durations during speech reflect word selection and phonological encoding. *Cognition* 82(1):B1–B14
- Griffin ZM, Bock K (2000) What the eyes say about speaking. *Psychol Sci* 11(4):274–279
- Griffin ZM, Oppenheimer DM (2006) Speakers gaze at objects while preparing intentionally inaccurate labels for them. *J Exp Psychol Learn Mem Cogn* 32(4):943–948. doi:[10.1037/0278-7393.32.4.943](https://doi.org/10.1037/0278-7393.32.4.943)
- Meyer AS, Sleiderink AM, Levelt WJM (1998) Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66(2):B25–B33. doi:[10.1016/S0010-0277\(98\)00009-2](https://doi.org/10.1016/S0010-0277(98)00009-2)
- Nüssli M-A (2011) Dual eye-tracking methods for the study of remote collaborative problem solving. PhD thesis, École Polytechnique Fédérale de Lausanne
- Richardson DC, Dale R (2005) Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn Sci* 29(29):1045–1060
- Sangin M (2009) Peer knowledge modeling in computer supported collaborative learning. PhD thesis, École Polytechnique Fédérale de Lausanne
- Sangin M, Molinari G, Nüssli M-A, Dillenbourg P (2008) How learners use awareness cues about their peer's knowledge: insights from synchronized eye-tracking data. In: ICLS, vol 2, pp 287–296
- Sangin M, Molinari G, Nüssli M-A, Dillenbourg P (2011) Facilitating peer knowledge modeling: effects of a knowledge awareness tool on collaborative learning outcomes and processes. *Comput Hum Behav* 27(3):1059–1067. doi:[10.1016/j.chb.2010.05.032](https://doi.org/10.1016/j.chb.2010.05.032)
- Zelinsky GJ, Murphy GL (2000) Synchronizing visual and language processing: an effect of object name length on eye movements. *Psychol Sci* 11(2):125–131

# Chapter 6

## Effectiveness of Gaze-Based Engagement Estimation in Conversational Agents

Ryo Ishii, Ryota Ooko, Yukiko I. Nakano, and Tokoaki Nishida

**Abstract** In face-to-face conversations, speakers monitor the listener's gaze to check whether the listener is engaged in the conversation. The speaker may change the conversational strategy if the listener is not fully engaged in the conversation. In this chapter, we propose an algorithm to estimate the user's conversational engagement based on various types of gaze information, such as gaze shift patterns, gaze duration, amount of eye movement, and pupil size. By applying the proposed algorithm, we implement an agent that can change its conversational strategy according to the user's conversational engagement. We also evaluate the agent system by investigating how the agent's awareness of the user's engagement affects the user's verbal and nonverbal behaviors as well as the subjective impressions of the agent. First, based on an empirical study, we identify useful information for estimating user engagement, and establish an engagement estimation model using a decision tree technique. The model can predict the user's disengagement with an accuracy of over 70 %. Then, the model is implemented as a real-time engagement-judgment mechanism and is incorporated into a multimodal dialogue manager in a conversational agent. Finally, our evaluation experiment reveals that probing questions by the engagement-sensitive agent successfully recover the subject's conversational engagement, change the gaze behaviors of the subject, and elicit more verbal

---

R. Ishii · T. Nishida  
Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan

R. Ishii  
e-mail: [ishii.ryo@lab.ntt.co.jp](mailto:ishii.ryo@lab.ntt.co.jp)

T. Nishida  
e-mail: [nishida@i.kyoto-u.ac.jp](mailto:nishida@i.kyoto-u.ac.jp)

R. Ishii · R. Ooko · Y.I. Nakano (✉)  
Department of Computer and Information Science, Seikei University, Tokyo, Japan  
e-mail: [y.nakano@st.seikei.ac.jp](mailto:y.nakano@st.seikei.ac.jp)

R. Ooko  
e-mail: [dm116212@cc.seikei.ac.jp](mailto:dm116212@cc.seikei.ac.jp)

R. Ishii  
NTT Communication Science Laboratories, NTT Corporation, Kanagawa, Japan

contribution. Moreover, such timely probing questions also improve the subject's impression of the agent.

## 6.1 Introduction

In conversation, nonverbal behaviors work in a complementary manner with verbal behaviors to convey the meaning of utterances. One of the essential aspects to realizing such a process is the conversation participation attitude, or engagement. According to the definition by Sidner et al. (2004), engagement is the process by which two (or more) participants establish, maintain, and end their perceived connection. For example, in order to maintain the communication, speakers continue to check whether the listener positively participates in the conversation, while paying attention to the speaker. On the other hand, listeners demonstrate their participation attitudes through nonverbal behaviors, such as nodding and eye gaze, and express their desire to continue communication (Argyle et al. 1973).

Therefore, in order to establish better communication partnership between human users and conversational humanoids, the system should be able to recognize such nonverbal signals and estimate the user's engagement state. Automatically judging whether the user is actively participating in the conversation with the system may be useful in adaptively determining the system behavior. For example, when the system detects the user's disengagement from the conversation, the system should encourage the user's active participation in a conversation or change the topic. However, few studies have examined automatic interpretation of the user's attitudes based on the gaze information, even though off-the-shelf eye tracking technologies are sufficient to recognize the user's eye gaze.

Thus, in an attempt to improve the smoothness of communication between the user and the agent, the present study proposes an engagement estimation method. The proposed method is implemented in a conversational agent, and the autonomous agent system is evaluated. In particular, we focus on the user's gaze behaviors while communicating with virtual agents, and address the following issues:

- (1) Identification of gaze behaviors that are distinctively observed when users are disengaged from the conversation based on the analysis of the correlation between various types of gaze information and the human judgment of engagement.
- (2) Based on empirical results, we propose an engagement estimation method, develop a mechanism that can detect user disengagement in real time based on gaze information, and implement an autonomous conversational agent by incorporating the engagement estimation mechanism.
- (3) An evaluation experiment to show how probing questions by the engagement-sensitive agent affect the subject's gaze behaviors and verbal behaviors and improve the impression of the agent is performed.

In the next section, we will describe how this study is related to previous research. As such, we analyze our corpus and propose an engagement estimation model based

on the analysis. After describing the implementation of the system, an evaluation experiment is conducted. Finally, we discuss the results of the evaluation experiment and areas for future research.

## 6.2 Related Research

In communication science and psychology, a number of studies have investigated functions of eye-gaze in face-to-face communication. Kendon (1967) observed eye-gaze behaviors using the ethnomethodological method and discussed various types of eye-gaze functions. Psychological studies reported that eye gazing, specifically accompanied by head nods, serves as positive feedback to the speaker, and demonstrates that the listener is paying attention to the conversation (Clark 1996; Argyle and Cook 1976). This type of mutual gaze also contributes to smooth turn-taking (Novick et al. 1996). In contrast, when conversational participants share the same physical environment and their task requires complex reference to, and joint manipulation of physical objects, joint attention between the participants is a positive signal of conversational engagement (Argyle and Graham 1977; Anderson et al. 1997; Whittaker 2003).

Although in previous studies on interactive systems, the main application of head/eye tracking technology is to estimate the user's interest (Iqbal and Bailey 2004; Qvarfordt and Zhai 2005; Iqbal et al. 2005; Eichner et al. 2007; Nakano and Nishida 2007), some studies have been more directly related to sensing communicative signals displayed by gaze, which contribute to interaction management between the user and the agent. Nakano et al. (2003) proposed a gaze model for nonverbal grounding in conversational agents and used a head tracker to implement an agent that can judge whether the information provided by the agent is grounded.

More recently, Bohus and Horvitz proposed a method of predicting the user's engagement intention in multiparty situations using a head tracker (Bohus and Horvitz 2009). They focused on predicting whether the user will be engaged in the conversation, but not on judging whether the user is engaged in the ongoing conversation, to maintain the communication. In human-robot interaction, Morency et al. (2007) and Rich et al. (2010) used a head tracker to recognize a user's gaze direction and head nods, and exploited the recognized user's behaviors in order to judge whether the user is engaged in conversation with the robot.

The results of these studies suggest that off-the-shelf eye tracking systems are sufficiently accurate and stable to be used in complex agent systems. They also suggest that the user's gaze direction can be roughly estimated from the head direction, as measured by a head tracker. Thus, we believe that combining these sensing technologies with a dialogue management mechanism will enable conversational agents to become more sensitive to the user's conversational engagement.

Based on the above research, the present study attempts to build an information-providing agent that explains products on a display as a virtual salesperson, which requires that the agent accurately sense the user's attentional behavior. In order to achieve this goal, we use an eye (pupil) tracker to measure gaze information more accurately and estimate the user's engagement during conversation with the agent.

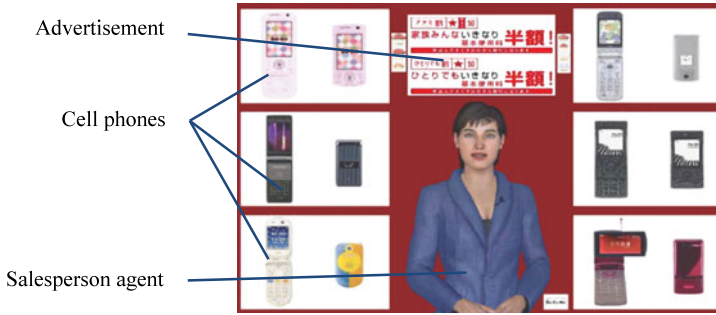


Fig. 6.1 Agent screen shot

### 6.3 Corpus Collection and Analysis

With the goal of establishing an engagement estimation model, first, we analyze gaze data to identify useful parameters for estimating conversational engagement. Since it was reported that gaze shifting patterns and gaze duration correlate with user interest (Qvarfordt and Zhai 2005), we herein investigate whether these two types of information are also useful in estimating the user's engagement in a conversation. If the user is not engaged in the conversation, her/his gaze may move significantly because of distractions from the conversation. In contrast, if the user is positively engaged in the conversation, the gaze remains fixed on the same object for a longer time because the user carefully looks at the object. Moreover, it is widely known that pupil size becomes larger when people are looking at something interesting or exciting (Hess 1965). Therefore, it is assumed that the pupil size may become larger when the user is engaged in the conversation. In contrast, if the user is not engaged, the pupil size may become smaller because the user is looking at the object without substantial interest.

By investigating the correlation between these various types of eye gaze information and human judgment of engagement, we will determine gaze parameters for generating an estimation model for the user's conversational engagement.

#### 6.3.1 Corpus Collection

##### 6.3.1.1 Wizard-of-Oz Experiment in Human-Agent Conversation

In order to collect eye gaze data in human agent interaction, we conduct a Wizard-of-Oz experiment, in which a female animated character is displayed on a 120-inch rear-projection screen (Fig. 6.1). The female animated character acts as a salesperson at a mobile phone store. In the experiment, a subject acts as a user (hereinafter referred to as the "user"). The user listens to the agent's explanation of six cell phones, each of which lasts approximately three to five minutes. The entire process

**Fig. 6.2** Snapshot of video for annotation



takes approximately 20 minutes if the user listens to all of the explanations (109 utterances, each having an average duration of 10 sec). The user was allowed to ask questions about cell phone functions and other yes-no questions and was allowed to request to change the topic to the next cell phone. Since the purpose of the data collection experiment is to collect the typical behaviors of a subject when the subject is disengaged, the agent's explanations were required to be less interesting, or boring, to the subjects. Therefore, we created agent explanations with longer utterances and less attention-grabbing nonverbal behavior animations. When the wizard (an experimenter) produced the agent's utterances by operating a GUI, the agent looked at the target cell phone most of the time and looked at the subject every 10 utterances. The agent repeated this explanation style for the entire session in order to provide the user with boring animation contents.

### 6.3.1.2 Collected Corpus

We collected 10 conversations from 10 subjects. The average length of the conversations was 16 minutes. We created a multimodal corpus containing the following verbal and nonverbal data:

- *Verbal data*: The user's speech was transcribed from the recorded speech, and the agent's utterances were extracted from the log of the Wizard-of-Oz system. The total number of utterances of the agent was 951, and that of the user was 61.
- *Nonverbal data*: The agent's gestures and gaze behaviors were extracted from the Wizard-of-Oz system log. We collected the user's gaze data using a Tobii X50 eye-tracker.
- *Human judgment of engagement*: Another 10 people were recruited as video annotators. They were asked to watch the video of the subjects in the Wizard-of-Oz experiment and to mark the times at which the subject on the video looked disengaged from the conversation. Figure 6.2 shows a snapshot of a video viewed by the annotators. In order to see the subject's gaze and facial expressions, the annotators watched videos that captured the front face of the subjects. In addition, the agent's animation synchronized with the subject's video was also shown to the

annotators. We used the Anvil video annotation tool to annotate the video (Kipp 2001).

Since the video data was recorded at 30 fps, the annotation results were discretized into 1/30-sec time frames. A disengagement score of 0 to 10 was assigned to each time frame by counting how many of the 10 annotators marked the frame as disengagement. For example, if none of the annotators marked the frame as disengagement, the score is 0. This suggests that the subject may be fully engaged in the conversation. Through this process, a total of 246,338 disengagement score data for 1/30-sec time frame were collected. We believe that collecting engagement judgments from 10 annotators provides an evaluation criterion that is more reliable and stable and that provides better ground truth.

### 6.3.2 Corpus Analysis

In this section, we analyze various types of gaze data to determine the engagement estimation parameters.

#### 6.3.2.1 Analysis of Gaze Transition Patterns

It has been found that, in estimating user interest with respect to visual stimulus, gaze transition pattern is more useful than gaze state (Qvarfordt and Zhai 2005). Therefore, we assume that looking at gaze transition is also useful in estimating conversational engagement. In order to analyze the gaze transition patterns, we created gaze direction transition 3-grams using the following labels:

- T: Looking at the target object of the agent’s explanation, which is the discourse focus.
- M: Mutual gaze, in which the subject establishes eye contact with the agent.
- AH: Looking at the agent (non-mutual gaze), in which the agent gazes away from the subject when the subject gazes at the agent.
- F: Looking at non-target objects, such as other cell phones or an advertisement poster ( $F1 \neq F2 \neq F3$ ), where F1, F2, and F3 indicate different objects. For example, while the agent is explaining Cell phone A, the user is looking at non-target Cell phone B, followed by non-target Cell phone E, and then looking again at non-target Cell phone B. In such a case, the gaze transition of the user is indicated as F1-F2-F1.

As described in Sect. 6.3.1.1, the agent looked at the user every 10 utterances in order to show simple repeated behaviors. If the subject is looking at the agent at that time, it is presumed that a mutual gaze is established between them, and the label M is assigned to the corresponding time frames. During the remainder of the time, the agent is looking away from the subject in the general direction of the target

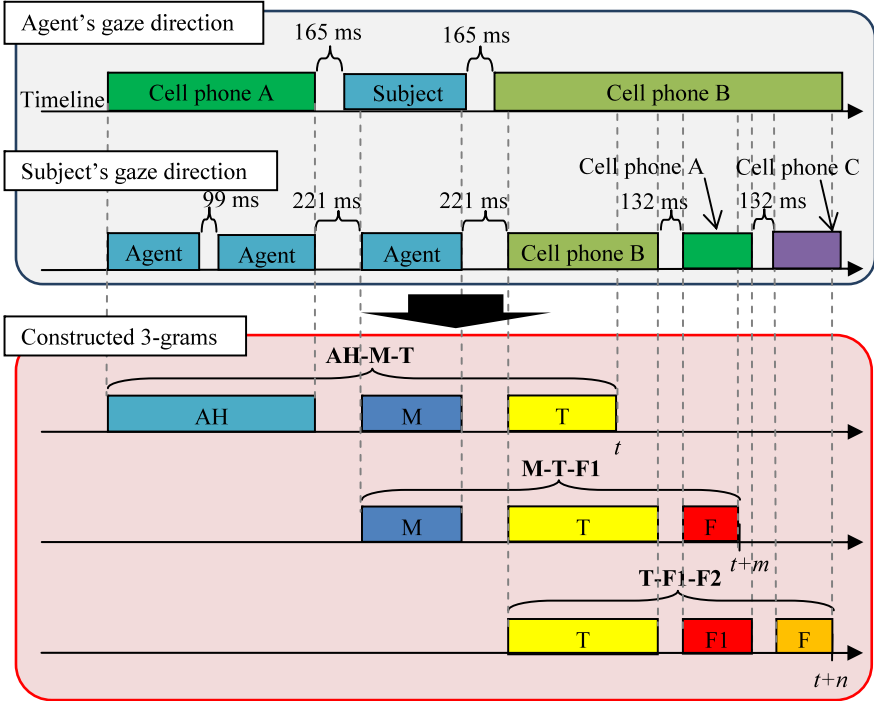


Fig. 6.3 3-gram construction

cell phone, and it is presumed that joint attention is established between the subject and the agent when the subject’s gaze label is T.

Figure 6.3 shows how to construct a gaze transition 3-gram. Since the eye-tracker fails to measure the pupils’ movement during blinks, small blanks often occur in gaze data. For this reason, we counted two consecutive gaze data as a single continuous eye gaze if the same object was continuously looked at in both data and the data blank was less than 200 ms in duration.

For example, as shown in Fig. 6.3, suppose that the agent’s gaze direction shifts as follows: Cell phone A-(165-ms blank)-Subject-(165-ms blank)-Cell phone B. The subject’s gaze shifts are shown in the second line. In this example, the gaze-3-gram is AH-M-T at time  $t$ . Since the first two blocks are labeled AH and the blank between these blocks is 99 ms, the first two blocks are merged. Then, the next block is relabeled M, because this block overlaps with the agent’s looking at the subject. The third block (originally, the fourth block) is labeled as T because the subject is looking at Cell phone B, which is the target object that the agent is explaining about. At time  $t + m$ , the subject is looking at Cell phone A, which is not the target object. Thus, the third block for constructing a 3-gram is labeled F1, and by combining F1 with the last two blocks, the M-T-F1 3-gram is assigned to this time frame. Likewise, at time  $t + n$ , the current gaze is F (Cell phone C) and



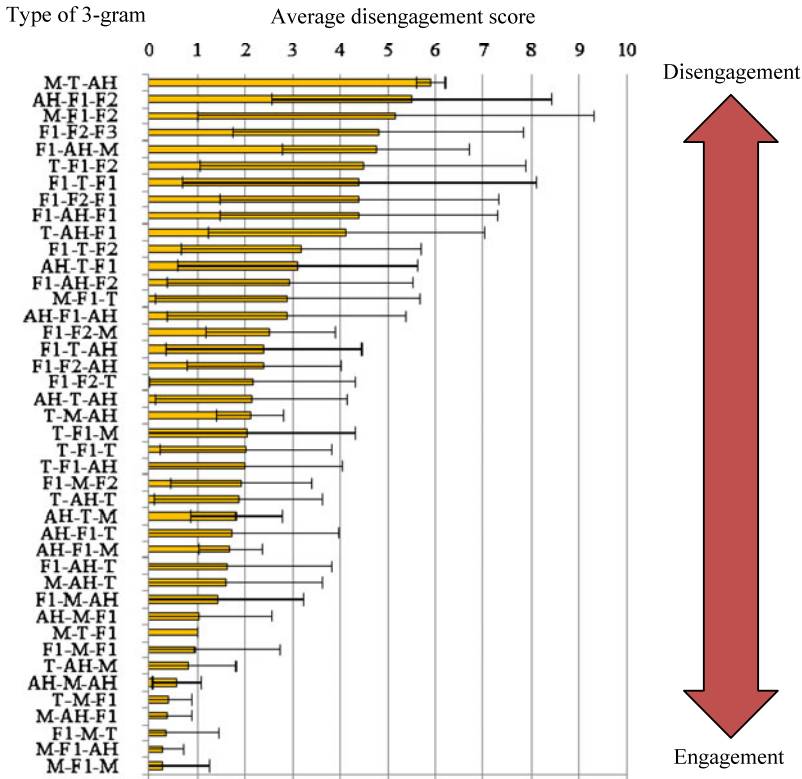


Fig. 6.4 Relationship between eye-gaze 3-grams and disengagement score

the last two gaze blocks are T and F (Cell phone A), respectively. Consequently, the T-F1-F2 gaze 3-gram is assigned to this time frame.

If the gaze blank between two gaze blocks is longer than 1 sec, then we ignore such a sequence as an incomplete 3-gram and start a new 3-gram from the next gaze data. In our corpus, most of the gaze data can be used to construct 3-grams, and incomplete 3-grams are rare. We obtained a total of 140,819 1/30-sec data points.

Using the 3-grams described above, we investigate the correlation between the 3-gram type and the disengagement score for each 1/30 time frame. The average disengagement score was calculated for each 3-gram pattern. For example, 1,085 data points were assigned to the AH-F1-F2 3-gram, and the average disengagement score for these data points was 5.50. The average disengagement scores for all 3-gram types are shown in Fig. 6.4. The x-axis shows the average disengagement score. The first quartile and the third quartile are also shown on the bar chart. The y-axis shows the 3-gram types. The average disengagement score differs considerably depending on the 3-gram type. The 3-gram with the highest score is M-T-AH, the average score of which is 5.9. The 3-gram with the lowest score is M-F1-M, the average score of which is 0.26. These results suggest that 3-grams with higher

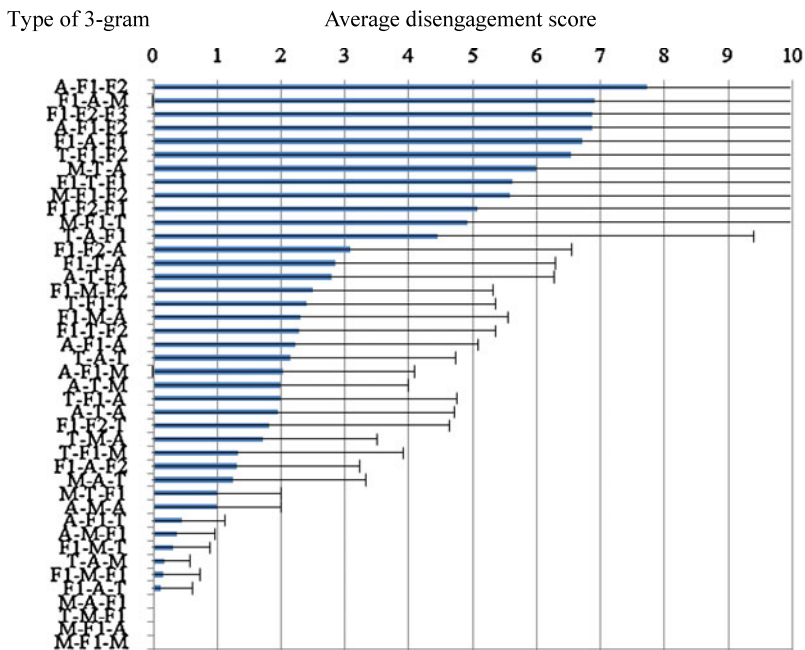


Fig. 6.5 Eye-gaze 3-gram and disengagement score for a long duration

scores frequently co-occurred with the subject’s disengagement states. Therefore, the average disengagement score for each 3-gram may be a useful parameter for estimating conversational engagement.

### 6.3.2.2 Analysis of Gaze Duration

In this section, focusing on gaze duration, we analyze the correlation between 3-gram duration and user engagement. Specifically, we subcategorize each 3-gram type into three sub-types, according to the total duration of all constituents of the 3-gram. The thresholds for subcategorization were determined based on the average ( $\mu$ ) and the standard deviation ( $\sigma$ ) for the cumulative duration of all three constituents:

- Long duration:  $t \geq \mu + \sigma/2$
- Middle duration:  $\mu + \sigma/2 > t \geq \mu - \sigma/2$
- Short duration:  $t < \mu - \sigma/2$

Here,  $t$  is the time duration from the start of the first constituent of the 3-gram to the current time. If  $t$  is greater than  $\mu + \sigma/2$ , then the 3-gram for the data point at time  $t$  is “long”. We calculated the average disengagement scores for each subcategory. Figure 6.5 shows the averages for the long-duration categories. The  $x$ -axis shows the average disengagement score, and the  $y$ -axis shows the 3-gram

type. For example, in the M-F1-T 3-gram, the average duration ( $\mu$ ) was 7.85 sec, and  $\sigma/2$  was 2.03 sec. In subcategorizing this 3-gram, the average disengagement scores for the long-duration category ( $t \geq 9.88$  (sec)), moderate-duration category ( $7.85 > t \geq 5.82$  (sec)), and the short-duration category ( $5.82 \geq t$  (sec)) were 4.93, 3.47, and 0.86, respectively. Since the average disengagement score in the original M-F1-T 3-gram was 2.89, the M-F1-T 3-gram with a long duration has a much higher disengagement score than the original M-F1-T 3-gram. Thus, we expect that by taking the 3-gram duration into consideration, the correlation between gaze 3-grams and disengagement score will be further clarified.

### 6.3.2.3 Analysis of the Amount of Eye Movement

It may be assumed that if the subject is not engaged in the conversation, his/her gaze moves significantly because he/she more frequently looks around at other objects, such as advertisements or cell phones that are not currently being explained. We herein investigate whether the amount of eye movement increases when the subject is not engaged in the conversation.

We calculated the moving average of a 400-ms window for the eye movement. For instance, when the user was engaged (the average disengagement score is 2.38), the amount of eye movement was small (the distance ranged from 9.01 to 10.0 pixels). In contrast, the amount of eye movement increased (the distance ranged from 57.01 to 58.0 pixels) when the subject was fully distracted (average disengagement score: 4.0). The correlation coefficient between the disengagement score and the amount of eye movement was 0.76, which is very high. Thus, we expect that the amount of eye movement may be a useful parameter for estimating conversational engagement.

### 6.3.2.4 Analysis of Pupil Size

Pupil size is known to increase when people are looking at something interesting and exciting. In order to examine whether pupil size is also useful for estimating engagement, we analyzed the correlation between pupil size and user conversational engagement. The correlation coefficient between the disengagement score and the pupil size was  $-0.77$ . As the disengagement score increases (i.e., the more the subject is disengaged), the pupil size decreases. Therefore, the pupil size may be a useful parameter for estimating the conversational engagement.

## 6.4 Engagement Estimation Model

The analysis results described in the previous sections indicate that 3-grams, eye-gaze duration, eye movement distance, and pupil size may be useful predictors of a user's engagement in a conversation. In this section, using a decision tree technique, we establish an engagement estimation model using these parameters.

**Table 6.1** Disengagement score analysis

	Score								
	1	2	3	4	5	6	7	8	9
Average max. score	3.34	3.90	5.27	5.89	6.47	7.47	8.12	8.80	9.39
Difference	2.34	1.90	2.27	1.89	1.47	1.47	1.12	0.80	0.39

### 6.4.1 Training Data

Following the analysis in Sect. 6.3, we use the data for 1/30-sec time frames as one case of training data and apply decision tree learning. Each case consists of five eye-gaze feature values and the user’s conversational engagement state (engaged/disengaged) as the supervising feature. We set a threshold for the disengagement score in order to determine the supervising feature value, i.e., engaged or disengaged. Since the disengagement score is the number of annotators marking a given time frame as a disengagement state, it was assumed that the start time of disengagement marking differed depending on the annotator. Therefore, the disengagement score shifts up and down, and in some cases, the peak value was very low, e.g., 1 or 2. In other cases, the peak value was very high. Table 6.1 shows the average peak values for each score. For instance, the average peak value for shifts including score 2 was 3.90, whereas for score shifts including score 3, the average peak value was 5.27. Therefore, when the disengagement score was 3 at any given time, this movement had a higher possibility of reaching higher scores, as compared to cases with score 2. Since the difference in peaks between shifts with score 2 and shifts with score 3 was the largest (2.27) and the average peaks for shifts with score 3 was over 5 (i.e., more than half of the annotators made a “disengagement” judgment), we set the disengagement threshold to 3. In order to enable a binary judgment (engaged or disengaged), we assigned supervising label as follows:

- *Engaged*:  $0 \leq \text{disengagement score} \leq 2$
- *Disengaged*:  $3 \leq \text{disengagement score} \leq 10$

Since the disengagement score for each time frame is an integer value, the threshold is set between 2 and 3. By applying this threshold, we obtained 82,703 engagement cases and 42,500 disengagement cases.

Based on the analysis in Sect. 6.3, we used the following four features in our estimation model:

- *3-gram*: The current gaze transition 3-gram, as shown in Fig. 6.3. The feature value is a 3-gram label specified by the current gaze data and the last two gaze data.
- *Duration of 3-gram*: In addition to the duration from start to finish of the current 3-gram, the duration of each constituent was also used as a feature in decision tree learning.
- *Eye movement distance*: The amount of eye movement for the last 400 ms.
- *Pupil size*: The average pupil size of both eyes for that time frame.

**Table 6.2** Evaluation results

Model	Result					
	Engagement			Disengagement		
	Precision	Recall	F-measure	Precision	Recall	F-measure
3-gram	0.828	0.940	0.880	0.603	0.299	0.400
3-gram + Dr	0.866	0.803	0.833	0.659	0.722	0.689
3-gram + Ds	0.822	0.925	0.870	0.647	0.436	0.521
3-gram + PS	0.827	0.913	0.868	0.650	0.476	0.549
All	0.880	0.829	0.854	0.685	0.723	0.704

Thus, the training data consists of these four eye-gaze feature values and the user’s conversational engagement state (engaged/disengaged) as supervising data.

### 6.4.2 Tested Models

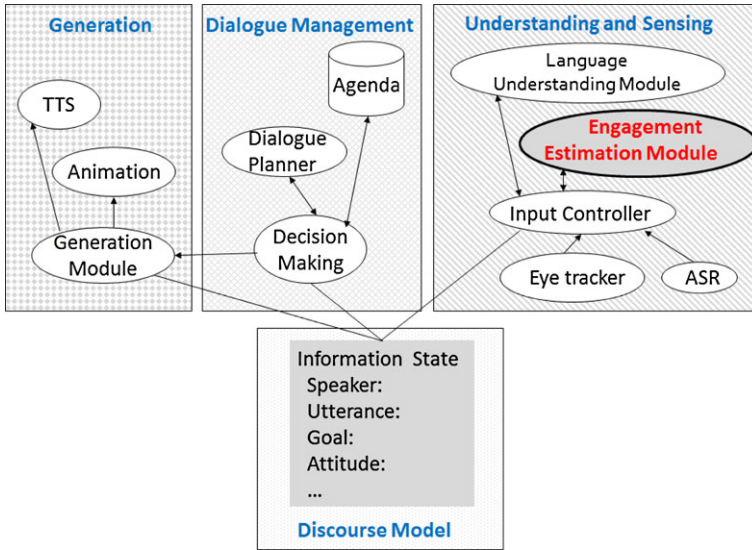
We combined these features to test the following five estimation models:

- *3-gram*: Estimation using only 3-gram labels
- *3-gram + Dr*: Estimation using 3-gram and eye-gaze duration
- *3-gram + Ds*: Estimation using 3-gram and eye movement distance
- *3-gram + PS*: Estimation using 3-gram and pupil size
- *All-parameter model (ALL)*: Estimation using 3-gram, duration, eye movement distance, and pupil size

We employed leave-one-out, 10-fold cross validation. Since we have 10 subjects, we chose nine of the subjects to obtain training data and the remaining subject was used to obtain the test data. This procedure was repeated 10 times, and the average estimation accuracy was calculated.

### 6.4.3 Evaluation of Engagement Estimation Methods

The results of decision tree learning are shown in Table 6.2. We used J48 in the WEKA implementation (Remco et al. 2010). In the overall evaluation, the F-measure of the all-parameter model (ALL) is found to be 0.854 for positive engagement and 0.704 for disengagement, which is the best score among all of the models. This suggests that all of the parameters contribute to estimating the user’s conversational engagement. The performance of the 3-gram + Dr model is much better than that of the 3-gram model. This suggests that gaze duration is a strong predictor of user engagement. Compared to the 3-gram model, the performance of the 3-gram + Ds model and that of the 3-gram + PS model are much better, but are not as good



**Fig. 6.6** System architecture

as that of the 3-gram + Dr model. This suggests that the eye movement distance and the pupil size are useful in estimating conversational engagement, but are not as effective as duration information.

## 6.5 Implementation of an Engagement-Sensitive Conversational Agent

By incorporating the all-parameter model obtained in the previous section into a fully autonomous dialogue system, we create an engagement-sensitive conversational agent. The system architecture is shown in Fig. 6.6, and the primary components are described below.

### 6.5.1 Understanding and Sensing

User's verbal and nonverbal behaviors are sensed and interpreted in the following modules. In addition, we also implemented a simple language understanding.

- **Input Controller:** The input controller receives the recognition results from the julius-4.0.2 speech recognition system (ASR)<sup>1</sup> and eye gaze data from the Tobii X-120 eye tracker. The eye tracker measures the user's gaze points at 50 Hz.

<sup>1</sup>julius-4.0.2. Available from <http://julius.sourceforge.jp/forum/viewtopic.php?f=13&t=53>.

The input controller also obtains the interpretation results from language understanding and engagement estimation and sends the interpretation results to the discourse model.

- *Engagement Estimation Module*: We implemented the engagement estimation method proposed as an engagement estimation module. This module receives eye tracking information through the input controller and uses the gaze information to judge whether the user is engaged in the conversation with the agent. In the current implementation, disengagement judgments are calculated for a 500-ms window. If the system judged that the user is disengaged 40 % of the time, i.e., more than six of the 16 estimations during the 500-ms period, the system judges that the user is disengaged from the conversation. Then, the judgment results are sent back to the input controller to update the dialogue state.

### 6.5.2 *Discourse Model*

The discourse model maintains the state of the dialogue. In the current system, gaze information is updated 50 times per second. On the other hand, verbal information is updated upon each utterance, which is normally several seconds long. In order to keep track of the dialogue state, we use the concept of the information state (IS) (Matheson et al. 2000) and modified the IS to manipulate such heterogeneous verbal and nonverbal information. Subscription and trigger relationships are defined in a configuration file to specify which component subscribes to which information and which information triggers which component. For example, when the gaze information is updated, a message is sent to the engagement estimation module, which processes the message to judge whether the user is engaged in the conversation.

### 6.5.3 *Dialogue Management and Generation*

The decision making module decides the agent's next action by referring to the IS and the agenda. The agenda is implemented as a stack and is updated by the dialogue planner. The dialogue planner receives a user's request for explaining a cell phone as input and generates communicative goals, which are added to the agenda. When the engagement estimation module detects user disengagement and reports this to the IS, the decision making module generates a probing question, such as "Do you have any questions?" or "Would you like to move on to the next cell phone?"

Once the agent's action is determined, multimodal output is produced using TTS software and an animation system. Canned agent's speech is synthesized using Hitachi Hit-Voice TTS and is saved as a .wav file. A sequence of animation commands for each speech is saved as a script file, which is automatically generated by the CAST system (Nakano et al. 2004). Animation scripts consist of a sequence of animation commands along with the time at which the animation should be executed. The animation scripts are interpreted by the Haptrek animation system to generate agent animations according to the specified timing.

## 6.6 Evaluation

In order to examine whether the agent's ability to estimate user engagement affect human-agent interaction, we performed an evaluation experiment.

### 6.6.1 Subjects and Task

Three female and seven male subjects participated in the experiment. The subjects did not participate in the Wizard-of-Oz corpus collection experiment in Sect. 6.3.1.1, and the subjects were not the annotator for disengagement judgment in Sect. 6.3.1.2. The subject's task was the same as in the previous experiment, namely, listening to the agent's explanation and guessing which cell phone models are the most popular among female high school students or businessmen. A list of questions that the subject was allowed to ask (related to price, game functions, slenderness, one-segment broadcasting function, and display size) was displayed in front of the subject.

The subjects wore a headset microphone for speech input. In the experiment, however, the user's speech was interpreted by an experimenter in order to avoid speech recognition errors, which would seriously influence the quality of the interaction.

### 6.6.2 Experimental Design and Hypotheses

If the system successfully detects the subject's disengagement status and notifies that system is aware of their attitude, the subjects may change their verbal and non-verbal behaviors. To examine the effectiveness of the system's awareness of the subject's attitude, we investigate the subject's response to the system's probing questions. For this purpose, we set the following two experimental conditions:

- *Probing based on engagement estimation (engagement estimation condition)*: The agent generates probing questions when the Engagement Estimation Module detects the user's disengagement. This is the proposed system.
- *Periodic probing (periodic probing condition)*: The agent asks probing questions once every 10 utterances. The agent's behaviors are mostly the same as those in the Wizard-of-Oz experiment described in Sect. 6.3.1.1, except for periodically producing probing questions. Since the frequency of probing questions asked by the agent may affect the subject's response, the frequency was determined based on the average probing frequency of the proposed system measured in a preliminary experiment. Thus, the total number of probing questions from the agent was assumed to be approximately equal between these two conditions.

We employed a within-subject design, and each subject experienced both conditions. In order to cancel the order effect, half of the subjects started with the engagement estimation condition and the other half started with the periodic probing



condition. By comparing these two conditions, we test the hypothesis that the engagement estimation condition can probe the subject with a more proper timing than the periodic probing condition. To test this hypothesis, we employ the following measures:

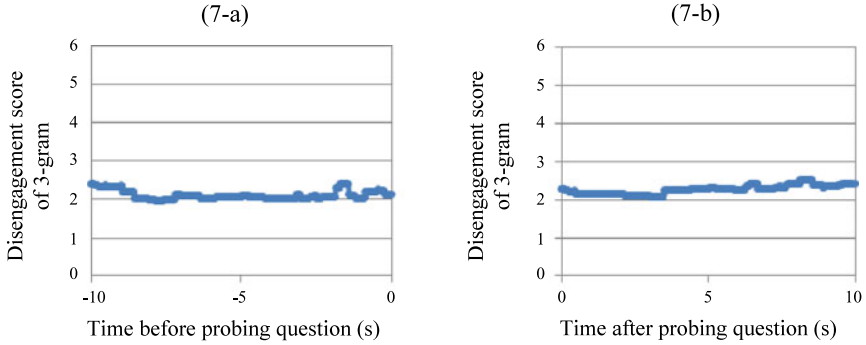
- (a) *Gaze behaviors*: If probing questions are effective in restoring the subject's engagement, the subject's gaze behaviors may be changed after the agent's probe. To test this hypothesis, we measure the gaze shifting pattern represented as 3-grams, eye movement distance, and pupil size in both conditions, and investigate whether these measured values are changed before and after the probing questions. If the results are consistent with what the result of the empirical studies reported in Sect. 6.3.2, then in the engagement estimation condition, 3-grams with higher average disengagement scores are frequently observed before probing and decrease after probing, the eye movement distance is longer before probing and becomes shorter after probing, and the pupil size decreases before probing and recovers after probing. If all of these hypotheses are supported, then the engagement estimation mechanism may work as expected, and, in the engagement estimation condition, the agent's probing questions are generated with the proper timing needed to recover the subject's engagement.
- (b) *Frequency of verbal contributions*: As a verbal measure, the frequency of verbal contributions from the subjects was counted. We assume that the subject is more likely to ask questions or request to change the topic during her/his turn following the agent's probing question if the question is presented with an appropriate timing. Thus, if the subject's verbal contributions become more frequent in the engagement estimation condition than that in the periodic probing condition, this may prove that the probing timing is more appropriate in the engagement estimation condition.
- (c) *Subjective measure*: We used a six-point Likert scale to ask the subjects about their impressions of the agent. The questionnaire contained 33 questions, which were classified into seven categories: awareness of engagement, appropriateness of behaviors, smoothness of conversation, favorability, naturalness of motion, humanness, and intelligence.

## 6.6.3 Results

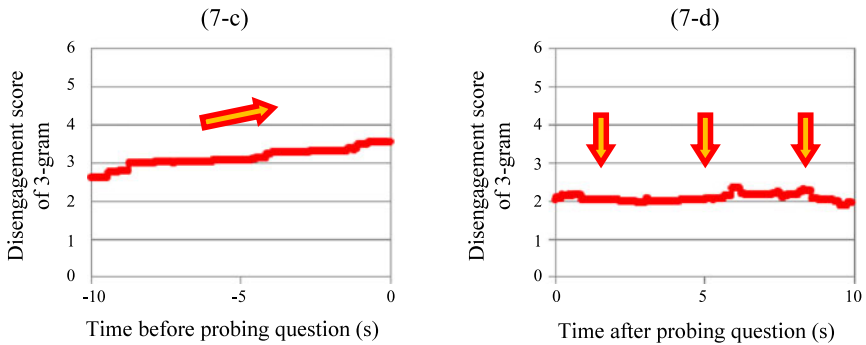
### 6.6.3.1 Gaze Behavior

In order to examine whether the subjects' gaze behaviors were changed after the agent's probing question, we investigated the subjects' gaze behaviors 10 sec before and 10 sec after the agent's probing question, since 10 sec is the average utterance length.

## &lt;Periodic probing condition&gt;



## &lt;Engagement estimation condition&gt;

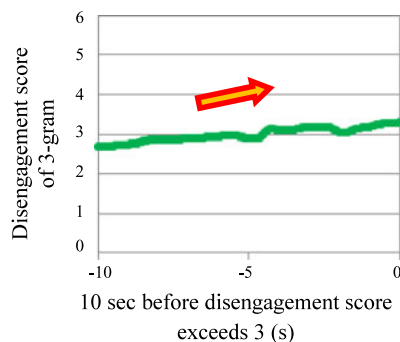


**Fig. 6.7** Changes in disengagement score of 3-grams

**(1) Changes in 3-Grams** We use the average disengagement scores for each 3-gram shown in Fig. 6.4 as the value of each 3-gram. If a 3-gram value is higher, this indicates that the 3-gram is more frequently observed when the user is disengaged. We split each subject's data into 33-ms time intervals and identified 3-grams observed in each time interval and calculated the average 3-gram value in a given time interval. Figure 6.7 plots the averages of 3-gram values for 10 sec before and after probing questions and shows how the 3-gram values are changed over the 20-sec period.

First, we calculated the averages of 3-gram values for 10 sec before probing questions and for 10 sec after the probe. As a result of the  $t$ -test, in the engagement estimation condition, the difference in average 3-gram values for these two time periods was statistically significant. The average for before probing was 3.12 (see Fig. 6.7-c), and that after probing was 2.06 (see Fig. 6.7-d),  $t(9) = 2.03$ ,  $p < 0.05$ . However, the difference was not statistically significant in the periodic probing condition. The average for the period before probing was 2.07 (see Fig. 6.7-a), and that for the period after probing was 2.24 (see Fig. 6.7-b). This suggests that in the engagement estimation condition, disengagement gaze patterns are more frequently observed before probing and such gaze patterns decreased after probing.

**Fig. 6.8** Changes in 3-grams in corpus data



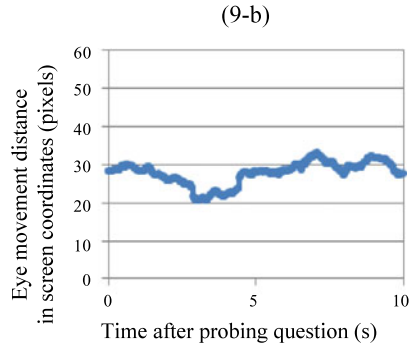
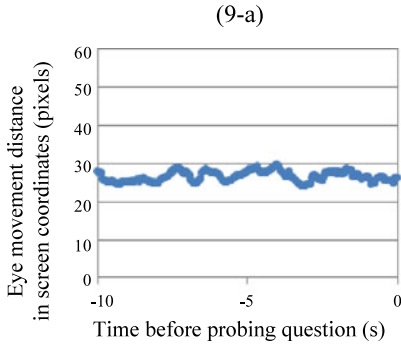
In the periodic probing condition, the average of 3-gram values did not change over time. In contrast, in the engagement estimation condition, the average of 3-gram values increased before probing questions, and was close to 4 just before probing.

We investigated whether a similar trend is found in the corpus analyzed in Sect. 6.3.2. We identified the time at which the disengagement score exceeded 3, and analyzed 3-gram data observed for the last 10-sec period. Figure 6.8 shows how the average 3-gram value changed during this 10-sec period. Note that the slope of the graph is very similar to that shown in Fig. 6.7-c. Thus, this result suggests that the changes in gaze patterns observed in the engagement estimation condition were quite similar to those in the corpus data.

**(2) Changes in Eye Movement Distance** We analyzed the eye movement distance 10 sec before and after probing questions, as shown in Fig. 6.9. The y-axis indicates the average amount of eye movement for each time interval. In the periodic probing condition, the average eye movement amount did not change over time. The average for the period before probing was 26.1 (see Fig. 6.9-a), and that for the period after probing was 27.3 (see Fig. 6.9-b). In contrast, in the engagement estimation condition, the average amount of eye movement before probing questions was larger than that after probing. The average for the period before probing was 29.7 (see Fig. 6.9-c), and that for the period after probing was 25.3 (see Fig. 6.9-d). The result of a paired  $t$ -test was statistically significant ( $t(9) = 6.32, p < 0.05$ ).

We also investigated the corpus analyzed in Sect. 6.3.2, where strong correlation was found between the amount of eye movement and disengagement. We identified the time when the disengagement score exceeded 3 and measured the amount of eye movement for the last 10 sec as shown in Fig. 6.10. As shown in the graph, before probing questions, the amount of eye movement increased over time (the correlation coefficient between time ( $x$ -axis) and the amount of eye movement ( $y$ -axis) was 0.76). At the time when a probing question was introduced, the average eye movement distance exceeded 40 pixels. Thus, we can claim that the results obtained in the corpus data and those in the engagement estimation condition are consistent.

<Periodic probing condition>



<Engagement estimation condition>

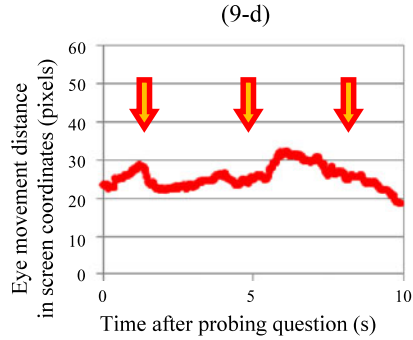
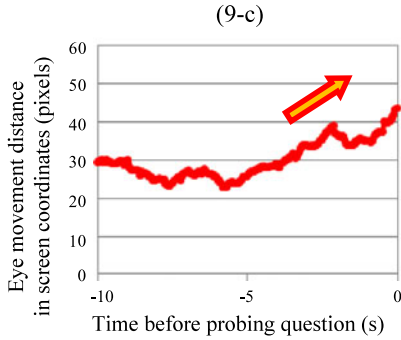
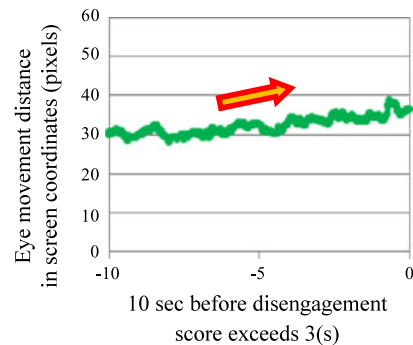


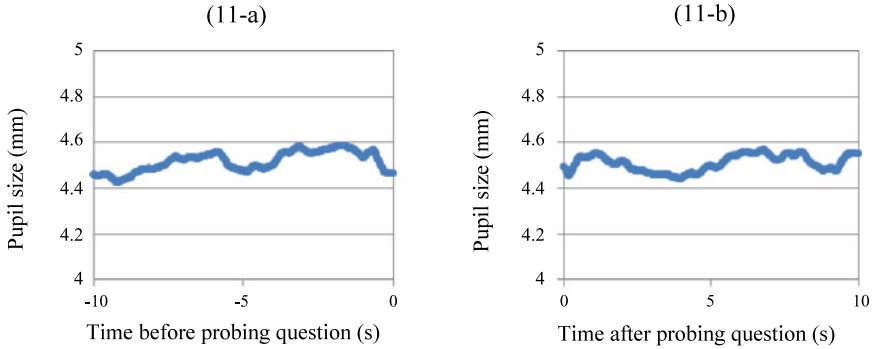
Fig. 6.9 Changes in eye movement distance

Fig. 6.10 Changes in eye movement distance in corpus data

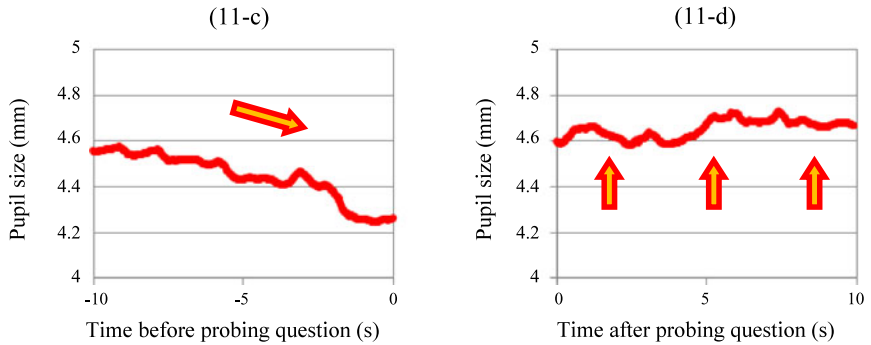


**(3) Changes in Pupil Size** Using a method similar to that described above, we also analyzed the pupil size 10 sec before and 10 sec after probing questions, as shown in Fig. 6.11. The y-axis indicates the average pupil size for each time interval. Under the periodic probing condition, the average pupil size did not change over time. The average for the period before probing was 4.52 (see Fig. 6.11-a), and

## &lt;Periodic probing condition&gt;



## &lt;Engagement estimation condition&gt;

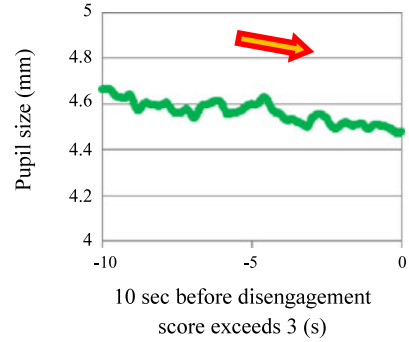


**Fig. 6.11** Changes in pupil size

that for the period after probing was 4.51 (see Fig. 6.11-b). In contrast, under the engagement estimation condition, the average pupil size before probing questions was smaller than that after probing. The average for the period before probing was 4.44 (see Fig. 6.11-c), and that for the period after probing was 4.65 (see Fig. 6.11-d). The result of a paired  $t$ -test was statistically significant ( $t(9) = 1.40$ ,  $p < 0.05$ ). These results suggest that under the engagement estimation condition, the agent produces probing questions with a proper timing and successfully re-captures the subject's interest.

We also investigated the corpus analyzed in Sect. 6.3.2, where we found that a smaller pupil size was strongly correlated with the disengagement. We identified the time at which the disengagement score exceeded 3 and measured the pupil size for the last 10 sec. As shown in Fig. 6.12, the pupil size decreased over time. At the time when a probing question was introduced, the average pupil size was approximately 4.4 mm. The correlation coefficient between time and pupil size was  $-0.94$ , which is very high. Thus, the results obtained under the engagement estimation condition are consistent with the results we found in the corpus analysis.

**Fig. 6.12** Changes in pupil size in corpus data



### 6.6.3.2 Difference in Verbal Behaviors

As another behavioral measure, we investigated the subjects' verbal behaviors. The sample of interaction under both conditions is shown in Fig. 6.13 and Fig. 6.14. We hypothesized that, in the engagement estimation condition, the agent's probing questions presented in a proper timing successfully elicit verbal contributions from subjects, such as asking a question or requesting a change of topic. Therefore, such behaviors are expected to be more frequently observed in the engagement estimation condition than in the periodic probing condition.

As shown in Fig. 6.13, in the periodic probing condition, even though the system detects the subject's disengagement, the agent did not ask probing questions at that time. Therefore, the subject asked a question or requested a change of topic even though he/she had to interrupt the agent. On the other hand, even when the subjects were fully engaged in the conversation, the agent produced probing questions. In such situations, the subjects did not make use of the opportunity to change the topic. In contrast, in the engagement estimation condition, the agent asked probing questions immediately after detecting the subject's disengagement from the conversation. As shown in Fig. 6.14, at this time, the subject took advantage of this opportunity and asked to change the topic.

Figure 6.15 shows the average ratios of (1) subject's asking a question and (2) requesting a change of topic with respect to the total number of the agent's probing questions. In the engagement estimation condition, the subjects asked questions 36.1 % of the time when an opportunity was presented, but in the periodic probing condition they did so only 19.0 % of the time. A statistical trend was observed by means of a two-tailed  $t$ -test ( $t(9) = 1.82, 0.05 < p < 0.1$ ). Similarly, under the engagement estimation condition, the user changed the conversation topic 55.6 % of the time when an opportunity was presented, but did so only 21.4 % of the time under the periodic probing condition. This difference is statistically significant ( $t(9) = 2.42, p < 0.05$ ). These results indicate that in the engagement estimation condition, the subjects are more likely to provide a verbal contribution when the agent poses probing questions, which suggests that the proposed agent system poses probing questions with a proper timing. These results support our hypothesis

- 
- (16) A: This model is only 16.8 mm thick, the thinnest of all the 904i series phones.  
**<The system detects a user's disengagement state>**
- (17) A: The model also has a recently-developed assisting slide that makes the cell phone body smooth.
- (18) A: The first time I picked it up, the slide felt very comfortable to me.
- (19) A: If you are looking for an attractive and comfortable cell phone, I recommend this model.
- (20) A: And, this model has...
- <A subject interrupts the agent's utterance and asks her to change the topic to a different cell phone>**
- (21) S: Could you change the topic and talk about another cell phone?
- (22) A: Yes, let me resume my explanation.
- <The topic is changed after a user's explicit request>**
- (23) A: At right center is the "SIMPURE L2", which is a worldwide model and can be used anywhere in the world.
- (24) A: Its body is only 17.6 mm thick and fits easily into your hand. It is beautifully solid and very usable.
- (25) A: It also features a comfortable slide.
- (26) A: Its main feature is that unlike other phones it can be used in overseas countries.
- (27) A: So when you use it overseas, you can use the same number and email address that you use in Japan.
- <After speaking 10 utterances about the cell phones, the agent asks probing questions>**
- (28) A: Would you like to move on to the next cell phone?
- (29) S: No. I want to hear more about this model.
- (30) A: Ok, I will keep explaining the model.
- <The topic continues>**
- (31) A: When you go overseas, you don't need to worry about getting a new cell phone.
- 

**Fig. 6.13** Periodic probing condition

- 
- (4) A: At your upper right is the "Raku-raku phone IV" model, which is the newest model of the "Raku-raku" phone series and has a GPS feature.
- (5) A: The model is an advanced cell phone that incorporates the concepts of friendliness, security, and usability to make it universally user-friendly.
- (6) A: This model has especially been strengthened with the inclusion of a security function with GPS.
- <The system detects a user's disengagement state and asks probing questions>**
- (10) S: Would you like to move on to the next cell phone?
- (11) A: Yes, please.
- (12) A: Let us resume the explanation.
- <The topic is changed>**
- (11) A: At the left center is the "SH904i", which has a finger-sensitive touch pad and a three-inch wide screen.
- 

**Fig. 6.14** Engagement estimation condition

and serve as evidence that the proposed engagement estimation method is valid and that the implemented estimation mechanism works quite well in a real-time agent system.

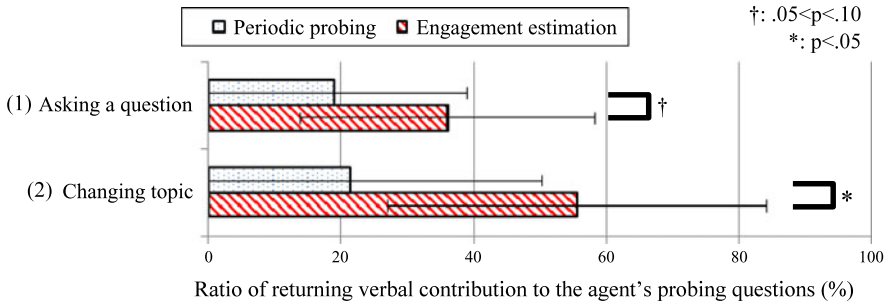


Fig. 6.15 Ratios of returning verbal contribution to the agent's probing questions

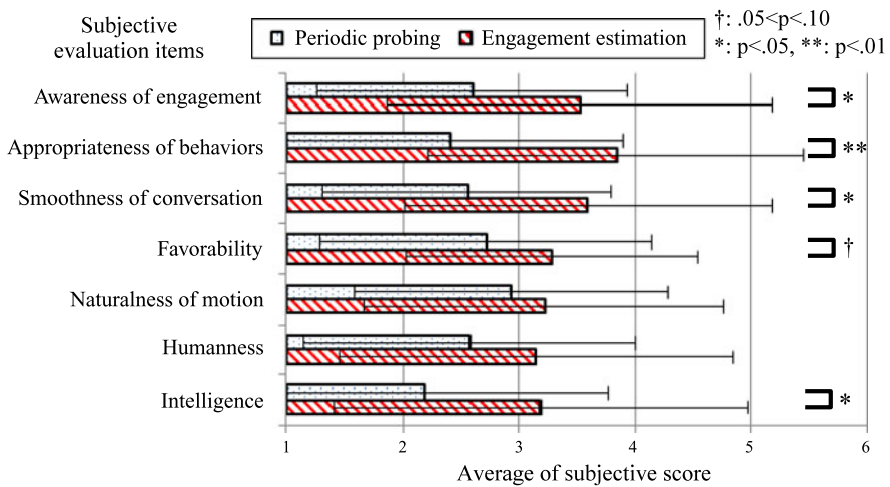


Fig. 6.16 Subjective evaluation results

### 6.6.3.3 Subjective Evaluation

As a subjective evaluation, the averages were calculated for each question category in the questionnaire. As shown in Fig. 6.16, all of the scores were higher in the engagement estimation condition than in the periodic probing condition. For “awareness of engagement”, “appropriateness of behaviors”, “smoothness of conversation”, “favorability”, and “intelligence”, we found a statistical significance or trend in two-tailed *t*-tests ( $t(9) = 2.91, p < 0.05$  for “awareness of engagement”,  $t(9) = 3.28, p < 0.01$  for “appropriateness of behaviors”,  $t(9) = 2.85, p < 0.05$  for “smoothness of conversation”,  $t(9) = 2.13, 0.05 < p < 0.10$  for “favorability”, and  $t(9) = 2.86, p < 0.05$  for “intelligence”).



### 6.6.4 Discussion

In the evaluation experiment, we focused on the effect of the agent's probing questions. Although there are several other ways to re-establish engagement (e.g., attracting user's attention through the use of gestures), the results of the present study clearly indicate that the agent's simple feedback, which only expresses the awareness of engagement, has a significant effect on the subjects' verbal and nonverbal behaviors, as well as their subjective impression.

First, in the engagement estimation condition, disengagement gaze patterns were more frequently observed before probing, and such gaze patterns decreased after probing. This result was supported by the results of analyzing individual gaze behaviors. The probing questions by the engagement-sensitive agent successfully decreased the eye movement distance and widened the pupil size. In contrast, under the periodic probing condition, these values were not very different before and after a probing question. These results suggest that the engagement-sensitive agent generates probing questions with a proper timing and that these questions recover the subjects' engagement. Under the periodic probing condition, the agent is assumed to have asked probing questions even when the subjects were fully engaged in the conversation. Thus, the participation attitudes of the subjects did not change.

More interestingly, we found consistent results between the engagement estimation condition in the evaluation experiment, in which the disengagement judgment was performed using the proposed agent system, and the corpus analysis described in Sect. 6.3.2, in which the disengagement judgment was performed by human annotators. This result suggests that the parameters used in our estimation model were properly selected and that the system's judgment of disengagement is quite similar to judgment by human.

In measuring the verbal behavior, it was found that the subjects have asked questions and changed the topic more frequently when the agent asked probing questions under the engagement estimation condition. This result also supports our hypothesis that, under the engagement estimation condition, probing questions were asked at the right moment according to the user's disengagement state.

For the subjective evaluation, the subjects felt that the proposed system was more aware of the subject's engagement and that the conversation with the agent was smoother. In addition, the subjects felt that the agent's behaviors were appropriate. Since the variation of the agent's probing was the same under both conditions, it is assumed that the subjects felt that the timing of the agent's behaviors was more appropriate under the engagement estimation condition than under the periodic probing condition. Interestingly, even though the agent utterances were completely the same under both conditions, the subjects felt that the proposed agent was smarter than the periodic probing agent. However, we found no difference in the humanness and naturalness of the agent's motion. These findings suggest that the timing of the agent's behaviors affects the subject's impression of the agent's intelligence, but not the naturalness as a human. In summary, the agent's verbal behaviors presented with a proper timing improve the user's impression of the agent's nonverbal expressions and selecting the agent's behaviors according to the results of engagement estimation is effective in human-agent interaction.

## 6.7 Conclusion and Future Work

In this chapter, we first collected a human-agent interaction corpus in a Wizard-of-Oz experiment and then analyzed the corpus concerning gaze 3-gram, gaze duration, amount of eye movement, and pupil size. The analyses revealed that all of these values were correlated with human observational judgment of conversation engagement. Based on these corpus analysis results, we used these factors in decision tree learning and found that the model using all of these factors performed the best. We then incorporated the model into a conversational agent serving as a salesperson. In an evaluation experiment, we compared the proposed model and the periodic probing system and found that our agent system generates probing questions with a proper timing, which demonstrates that the proposed engagement estimation mechanism can judge conversation engagement quite well. The engagement estimation mechanism also works well in a complex conversational agent system in real time and is useful for improving the quality of user-agent interaction.

As future directions for our work, we intend to improve the accuracy of the proposed model for estimating disengagement states in particular, because the F-measure of our current model is still only 0.7. Moreover, eye-tracking is not very robust because gaze data cannot be measured when the user's head moves significantly. Thus, in order to improve engagement judgment robustness, it is necessary to combine gaze information with other nonverbal information, such as facial expressions and head nods, because these behaviors are used as a feedback from listeners. On the other hand, it is also necessary to simplify the model by trying to use other possible gaze patterns, such as uni-grams and bi-grams.

Finally, we need to address issues related to how to select the most appropriate agent actions according to the user's engagement states. In addition to asking probing questions, there may be other possibilities for re-acquiring user engagement, such as asking the user's preference or telling the user to disregard other objects. More basic research is necessary in order to select effective agent actions.

**Acknowledgement** This study was funded in part by JSPS under a Grant-in-Aid for Scientific Research (S) (19100001).

## References

- Anderson AH et al (1997) The effects of face-to-face communication on the intelligibility of speech. *Percept Psychophys* 59:580–592
- Argyle M, Cook M (1976) *Gaze and mutual gaze*. Cambridge University Press, Cambridge
- Argyle M, Graham J (1977) The Central Europe experiment—looking at persons and looking at things. *J Environ Psychol Nonverbal Behav* 1:6–16
- Argyle M et al (1973) The different functions of gaze. *Semiotica* 7:19–32
- Bohus D, Horvitz E (2009) Learning to predict engagement with a spoken dialog system in open-world settings. In: *SIGdial'09*, London, pp 244–252
- Clark HH (1996) *Using language*. Cambridge University Press, Cambridge
- Eichner T et al (2007) Attentive presentation agents. In: *The 7th international conference on intelligent virtual agents (IVA)*, pp 283–295

- Hess EH (1965) Attitude and pupil size. *Sci Am* 212:46–54
- Iqbal ST, Bailey BP (2004) Task-evoked pupillary response to mental workload in human-computer interaction. In: CHI'04, Vienna, pp 1477–1480
- Iqbal ST, Xianjun SZ, Bailey BP (2005) Towards an index of opportunity: understanding changes in mental workload during task execution. In: CHI'05, Portland
- Kendon A (1967) Some functions of gaze direction in social interaction. *Acta Psychol* 26:22–63
- Kipp M (2001) Anvil—a generic annotation tool for multimodal dialogue. In: The 7th European conference on speech communication and technology, pp 1367–1370
- Matheson C, Poesio M, Traum D (2000) Modelling grounding and discourse obligations using update rules. In: 1st annual meeting of the North American chapter of the Association for Computational Linguistics (NAACL2000), pp 1–8
- Morency L-P et al (2007) Head gestures for perceptual interfaces: the role of context in improving recognition. *Artif Intell* 171(8–9):568–585
- Nakano YI, Nishida T (2007) Attentional behaviors as nonverbal communicative signals in situated interactions with conversational agents. In: Nishida T (ed) *Engineering approaches to conversational informatics*. Wiley, New York, pp 85–102
- Nakano YI et al (2003) Towards a model of face-to-face grounding. In: The 41st annual meeting of the Association for Computational Linguistics (ACL03), Sapporo, Japan, pp 553–561
- Nakano YI et al (2004) Converting text into agent animations: assigning gestures to text. In: Human language technology conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Boston, companion volume, pp 91–102
- Novick DG, Hansen B, Ward K (1996) Coordinating turn-taking with gaze. In: *ICSLP-96*, Philadelphia, vol 3, pp 1888–1891
- Qvarfordt P, Zhai S (2005) Conversing with the user based on eye-gaze patterns. In: The conference on human-factors in computing systems (CHI 2005), pp 221–230
- Remco R, Bouckaert EF, Hall MA, Holmes G, Pfahringer B, Reutemann P, Witten IH (2010) WEKA—experiences with a java open-source project. *J Mach Learn Res* 11:2533–2541
- Rich C et al (2010) Recognizing engagement in human-robot interaction. In: *ACM/IEEE international conference on human-robot interaction*, pp 375–382
- Sidner CL et al (2004) Where to look: a study of human-robot engagement. In: *ACM international conference on intelligent user interfaces (IUI)*, pp 78–84
- Whittaker S (2003) Theories and methods in mediated communication. In: Graesser A, Gernsbacher M, Goldman S (eds) *The handbook of discourse processes*. Erlbaum, Hillsdale, pp 243–286

# Chapter 7

## A Computational Approach for Prediction of Problem-Solving Behavior Using Support Vector Machines and Eye-Tracking Data

Roman Bednarik, Shahram Eivazi, and Hana Vrzakova

**Abstract** Inference about high-level cognitive states during interaction is a fundamental task in building proactive intelligent systems that would allow effective offloading of mental operations to a computational architecture. We introduce an improved machine-learning pipeline able to predict user interactive behavior and performance using real-time eye-tracking. The inference is carried out using a support-vector machine (SVM) on a large set of features computed from eye movement data that are linked to concurrent high-level behavioral codes based on think aloud protocols. The differences between cognitive states can be inferred from overt visual attention patterns with accuracy over chance levels, although the overall accuracy is still low. The system can also classify and predict performance of the problem-solving users with up to 79 % accuracy. We suggest this prediction model as a universal approach for understanding of gaze in complex strategic behavior. The findings confirm that eye movement data carry important information about problem solving processes and that proactive systems can benefit from real-time monitoring of visual attention.

### 7.1 Introduction

Effective modeling of human behavior and cognition is one of the primary challenges for building adaptive and proactive systems. Traditional data collection methods, such as interaction logs or verbal protocols, are often not reliable or applicable. For instance, it has been frequently argued that tasks such as reading, mental computations, and problem solving are hard to be assessed by methods such as verbal protocol (Surakka et al. 2003). In this chapter we focus on eye-tracking as a rich source of data for prediction of human cognitive states and actions.

---

R. Bednarik (✉) · S. Eivazi · H. Vrzakova  
University of Eastern Finland, Yliopistokatu 2, P.O. Box 111, 80101 Joensuu, Finland  
e-mail: [roman.bednarik@uef.fi](mailto:roman.bednarik@uef.fi)

S. Eivazi  
e-mail: [shahram.eivazi@uef.fi](mailto:shahram.eivazi@uef.fi)

H. Vrzakova  
e-mail: [hana.vrzakova@uef.fi](mailto:hana.vrzakova@uef.fi)

Modern eye-tracking research tends to rest on the eye-mind hypothesis (Just and Carpenter 1976); eye-tracking data are commonly considered as a measure of overt visual attention and that, according to the hypothesis, is linked to the internal processing. Understanding of the relations between eye movements and human cognition has indeed proven fruitful in many domains, such as reading comprehension, visual search, selective attention, and studies of visual working memory (Kaller et al. 2009).

Eye tracking is thus considered as a technology that allows an unobtrusive, robust and real-time user behavioral data collection. The capture of the ongoing visual and cognitive processes is achieved through registering the eye-movements of users and computational approaches for processing the data. The increased availability of eye-trackers makes eye-tracking also feasible as an input device in gaze-aware interfaces (Hsu et al. 2003). For example, the technology has been applied in eye-typing (Meyer et al. 2003), object pointing and selection (Salvucci 2001), gaming (Smith and Graham 2006), or interaction with problem solving (Bednarik et al. 2009).

Previous research shows that eye movements during interaction with complex visual stimuli are often regular and systematic (Yarbus 1967; Rayner 1998). The existence of detectable and stable patterns in eye-movements motivates researchers in creating of cognitive models of user behavior. For example, expertise differences have frequently been linked to the differences in the eye-movement patterns.

In the domain of user modeling, Loboda and Brusilovsky (2010), Bednarik (2005) and Conati and Merten (2007) argued that eye tracking can be applied for improving the accuracy of prediction models. Loboda and Brusilovsky highlighted the advantages of eye movement data for on-line assessment of user meta-cognitive behavior. Conati and Merten (2007) showed that eye-tracking data improves the performance of probabilistic models in online assessment.

Despite the great potentials, it is not yet well understood, however, how abundant, low-level raw eye-tracking data can be employed for modeling of high-level user internal states. In this chapter we describe the design and components of a system that employs eye-tracking data in an offline manner to model user performance and cognitive activities in an interactive problem solving task. The presented experiments investigate three prediction models aiming to provide a recognition and unambiguous interpretation of eye gaze patterns. We describe the design of the approach to process raw eye-movement data into features that can be fed to the computational models in ways that would allow providing new intelligent user interfaces with behavioral predictions about user strategies and performance.

### ***7.1.1 Related Work***

People apply a range of strategies when they have to make a choice or decision to achieve their goals. Understanding these processes as they occur with interactive interfaces is not an easy task, but at the same time, it is a central research problem to tackle on the way towards more intelligent interactive systems. Understanding users'

plans and goals in real time would enable us to significantly improve the interaction. Therefore, in order to create interfaces that are more sensitive and proactive to user's needs, the user cognitive states must first be invariably recognized.

Ericsson and Simon (1993) assumed that think aloud reports are a reflection of the cognitive processes that generate user's behavior and action. In real-time systems, however, data collection using verbal protocol methods is problematic for several reasons: think aloud utterances are often incoherent (Ericsson and Simon 1993), verbalizing thought is not a natural procedure in everyday situations, many inner processes are unconscious, and the rate of thoughts is typically faster than one is able to articulate. According to van Someren et al. (1994) in many cases it is possible to combine think aloud method with other data collection methods, in such a way that think aloud method is employed to provide primary data and later this data can be used to support and promote analysis using complementary methods.

Another data collection approach frequently applied to get insights into cognition is eye tracking. Eye tracking offers several advantages over other protocols and it has been proposed as particularly feasible for assessment of user strategies (Goldberg and Kotval 1999). Fore mostly, eye tracking is non-invasive, non-intrusive and typically does not require user cooperation and conscious awareness. Glöckner and Herbold (2010) argued that in a problem solving experiment, recording data with eye tracking methods decreases the chance of influence on the decision processes of users. They considered eye movement-based analysis as an evaluation technique that enhances the traditional performance data such as think-aloud protocols and walk-through evaluations of computer interfaces.

With few notable exceptions (e.g. Anderson et al. 2004) it is generally accepted that eye movements, fixations and the derived measures provide information about cognitive processes. For instance, Velichkovsky (1999) claimed that fixation durations increase during solving a problem with increasing the level of cognitive processing. Thus, short fixations are related to more superficial levels of processing (e.g. screening or perception), whereas longer fixations are related to deeper processing, such as deliberate consideration of information and planning (Glöckner and Herbold 2010).

Both user expertise and cognitive states have been previously modeled through an eye-tracking data analysis. Based on a machine learning classification, Liu et al. (2009) explained differences between experts and novices in building concept maps. Participants constructed collaboratively concept maps of the content for 20 minutes as their eye-movement data were recorded. Results showed 96 % recognition rate for two distinct clusters of experts and novices. The authors reported that while high-skilled participants concentrated on specific concepts longer, low-skilled participants had shorter attention spans and scattered gazes.

Liang et al. (2007) claimed that a general Support Vector Machine (SVM) is a suitable machine learning method for classification of human behavior, especially for detecting cognitive states with eye movement data. Authors demonstrated that driver distraction can be detected using driver performance measures and eye movement measures in real time.

In another study, Simola et al. (2008) applied Hidden Markov Models to predict what task a user is currently conducting, out of three information search tasks: word

search, searching for an answer, or choosing the most interesting title from a list. The model was trained on eye-tracking data and achieved an accuracy of 60.2 %.

Several other recent studies employed a machine learning approach to analyze eye-tracking data. Vidal et al. (2011) applied k-Nearest Neighbor, as a machine learning method, to distinguish feature space of fixations, saccades and smooth pursuits, together with EOG signal. Ishii and Nakano (2008) used eye movements features and an SVM to create an intelligent agent that is able to evaluate engagement in multi-user conversations.

Xu et al. (2008) proposed a personalized online content recommendation system based on acquiring individual user's visual attention over their previously read documents, browsed images and watched videos. Using a data mining pipeline, they predicted user's attention on unseen materials.

A personalized notification system reacting according to eye movements was also studied by Bailey and Iqbal (2008). Their system benefited from the correlation between visual attention to proceeded task and changes in pupillary responses as a source for adaptive notification system. With a proper alignment of user's eye movements and tasks, notification can be then presented in less disturbing moments. Xu et al. (2009) also employed user's visual attention during reading as a cornerstone for a document summarization algorithm. The hybrid summarization process suggested candidates for text summarization using the prediction of visual attention and word semantics analysis. Vrochidis et al. (2011) studied the potentials of eye-movements as a source of implicit feedback in video retrieval tasks. They built a recommendation system for finding similar topics in videos based on Support Vector Machines.

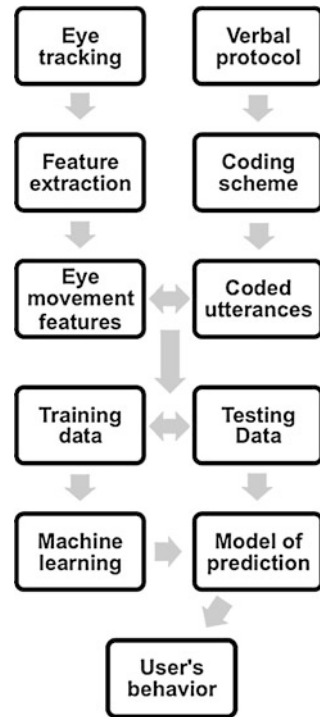
Finally, few studies have employed also pupillary data as a source for machine learning-based prediction of relevant events. For example, Bednarik et al. (2012) performed interaction intention prediction based on computational models learned from eye-movements and pupillary responses. The employed SVM classifiers achieved accuracy of about 75 %.

## 7.2 Mapping Gaze to Behavior

Modeling internal cognitive states using computational approaches is an active research topic, however complex problem solving is a domain not previously explored in greater detail using eye-movement tracking. Yet, in order to support the user's interaction with an interface in an intelligent way, the IUIs have to accurately tap into the sequence of thoughts and actions of the users.

In this study we thus employ eye-tracking to reveal such relevant information from user's ocular behavior. At the core of the method described here, gaze data are associated with human cognition states by aligning them with the annotations of think aloud protocol, as a ground truth. The presented method progresses according to the following main steps: verbal protocol analysis of the cognition, feature extraction and mapping to the verbal protocols, application of a machine learning method

**Fig. 7.1** Procedures of the proposed mapping. Adapted from Eivazi and Bednarik (2010)



for building associations between the two, and classification using the learned models.

In detail, first, we code all think-aloud data recorded from user's speech during interaction. We suggest to apply for example the coding scheme proposed by O'Hara and Payne (1998) that is based on the Ericcson (1975)'s approach and has also been applied with modifications in other studies (Morgan et al. 2007; Davies 2003). The details of the coding as applied for the case dataset are presented in Sect. 7.3.1.1.

In the second phase, we perform mapping of gaze-based data to qualitative differences in the corresponding think-aloud protocols. We compute a large set of eye-tracking features that are informed by the theories of cognition and visual attention, and for each data-point in the think-aloud protocol we build a corresponding vector of these features. In the last stage, we present the inference task as a typical classification problem and we apply machine learning and pattern recognition methods to solve it. Figure 7.1 presents the computational architecture of the proposed approach.

The mapping system described above enables us to 1) investigate the relationships between high-level cognitive traits and low-level eye-tracking data, and 2) propose a prediction real-time model to recognize user's cognitive states and user's performance. Future interactive systems can make use of such automatic modeling and classification methods.



**Fig. 7.2** Target configuration. Adapted from Eivazi and Bednarik (2010)

1	2	3
8		4
7	6	5

Because it is known that preprocessing stage is highly important for successful performance of prediction models such as the one employed here (Graf and Borer 2001) and that normalization is the central part of the process when SVM is used (Chang and Lin 2011), in the following experiments we specifically investigate the effects of various normalization and windowing approaches on the performance of the proposed system.

### 7.3 Method

In order to answer the question whether gaze data can be used to classify and predict human strategies and performance we choose the classical 8-tiles puzzle game. We employ the data collected from the experiment of Bednarik et al. (2009). Similar settings have been used in numerous studies investigating interactive human problem solving.

To achieve the goal of predicting user's actions and performance through the eye movement data, two main analysis techniques had to be carried out. First, outcome measures had to be defined and computed, including feature extraction and clustering of the features. Second task consisted of creation and validation of the prediction model. We will next briefly describe the settings of the original study and then the methods of data labeling, feature extraction and model building.

In the original study, the authors had instructed a group of participants to think aloud while solving the 8-tiles puzzle game. Each tile in the puzzle had dimensions of  $200 \times 200$  pixel, and each tile had a width of 5.29 cm and height was 5.29 cm, measured on the screen.

Fourteen participants solved three trials of the game using computer mouse as an interaction method. They started with a warm-up puzzle and a think aloud practice and then continued for three unique initial configurations of the puzzle game. The three configurations were comparable in the level of complexity and were presented in random order. The target goal state of the puzzle is shown in Fig. 7.2; the goal state was visible to the participants at bottom left side of the screen all times.

Figure 7.3 present the three initial states of the puzzle game. In addition to participants' voice protocols, eye movements were recorded using Tobii ET 1750 eye tracker. The resolution of the 17 inches screen was  $1280 \times 1024$  and the viewing distance 60 cm (Bednarik et al. 2009).

2	1	6	8		2	6	8	1
4		8	5	3	1		7	3
7	5	3	7	6	4	5	4	2

Fig. 7.3 Initial configurations. Adapted from Eivazi and Bednarik (2010)

### 7.3.1 Data Analysis

For the purposes of this experimentation, two feature extraction methods were performed on eye movements sequence. The first method computed features using windows with dynamic durations and the second method computed features for a fixed window interval. The dynamic window size was contingent with the length of the respective utterance, whereas the fixed window size was systematically modified from 200 ms to 7400 ms in 200 ms steps to evaluate the effects of the window size on the prediction performance.

#### 7.3.1.1 Outcome Measures

To address the first problem of outcome measures, verbal data were classified into six categories based on O’Hara and Payne (1998) with a slight modification. The classification categories described qualitatively the following different utterances: *Cognitions* referred to statements describing what concrete and specific information a participant is currently attending to and what information he is processing. Examples of cognition statements would be “... number 3... ok 1, 2, 8, 7” or “No, you are messed up”.

*Evaluations* were conceptually similar to cognitions, however they were less accurate about the object of interest. In addition, when participants were referring to how well they performed or what is the general situation in the problem-space, we coded that utterance as belonging to evaluations. Examples of evaluative statement would be “... this one will mess my things” or “... I am doing something...a mistake over here”.

*Plans and planning* were utterances containing a description or reference to plan development, its specific goals and detailed actions to be taken next. Examples of planning statements would be “... I have to reshuffle the first row by changing the position of 1 and 4” or “... I want to get the 6 from the upper left corner out there”.

*Intentions*, on the other hand, were utterances describing the general aims, without a specific descriptions how to achieve them. Examples of intention statements would be “... I will start playing little bit around to sort the left part” or “... how do I rotate these whole thing”.

**Table 7.1** Fixation-based features

Variable	Feature	Description	Unit
Fixation	Count*	Number of fixation	n
	Polygon area	Area covered by fixations	px2
Fixation duration	First	Features describing distribution of fixation duration (= total 8 features)	ms
	Last		
	Minimal		
	Maximal		
	Sum*		
	Mean*		
	Median		
	Standard deviation		
Distance between fixations	First	Euclidean distance between consecutive fixations (= total 8 features)	degree
	Last		
	Minimal		
	Maximal		
	Sum*		
	Mean*		
	Median		
	Standard deviation		
Angle between fixations	Mean	Angle between consecutive fixations (= total 6 features)	degree
	Minimal		
	Maximal		
	First two		
	Last two		
	First and last		

*Concurrent move* utterances referred to description of the changes in the problem along the manipulation with it. Examples of concurrent move statements would be “... 4 will be there” or “... there will be 8 coming here and the 7 coming here”.

Finally, we applied a category of not applicable for other utterances; however, we do not consider those data in this analysis. More detailed description can be found in O’Hara and Payne (1998).

The unit of analysis was one sentence. Two independent coders conducted the coding and achieved the inter-rater agreement of 86 %. Eye-gaze replay was used to resolve particular difficulties and unclear codes.

Of all three trials and all participants, the coding yielded a total of 1389 labeled utterances, of which 281 data points belonged to Cognition states, 397 data points to Evaluation activities, 130 data points to Planning, 247 data points to Intention re-

**Table 7.2** Saccade-based features

Variable	Feature	Description	Unit
Saccade	Count	Number of saccades	n
	Polygon area	Area covered by saccades	px <sup>2</sup>
Saccade duration	Sum	Features describing distribution of saccade duration (= total 4 features)	ms
	Mean		
	Median		
	Standard deviation		
Saccade amplitude	First	Euclidean distance between consecutive fixations (= total 8 features)	degree
	Last		
	Minimal		
	Maximal		
	Sum		
	Mean		
	Median		
	Standard deviation		
Saccade direction	First	Angle between consecutive saccades (= total 6 features)	degree
	Last		
	Minimal		
	Maximal		
	Sum		
	Mean		

lated utterances, and 334 utterances contained the descriptions of concurrent moves. The mean duration of a coded sentence was 6829 ms (SD = 8110).

The eye movement features that were used in this experiment are listed in Tables 7.1, 7.2 and 7.3. The features marked by (\*) were already used in the previous work by authors (Eivazi and Bednarik 2010). Similarly as the coded utterances, eye-movement data carried a timestamp, enabling their easy mapping to the verbal protocol. Furthermore, we partitioned the screen into areas of interest (AOIs). The user interface was partitioned into nine AOIs corresponding with the nine possible positions of tiles of the game, and one additional surrounding area for the remaining part of the screen. The goal state of the game was shown constantly at the left bottom of the screen. In total 49 features were computed.

### 7.3.1.2 Feature Construction, Extraction and Selection

The preprocessing of the raw eye movement data was performed using the velocity-based fixation identification algorithm (Salvucci and Goldberg 2000). The fixation

**Table 7.3** Interface-dependent features

Feature	Description	Unit
Total number of visited tiles	Total number of fixations, located on the tiles (including multi visits)	n
Unique number of visited tiles	Unique number of fixations, located on the tiles	n
Number of switches between tiles	Total number of fixation switches between the tiles	n
Number of visited goal area	Total number of fixations, located in the goal area	n

identification software was designed by authors in Matlab and the best settings for the algorithm was chosen after manually comparing the results using visual replay of the gaze. As a result, the velocity threshold 100 deg/s, the minimum fixation duration 100 ms, and the minimum distance between two gaze points of 30 pixels were applied on the dataset. The LibSVM Matlab toolbox of Hsu et al. (2003) was used to build and train the prediction model along with custom developed scripts in Matlab.

To describe the cognitive processes occurring during problem solving task, a large number of eye movement features were constructed using two window-based feature extraction methods: a dynamic window extraction and a fixed window size approach.

In the case of the dynamic window size feature extraction, each feature vector was computed from the eye-movement data belonging to the whole duration of the corresponding utterance. For example, during three seconds of single state such as evaluation, we computed the eye movement features using data from the entire interval.

In the case of the fixed window size feature extraction, each vector was computed from eye-movement data for the duration of a fixed window size. For example, during each 1000 ms window the eye movement features were computed and labeled according to the corresponding state.

Even though we constructed nearly fifty eye movement features, a question arises whether a more compact set would perform similarly and thus allow a more effective computation. Thus, we used a simple experimental evaluation method to select a subset of the features. The aim was to reduce the number of features which are sufficient to describe classes separately. The differences in individual features were observed by visually plotting the distributions of the features for each class. For modeling problem-solving behavior 10 out of 49 features were selected:

- number of fixations
- sum fixation duration
- area covered by fixations
- mean fixation distance
- unique number of visited tiles
- number of switches between tiles

- number of visited goal area
- sum saccade duration
- area covered by saccades
- sum saccade amplitude

And for modeling performance groups 20 out of 49 features were selected:

- number of fixations
- number of saccades
- maximal fixation duration
- sum fixation duration
- mean fixation duration
- area covered by fixations
- distance between first and last fixations
- minimal fixation distance
- maximal fixation distance
- sum fixation distance
- mean fixation distance
- total number of visited tiles
- number of visited goal area
- sum saccade duration
- area covered by saccades
- maximal saccade amplitude
- sum saccade amplitude
- mean saccade amplitude
- minimal change in saccade direction
- maximal change in saccade direction

### 7.3.1.3 Prediction Model

To address the second task (prediction model) we employ a well-established machine learning approach. Support vector machine (SVM) is a standard and frequently applied tool that has been previously shown performing well for various classification tasks (Meyer et al. 2003). SVM has been successfully used in detection, verification, recognition, and information retrieval from a range of datasets (Liang et al. 2007). Liang et al. (2007) presented three reasons that make SVM suitable for classification of human cognition states: first, it is rarely possible to represent cognitive states of humans by a linear model. SVM can compute nonlinear models as efficiently as the linear models. Second, SVM can be applied without prior knowledge before training. In addition, it can extract information from noisy datasets. Third, while traditional learning methods (e.g., logistic regression) only minimize training error, SVM minimizes the upper bound of the generalization error. This makes SVM able to produce more robust models. In our application, SVM is used as a supervised learning classification method.

In total, three prediction models were trained to predict user's actions and performance. The first model learns the patterns of problem-solving behaviors on a 5-class problem corresponding to all five states and corresponding eye movement features (hereafter we refer to this task as the *5-class task*). The second model deals with a simplified 2-class task of detecting planning and intention activities versus cognition and evaluation activities (hereafter *2-class task*).

The third model searches for patterns in data vectors originating from different performance groups (two classes, high- and low-performing participants) and eye movement features (we refer to this task as the *performance task*). The task is to predict to which performance group any given data vector belongs.

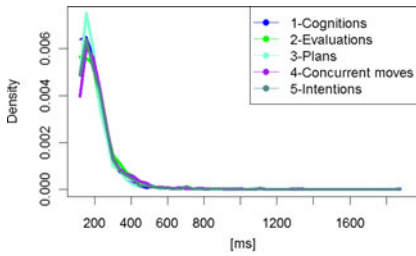
For the first model, the ground truth was labeled for each class (five coding states) in the sample data. For the second model, ground truth was created by merging the five coding states into two coding states. Finally for the last model, the ground truth was established by computing the task completion times.

The LibSVM Matlab toolbox developed by Hsu et al. (2003) was used to build the prediction models. Whole dataset was randomly partitioned into 70 % training data and 30 % testing data for which the gaze data was available. Normalization of data was applied in two ways on the balanced dataset: linear transformation within an interval [0, 1], and Z-score normalization. Both training and testing data were normalized with the same method. For linear transformation, first from the training data the minimum and maximum values were computed for each feature separately, and next the same minimum and maximum value were applied in testing data. Similarly in Z-score normalization, mean and standard deviation of each features were computed separately and next the same constants were applied in testing data.

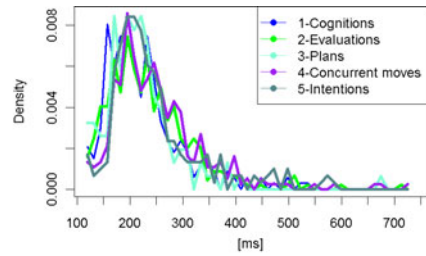
Similar to the Eivazi and Bednarik (2010) approach in order to find best hyper-parameter of the prediction model, an optimal cost ( $c$ ) and gamma ( $g$ ) value of the RBF kernel was estimated by a 2D grid search (Hsu et al. 2003). The prediction performance for each hyper-parameter setting was computed by 5-fold cross validation on the training data. The training data was randomly divided into five subsets. Consequently, one subset was tested by using the model based on the remaining data (four subsets). The objective function for performance measure was Equal Error Rate (EER) in 2-classes prediction models and for 5-classes prediction model the objective function was accuracy.

## 7.4 Results

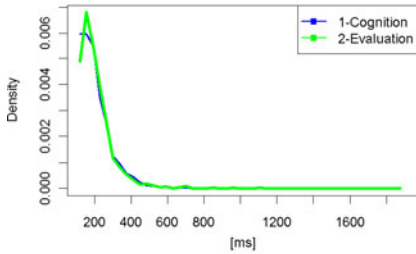
In this chapter we analyze user behavior during a problem-solving task. In particular, we analyze the eye movement data and features as subsets aligned with the categories of verbal protocols. Compared to Eivazi and Bednarik (2010) study, we changed the setting to a randomized and task-independent approach. Here, we systematically evaluate the effects of normalization method and the type and size of windowing for feature extraction.



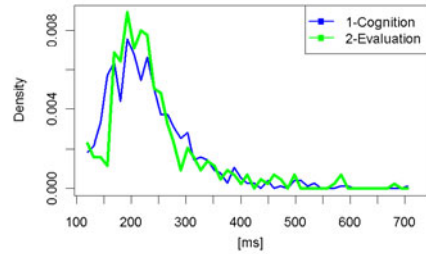
(a) 5-class classification task with the one second fixed window size setting



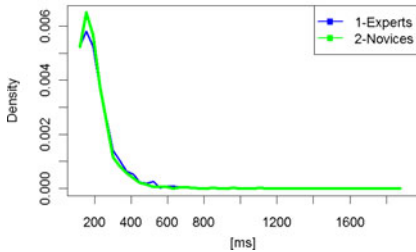
(b) 5-class classification task with the dynamic window size setting



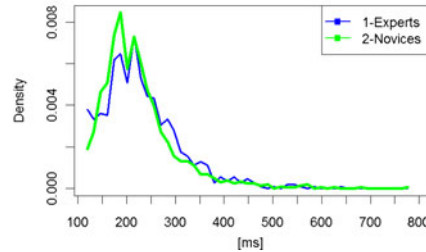
(c) 2-class classification task with the one second fixed window size setting



(d) 2-class classification task with the dynamic window size setting



(e) Performance classification task with the one second fixed window size setting



(f) Performance classification task with the four second fixed window size setting

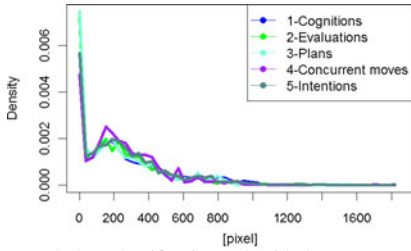
**Fig. 7.4** Histograms of mean fixation duration

### 7.4.1 Descriptive Analysis

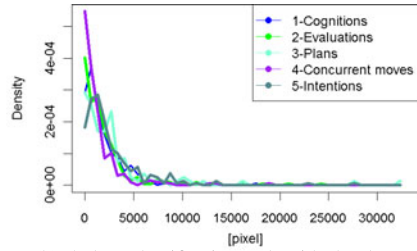
Compared to Eivazi and Bednarik (2010, 2011), we present a more complete set of features and a more detailed evaluation of different feature extraction parameters and normalization methods. In the following, we highlight some of the peculiarities of applying various settings of feature extraction window.

The differences in individual features related to the cognitive states were generally small and the features contained great variances. In addition, window size seems to have an effect on the shape of the feature distributions. To demonstrate visually the nuances, we plot two features under changing settings, see Fig. 7.4 and Fig. 7.5.

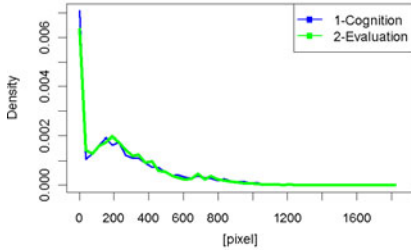




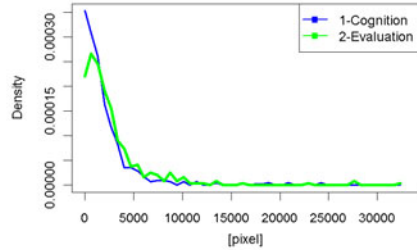
(a) 5-class classification task with the one second fixed window size setting



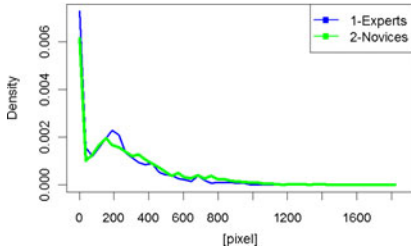
(b) 5-class classification task with the dynamic window size setting



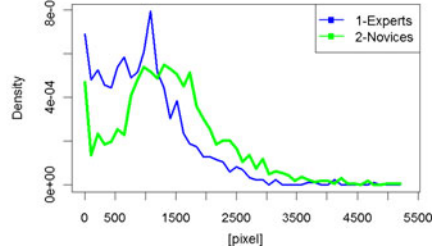
(c) 2-class classification task with the one second fixed window size setting



(d) 2-class classification task with the dynamic window size setting



(e) Performance classification task with the one second fixed window size setting



(f) Performance classification task with the four second fixed window size setting

**Fig. 7.5** Histograms of sum fixation distance

Figure 7.4 shows the distributions of mean fixation duration. In the left column, a window size of one second is applied, while the right column contains visualizations of dynamic window effect (Figs. 7.4b and 7.4d) and a four-second window (Fig. 7.4f). It seems that longer window sizes better highlight differences between the distributions of the mean fixation duration for the respective classes for all three problems, as the distribution plots and peaks overlap less.

Figure 7.5 presents the distributions of the sum of fixation distance feature under the three tasks and window sizes. The left column shows distributions when a relatively short one-second window was applied. Similarly as with the previous case, the effect of the window size during feature extraction is visible through better separation of the histograms when longer window size was applied, in particular for the performance classification task (Fig. 7.5f).

## 7.4.2 Classification Performance

### 7.4.2.1 Problem-Solving States

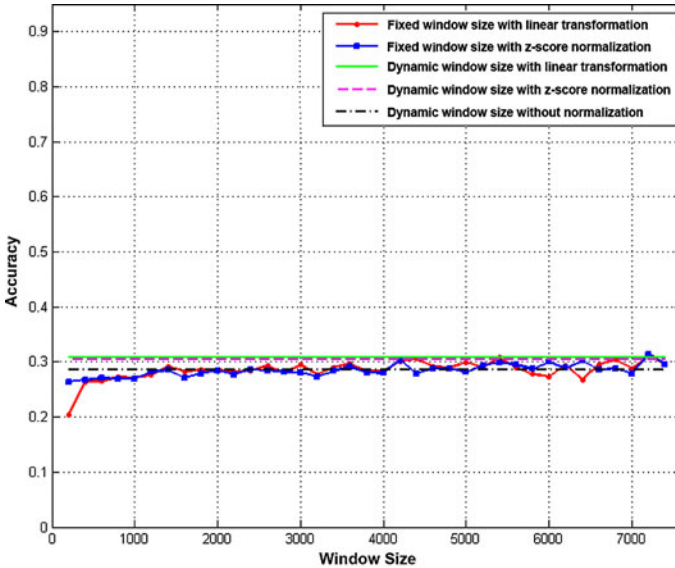
The performance on the multi-class classification task with five distinct classes has been compared using two sets of features, two normalization methods, and by systematically adjusting the window size. A baseline majority classifier would have an accuracy of 29 % in this case. The best resulting accuracy of 31 % was achieved when dynamic window size method and linear transformation method were applied together. When window size was fixed, the best accuracy was 32 % (window size of 4800 ms). While performance was slightly better when data were preprocessed using normalization, the type of normalization had no effect on the results. Figure 7.6a shows the effect of window size and normalization method on the prediction performance. The enumerations of the prediction performances for all tested window sizes and normalization settings are presented in Table 7.4.

Considering that a multi-class classification problem is far more challenging compared to a binary classification problem, we turn the problem into 2-class classification problem by merging the related classes. The first such resulting class includes samples from both *cognition* and *evaluation* classes and the second class includes samples from both *planning* and *intention* classes. Samples related to the concurrent move class were ignored in this task.

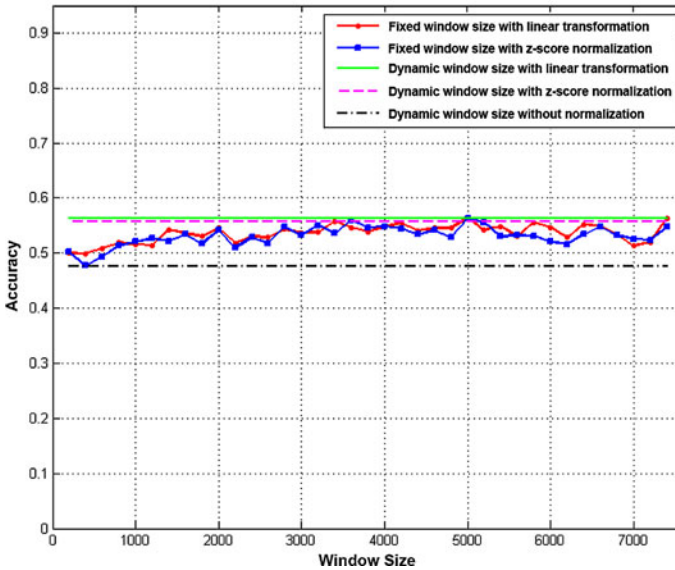
Similar to the multi-class classification task, using two sets of features, two normalization methods, and systematically adjusted window size the prediction performances for 2-class classification task have been compared. A baseline majority classifier would have an accuracy of 64 %. The best resulting accuracy of 60 % was achieved when dynamic window size method and linear transformation method were applied together. When window size was fixed 3400 ms, the best accuracy was 56 %. Normalization had a sizable effect on the performance, however, both methods of normalization performed nearly equally when the window size was fixed. When using dynamic window size, linear normalization showed a slight improvement over the Z-score normalization. Figure 7.6b shows the effect of window size and normalization method on the prediction performance. The prediction performances for all tested window sizes and normalization settings are presented in Table 7.5.

### 7.4.2.2 Problem-Solving Performance Classification

In this study, all users were divided into two groups: expert and novice. The average time for solving the puzzle was 234 seconds and thus the users who solved the puzzle less than 234 seconds were denoted as experts and other users as novices. In total, nine users belonged to the expert group and five users were regarded as novices.



(a) 5-class case, accuracy (higher is better)



(b) 2-class case, accuracy (higher is better)

**Fig. 7.6** Window size and normalization effects on prediction performance, **a** 5-class and **b** 2-class problem

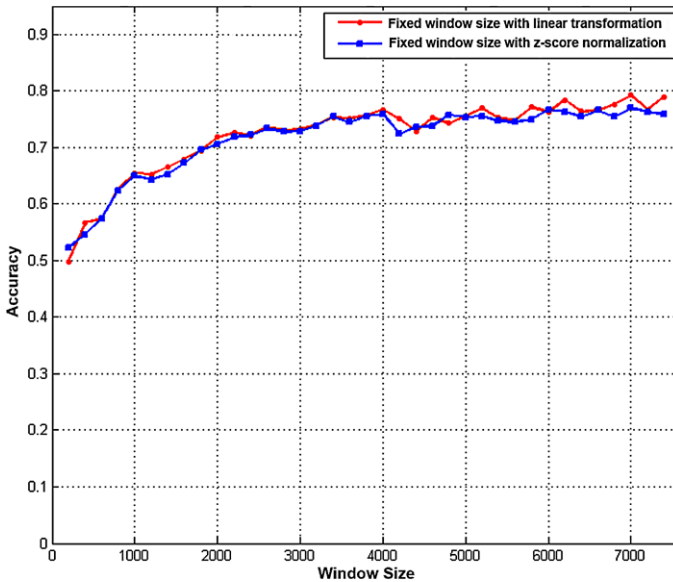
The performance of 2-class classification task with the expert and novice groups has been compared using two sets of features, two normalization methods, and a systematically adjusted window size. A baseline majority classifier would have an

**Table 7.4** The effect of window size and normalization setting on 5-class prediction performance

Window size [ms]	Accuracy (testing)			
	Linear normalization		Z-score normalization	
	feature subset	all features	feature subset	all features
200	20	18	26	26
400	26	27	27	26
600	26	27	27	27
800	27	27	27	27
1000	27	28	27	27
1200	28	27	28	28
1400	29	28	29	28
1600	28	29	27	27
1800	28	28	28	28
2000	28	29	28	28
2200	28	30	28	29
2400	29	29	29	29
2600	29	29	28	29
2800	29	27	28	28
3000	29	28	28	30
3200	28	27	27	28
3400	28	30	28	29
3600	30	30	29	28
3800	28	29	28	29
4000	28	28	28	29
4200	30	31	30	30
4400	31	31	28	31
4600	29	30	29	29
4800	29	31	29	28
5000	29	31	28	29
5200	30	31	29	32
5400	31	31	30	30
5600	29	30	29	31
5800	28	31	29	31
6000	27	31	30	29
6200	29	30	29	30
6400	27	31	30	29
6600	30	30	29	29
6800	30	32	29	30
7000	29	31	28	30
7200	31	31	31	31
7400	30	32	30	31
Average	28.43	29.19	28.43	28.92
Dynamic	31	31	31	30

**Table 7.5** The effect of window size and normalization setting on 2-classes prediction performance

Window size [ms]	Accuracy (testing)			
	Linear normalization		Z-score normalization	
	feature subset	all features	feature subset	all features
200	50	49	50	49
400	50	50	48	48
600	51	50	49	49
800	52	51	51	51
1000	52	52	52	54
1200	51	53	53	53
1400	54	53	52	53
1600	54	53	53	53
1800	53	51	52	52
2000	54	55	54	53
2200	52	52	51	55
2400	53	55	53	55
2600	53	54	52	56
2800	54	55	55	55
3000	54	55	53	54
3200	54	53	55	54
3400	56	54	54	56
3600	55	56	56	56
3800	54	54	55	56
4000	55	55	55	54
4200	55	53	54	54
4400	54	54	53	56
4600	55	55	54	54
4800	55	54	53	56
5000	56	53	56	54
5200	54	56	55	55
5400	55	55	53	55
5600	53	52	53	56
5800	56	54	53	55
6000	55	54	52	56
6200	53	52	52	52
6400	55	53	53	56
6600	55	54	55	54
6800	53	56	53	57
7000	51	53	53	55
7200	52	55	52	52
7400	56	55	55	56
Average	54	54	53	54
Dynamic	60	56	58	56



**Fig. 7.7** Window size and normalization effect on performance prediction

accuracy of 64 %. The best resulting accuracy of 79 % was achieved when 7000 ms fixed window size and linear transformation method were applied together. Linear normalization had always performed slightly better than Z-score transformation. Figure 7.7 shows the effect of window size and normalization method on the prediction performance. The prediction performances for all tested window sizes and normalization settings are presented in Table 7.6.

## 7.5 Discussion

In this chapter we presented an evaluation of window size and normalization methods for automatic classification of problem solving strategies from eye-tracking data. We presented the approach for such modeling, the design of features, two normalization methods, and performance results in three tasks.

The preliminary results from a previous experimentation on this dataset shown a 53 % accuracy on the prediction accuracy for the five classes task (Eivazi and Bednarik 2010) and 66 % accuracy on the prediction accuracy for the performance classification task (Eivazi and Bednarik 2011). While ten features were computed using primarily fixation data in the preliminary work, in the present work we significantly increased the number of features by five folds.

In our previous work, the baseline performance was established as a classification accuracy of a majority classifier. For the problem-solving states task, the baseline majority classifier had an accuracy of 27 % and for the problem-solving perfor-

**Table 7.6** The effect of window size and normalization setting on performance prediction performance

Window size [ms]	Accuracy (testing)			
	Linear normalization		Z-score normalization	
	feature subset	all features	feature subset	all features
200	50	49	52	50
400	57	54	54	53
600	57	58	57	54
800	63	60	62	56
1000	65	62	65	60
1200	65	61	64	61
1400	66	64	65	62
1600	68	65	67	64
1800	69	66	69	65
2000	72	69	71	68
2200	73	69	72	69
2400	72	69	72	68
2600	74	70	73	68
2800	73	70	73	69
3000	73	72	73	69
3200	74	71	74	69
3400	75	72	75	71
3600	75	72	74	72
3800	76	73	75	72
4000	77	73	76	72
4200	75	75	72	75
4400	73	74	74	72
4600	75	75	74	74
4800	74	73	76	73
5000	75	75	75	75
5200	77	76	75	75
5400	75	75	75	73
5600	75	75	74	74
5800	77	76	75	75
6000	76	77	77	74
6200	78	76	76	74
6400	76	76	75	75
6600	76	76	77	74
6800	78	76	75	74
7000	79	77	77	74
7200	77	76	76	75
7400	79	77	76	78
Average	72	70	71	69

mance classification task the baseline majority classifier had an accuracy of 55 % (Eivazi and Bednarik 2011).

In the Eivazi and Bednarik (2010) approach, the selection of training and testing data, the normalization approach, and a k-fold cross validation may not be generalizable to the real time prediction systems. This is due to the fact that the whole dataset was subdivided only into two training and testing datasets according to boundaries of the three problem-solving sessions. The training data was then derived from the first two trials and the remaining last trial was selected for testing purposes. It is thus questionable, whether such approach brings about some time-dependency and whether it does consider a truly random selection of data as in a real time system.

Furthermore, the normalization method originally involved was both session and class dependent, which is a possible approach given the data set is available as in an offline prediction. However, in real-time systems the target class label is not known beforehand, and thus the normalization phase cannot distinguish it. We adopted a class-independent approach in this chapter.

Finally, the comparison with previous experimentation needs to take into account the metrics of prediction performance. Namely, the number of classes in each trial was not balanced before, and therefore using accuracy as an objective function may not be adequate for more detailed studies. Moreover, using accuracy as a measure in k-fold cross-validation method may lead to choosing an not generalizable parameters for SVM prediction model. The results of the Eivazi and Bednarik test show large differences between cross validation and testing data accuracy, which we believe is the results of improper objective function.

### ***7.5.1 The Current Results***

The presented work is built upon a significantly improved machine learning pipeline. The described approach, however, did not achieve classification performances as high as reported before. In sum, in the 5-class problem, the best accuracy was 32 % using linear transformation method and 6800 ms window size. However, around the same 31 % accuracy was achieved also using dynamic window size. When the number of classes was reduced as in the 2-class problem, the best accuracy was 60 % using the dynamic window size and linear transformation method. Moreover, the accuracy was 56 % when 3400 ms window size with linear transformation method were used to train classifier. In the performance classification task, the highest accuracy was 79 % using linear transformation method with 7000 ms window size.

In sum, our results show that the linear transformation method suits better for most cases, however the prediction performance differences were not large. It seems that longer window-size works better than short windows, and that the dynamic window approach performs best. The challenge is, however, that in a real-time implementation one would need to wait longer for the classifier response. In addition, implementation of the dynamic windowing in the real-time is hard, since the boundaries between different labels would need to be predicted.



A visual comparison of Fig. 7.4 and Fig. 7.5 suggests that there may be an interaction between the windowing parameters, feature type and task at hand. We plan to investigate this phenomena in more detail in our future work.

Even though it is well known that oculomotor behavior is task-dependent to a certain degree (Yarbus 1967; Lipps and Pelz 2004), the way the present method builds on this observation seems not entirely effective. When dealing with classification of user strategies during problem solving, the individual differences in certain eye-movement measures are known to be large and temporally dependent (Bednarik et al. 2006; Bednarik and Tukiainen 2008). We believe that is one of the reasons for the observed performance results presented above.

## 7.6 Conclusions and Future Work

In this chapter we applied an SVM-based classification to predict problem-solving cognition states and user's performance. The goal was to evaluate whether eye tracking can be used to detect cognitive behavioral patterns for a purpose of proactive intelligent user interfaces. We clearly showed that normalization of eye-movement features is beneficial for classification performance. Although the method of normalization does not seem have significant effects on the overall performance, we recommend applying linear scaling that seemed to perform slightly better. Considering the effects of window size on extraction of the eye-movement features, we observed that even for same feature, the density distributions can vary depending on the size of the window. The prediction performance of the classification however generally improves with increased sizes of the extraction window.

Finally, we performed a naïve selection of features to show that computational complexity can be significantly decreased without impacting the performance. This observation opens new pathways for future research, in which we plan to investigate the relative contribution of the features. We hypothesize that the discriminative power of a feature should be evaluated across different window sizes as it seems that some features provide good performance at short windows while other features work best at longer extraction boundaries.

Compared to our previous attempt, we improved the methodology to a more rigorous settings that truly emulates the conditions of real-time prediction problem. The accuracy of cognitive activity classification was not extremely great and while one can see them as dissatisfying, we perceive them as setting the baseline of what is possible to achieve with applying the rigorous approach presented here.

## References

- Anderson JR, Bothell D, Douglass S (2004) Eye movements do not reflect retrieval: limits of the eye-mind hypothesis. *Psychol Sci* 15:225–231

- Bailey BP, Iqbal ST (2008) Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans Comput-Hum Interact* 14(4):1–28
- Bednarik R (2005) Potentials of eye-movement tracking in adaptive systems. In: Proceedings of the fourth workshop on the evaluation of adaptive systems, held in conjunction with the 10th international conference on user modeling (UM'05), pp 1–8
- Bednarik R, Tukiainen M (2008) Temporal eye-tracking data: evolution of debugging strategies with multiple representations. In: Proceedings of the 2008 symposium on eye tracking research & applications. ACM, New York, pp 99–102
- Bednarik R, Myller N, Sutinen E, Tukiainen M (2006) Analyzing individual differences in program comprehension. *Technol Instr Cogn Learn* 3(3/4):205
- Bednarik R, Gowases T, Tukiainen M (2009) Gaze interaction enhances problem solving: effects of dwell-time based, gaze-augmented, and mouse interaction on problem-solving strategies and user experience. *J Eye Movement Res* 3(1):1–10
- Bednarik R, Vrzakova H, Hradis M (2012) What you want to do next: a novel approach for intent prediction in gaze-based interaction. In: Proceedings of the 2012 symposium on eye-tracking research & applications, ETRA'12. ACM, New York
- Chang CC, Lin CJ (2011) LibSVM: a library for support vector machines. *Science* 2(3):1–39
- Conati C, Merten C (2007) Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. *Knowl-Based Syst* 20:557–574
- Davies SP (2003) Initial and concurrent planning in solutions to well-structured problems. *Q J Exp Psychol, A Hum Exp Psychol* 56(7):1147–1164
- Eivazi S, Bednarik R (2010) Inferring problem solving strategies using eye-tracking: system description and evaluation. In: Proceedings of the 10th Koli Calling international conference on computing education research, Koli Calling'10. ACM, New York, pp 55–61
- Eivazi S, Bednarik R (2011) Predicting problem-solving behavior and performance levels from visual attention data. In: Proceedings of 2nd workshop on eye gaze in intelligent human machine interaction at IUI, pp 9–16
- Ericsson KA (1975) Instruction to verbalize as a means to study problem solving process with the 8-puzzle: a preliminary study. Department of Psychology, University of Stockholm
- Ericsson KA, Simon HA (1993) Protocol analysis: verbal reports as data revised edition. MIT Press, Cambridge
- Glöckner A, Herbold AK (2010) An eye-tracking study on information processing in risky decisions: evidence for compensatory strategies based on automatic processes. *J Behav Decis Mak* 41(1):71–98
- Goldberg JH, Kotval XP (1999) Computer interface evaluation using eye movements: methods and constructs. *Int J Ind Ergon* 24:631–645
- Graf ABA, Borer S (2001) Normalization in support vector machines. In: Proceedings of the 23rd DAGM-symposium on pattern recognition. Springer, London, pp 277–282
- Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Technical report, National Taiwan University
- Ishii R, Nakano YI (2008) Estimating user's conversational engagement based on gaze behaviors. In: Proceedings of the 8th international conference on intelligent virtual agents (IVA'08), pp 200–207
- Just MA, Carpenter PA (1976) Eye fixations and cognitive processes. *J Cogn Psychol* 8:441–480
- Kaller CP, Rahm B, Bolkenius K, Unterrainer JM (2009) Eye movements and visuospatial problem solving: identifying separable phases of complex cognition. *Psychophysiology* 46:818–830
- Liang Y, Reyes ML, Lee JD (2007) Real-time detection of driver cognitive distraction using support vector machines. *IEEE Trans Intell Transp Syst* 8:340–350
- Lipps M, Pelz JB (2004) Yarbus revisited: task-dependent oculomotor behavior. *J Vis* 4(8):115
- Liu Y, Hsueh PY, Lai J, Sangin M, Nüssli MA, Dillenbourg P (2009) Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications. In: Proceedings of the 2009 IEEE international conference on multimedia and expo

- Loboda TD, Brusilovsky P (2010) User-adaptive explanatory program visualization: evaluation and insights from eye movements. *User Model User-Adapt Interact* 20:191–226
- Meyer D, Leischa F, Hornikb K (2003) The support vector machine under test. *Neurocomputing* 55:169–186
- Morgan PL, Waldron SM, King SL, Patrick J (2007) Harder to access, better performance? The effects of information access cost on strategy and performance. In: *Proceedings of the 2007 conference on human interface: part I*. Springer, Berlin, pp 115–125
- O'Hara KP, Payne SJ (1998) The effects of operator implementation cost on planfulness of problem solving and learning. *Cogn Psychol* 35:34–70
- Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychol Bull* 124(3):372–422
- Salvucci DD (2001) An integrated model of eye movements and visual encoding. *J Cogn Syst* 1(4):201–220
- Salvucci DD, Goldberg JH (2000) Identifying fixations and saccades in eye-tracking protocols. In: *Proceedings of the 2000 symposium on eye tracking research & applications, ETRA'00*. ACM, New York, pp 71–78
- Simola J, Salojärvi J, Kojo I (2008) Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cogn Syst Res* 9(4):237–251
- Smith JD, Graham TCN (2006) Use of eye movements for video game control. In: *ACM advancements in computer entertainment technology (ACE'06)*. ACM, New York, article no. 20
- Surakka V, Illi M, Isokoski P (2003) Voluntary eye movements in human-computer interaction. North-Holland, Amsterdam, p 471 (Chap 22)
- van Someren MW, Barnard YF, Sandberg JAC (1994) *The think aloud method: a practical guide to modelling cognitive processes*. Academic Press, San Diego
- Velichkovsky BM (1999) From levels of processing to stratification of cognition: converging evidence from three domains of research. Benjamins, Amsterdam
- Vidal M, Bulling A, Gellersen H (2011) Analysing EOG signal features for the discrimination of eye movements with wearable devices. In: *Proceedings of the 1st international workshop on pervasive eye tracking and mobile eye-based interaction, PETMEI'11*. ACM, New York, pp 15–20
- Vrochidis S, Patras I, Kompatsiaris I (2011) An eye-tracking-based approach to facilitate interactive video search. In: *Proceedings of the 1st ACM international conference on multimedia retrieval, ICMR'11*. ACM, New York, pp 43:1–43:8
- Xu S, Jiang H, Lau FC (2008) Personalized online document, image and video recommendation via commodity eye-tracking. In: *Proceedings of the 2008 ACM conference on recommender systems, RecSys'08*. ACM, New York, pp 83–90
- Xu S, Jiang H, Lau FC (2009) User-oriented document summarization through vision-based eye-tracking. In: *Proceedings of the 14th international conference on intelligent user interfaces, IUI'09*. ACM, New York, pp 7–16
- Yarbus AL (1967) *Eye movements during perception of complex objects*. Plenum, New York, pp 171–196 (Chap VII)

**Part III**  
**Gaze Awareness in HCI**

# Chapter 8

## Gazing the Text for Fun and Profit

Ralf Biedert, Georg Buscher, and Andreas Dengel

**Abstract** Reading digital books is becoming more and more common, and modern interface technologies offer a wide range of methods to interact with the user. However, few of them have been researched with respect to their impact on the reading experience. With a special focus on eye tracking devices we investigate how novel text interaction concepts can be created. We present an analysis of the eyeBook, which provides real time effects according to the read process. We then focus on multi-modal interaction and the usage of EEG devices for the recording of evoked emotions.

### 8.1 Introduction

Digital book sales have just recently surpassed (Miller and Bosman 2011) the number of paper books. Hundreds of thousands of smart phones, tablets and e-book readers are sold every day and with them new interaction techniques. With the advent of sophisticated touch screens, speech recognition processing power and storage, these devices are capable of providing dense multi medial experiences. At the same time the first eye tracking vendors began to target a mass market production and miniaturization of their units (Eisenberg 2011) and they likewise are high potential candidates for an integration into future generations of digital companions.

Plenty of research and development on interactive gaze-based applications has emerged since eye tracking has first been used for entertainment (Starker and Bolt 1990). However, we would like to focus the attention on a topic that is in our opinion

---

R. Biedert (✉) · A. Dengel  
German Research Center for Artificial Intelligence, Trippstadter Strasse 122,  
67663 Kaiserslautern, Germany  
e-mail: [rb@xr.io](mailto:rb@xr.io)

A. Dengel  
e-mail: [andreas.dengel@dfki.de](mailto:andreas.dengel@dfki.de)

G. Buscher  
Microsoft Bing, One Microsoft Way, Redmond, WA 98052, USA  
e-mail: [georgbu@microsoft.com](mailto:georgbu@microsoft.com)

quite underrepresented so far: text, and the entertaining potential eye tracking can provide. Text, as we argue in this respect, is special. Like no other stimulus have eye movements on text, commonly referred to as reading, been researched for more than hundred years (Rayner 1998). In fact, it was reading which motivated the development of first eye tracking methods.<sup>1</sup> Also, eye tracking has seen great progress during the last some forty years with the availability of simple, video-based remote eye tracking and computing devices capable of reacting to these measurements in less than the blink of an eye. What makes eye tracking on text so special today are a number of reasons. It is highly structured and imposes a certain behavior onto its readers and reveals their cognitive processes towards the machine. During the time the users are reading there is some sort of shared knowledge what's on their mind which the machine is likewise able to comprehend. In contrast to scenery perception or images the machine is able to elicit a huge amount of information of the given text, thanks to the widespread availability of search engines, semantic databases and the Internet as a whole. While we could already show that these properties allow for ground-breaking interfaces and information provision systems (Biedert et al. 2010; Buscher 2010), in this work we want to investigate how gaze tracking can be used to improve the *experience* part of the reading experience. We consider this to be of special concern due to two reasons. The sublime one is that literacy education and motivating (especially the young) to read has become a key issue recently since reading rates among youths are dropping in the U.S.<sup>2</sup> (Anonymous 2007). The more practical reason is the fact that most e-books sold today are, in fact, bought and consumed for pleasure nonetheless (Milliot 2011). Hence, focusing only on the practical aspects of improving information transmission would miss the point.

With these points in mind the rest of the chapter is structured as follows. In Sect. 8.2 we start by giving a general overview on related work in the domain of gaze-active textual interfaces. In Sect. 8.3 we present the foundations upon which we built our applications. This includes the technical frameworks and core structures, such our extensions to the current HTML standards,<sup>3</sup> as well as a real-time reading detection algorithm needed by most prototypes. In Sects. 8.4 and 8.5 we present two prototypes that facilitate gaze to enhance the reading experience, either as an uni-modal input, or as an integrated application in combination with speech recognition and handwriting interaction. Both serve the purpose of outlining how these multimodal gaze concepts are being perceived by users. Section 8.6 reports on our findings how the emotions evoked during reading can be recognized and linked to the underlying text. Eventually we argue that the real-time recording and aggregation of gaze and the books' evoked emotions can lead to an improvement of the texts themselves.

---

<sup>1</sup>Although, one might argue, that technology had nothing in common with the devices researchers have access to today.

<sup>2</sup>And likely in some other countries as well.

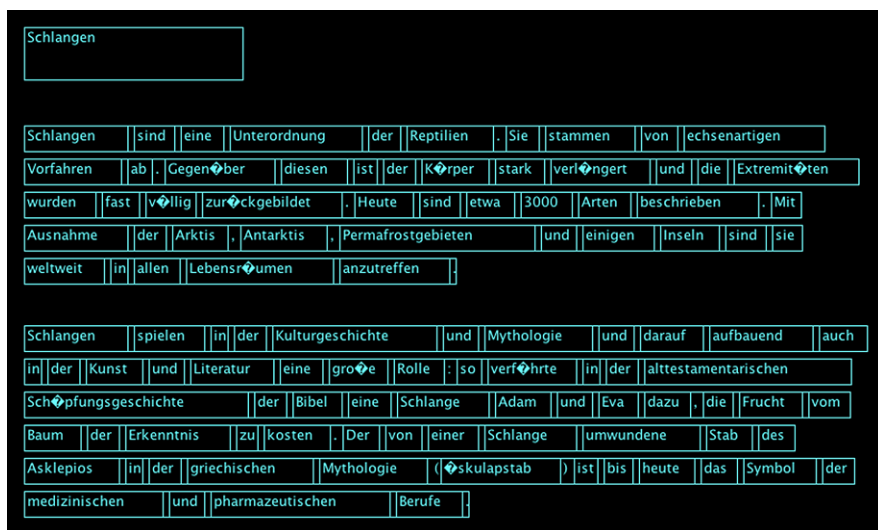
<sup>3</sup>See <http://www.w3.org/MarkUp/>.

## 8.2 Related Work

Tracking eye movements on text has a long history. In first experiments conducted during the 19th century, subjects reading text were monitored with the simplest means and the findings were basically of descriptive nature. Javal (1878), Landolt (1881) and Lamare (1892) were among the first to conduct *eye movement* studies (Wade and Tatler 2009), and eye movements on text were among the first aspects to be researched. While early experiments were of rather descriptive nature and provided early evidence that the eye moves in a series of *jerks* (i.e., saccades) while reading, the second half of the 20th century started to focus on cognitive aspects. Especially during the last thirty years of the last century the available tracking methods improved dramatically and with the availability of remote eye tracking devices and a computer-based evaluation of eye movements there was a dramatic increase in insights into the human perception and reading process (Rayner 1998). Sophisticated experiments could be performed with gaze-contingent stimuli, based on the subject's eye movements and behavior. Also, the first truly interactive eye tracking applications were implemented (Starker and Bolt 1990) in which eye tracking was used for entertainment. However, the real-time usage of gaze on text, for the sake of entertainment or information provision, has not explicitly been considered for a long time. The first application focusing on that aspect was iDict by Hyrskykari et al. (2000, 2006), which was designed to provide translations on comprehension problems detected in the reader's gaze patterns. In it translations tooltips (glosses) are presented after a certain dwell time on problematic words or provided in a gaze-responsive side bar.

## 8.3 Combining Gaze and Text

In order to properly implement and evaluate applications that can react on eye movements we need a way to integrate tracking data in a structured way, so that they can be combined with text *naturally*. While in many scenarios a rendered image that is subsequently studied in an eye tracking experiment is sufficient, there are challenges imposed by complex document layouts and textual interaction that require a more sophisticated solution. Hence, as the principal foundation for our approach we chose a browser as a rendering engine which we extend with gaze functionality. It already contains established means to structure and layout documents (i.e., HTML), as well as the necessary scripting facilities to react to external events (i.e., JavaScript). Into the browser we integrate a plugin that interfaces with the loading and setup process of a document, preprocessing and preparing it for the implementation of various interactive handlers as well as subsequent analysis. While some parts of the actual integration are merely implementation aspects, in this chapter we focus on key functionality actually needed by the applications and experiments that build upon it.



**Fig. 8.1** Internal representation of the document/application which is being used for internal processing and later evaluation. The geometry of all words and punctuations is accessible at runtime and can also be stored for later evaluation, along with incoming gaze, interaction, and EEG data

### 8.3.1 Document Preparation and Recording

After a document<sup>4</sup> is opened by a browser it transforms the structured HTML document into a DOM<sup>5</sup> tree which serves as the browser's basic data structure for rendering. Although the process itself is mostly straightforward the biggest challenge in this process is the transformation and handling of text and text nodes. While most of the resulting DOM model is, from an application's point of view, easily accessible and computable, e.g., in terms of size and layout, text nodes are usually treated as opaque blocks unavailable for detailed inspection. We address this problem by an intermediate step called *spanification* (Biedert et al. 2010) in which we break up each text node into individual words, forming DOM elements on their own, for which we can then again receive geometry information. This mechanism constructs an internal representation for which we can get the exact document coordinates and bounding boxes for textual elements, including punctuation and are able to store arbitrary meta-attributes to the individual elements during runtime, compare Fig. 8.1.

The preprocessing step also allows for another important property for the conduction of experiments, which is interaction recording. Taking advantage of the fact

<sup>4</sup>We use the terms *document* and *application* synonymously, since in our case each application has the nature of a (often highly dynamic) HTML/text document, and also all the HTML documents we work with usually have a considerable amount of processing logic built in or at least rely on them.

<sup>5</sup>See <http://www.w3.org/DOM/>.



that the relevant page geometry is known during runtime, the geometry information is serialized along with arriving gaze data, and for example mouse, interaction and EEG measurements, into a storage file that enables us to replay and automatically process the interaction session.

### 8.3.2 *Event Handling*

Apart from the general preparations for gaze-based interaction and recording we also have to consider how we can integrate eye tracking data into the document format elegantly. Since there is already an established mechanism for event handling in the DOM model<sup>6</sup> probably familiar to most web developers, we integrate gaze in a similar fashion. We propose a set of new event handlers (attributes) that can be attached directly to nodes in order to make them react to certain types of gaze patterns. The handlers we propose are:

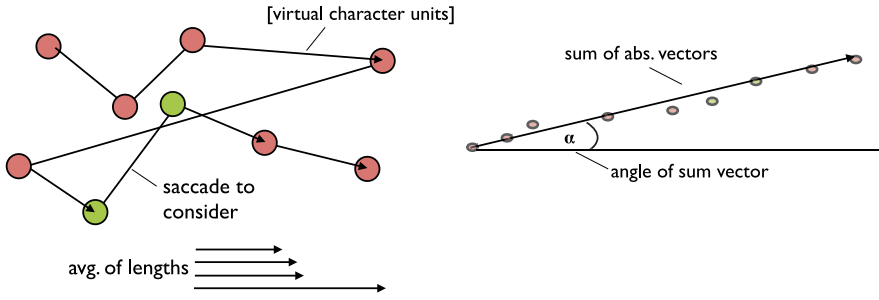
- `onGazeOver` The script annotated with this attribute is being executed when a fixation enters the element, or any embedded element, for the first time. Subsequent fixations on the element are ignored.
- `onGazeOut` The associated script is being executed when the first fixation is detected outside the tagged element, given that a fixation was detected within before.
- `onFixation` The associated code is executed on every detected fixation to the element.
- `onRead` Triggered when reading is detected (see Sect. 8.3.3 below) and one of the last saccades was moving over the specified element.

### 8.3.3 *Reading Detection*

As mentioned in the introduction, the integral part of our work is focused on text. Hence, a robust and reliable detection of reading is a prerequisite to properly implement many of our prototypes. Depending on the specific application there are a number of reasons why one would want to use such a reading detector. The most common one is to ensure that the application's reactions are bound to actual reading taking place. For example, one would want to trigger certain reactions such as acoustic effects only when a user actually reads a given paragraph, not merely when he looks at it occasionally or accidentally. While for some applications simple solutions, e.g., measuring raw average character progress, sometimes work they usually fail with a decrease in font size or increase of eye tracking noise. Thus, in this section we present our most thorough solution (Biedert et al. 2012) that forms likewise the foundation for a number of applications.

---

<sup>6</sup>See <http://www.w3.org/TR/DOM-Level-2-Events/events.html>.



**Fig. 8.2** Generation of the two feature vectors. The average forward speed (*left*) is computed as the normalized forward lengths of all saccades that pointed approximately into the reading direction. The angularity (*right*) reflects how horizontal or vertical the window  $w$  actually is. Both serve as the input for the successive classification

This approach uses *overall saccade shape* of a number of subsequent saccades (see Fig. 8.2). The algorithm's principal inputs are the incoming stream of gaze data, as well as the average font size of the area below the observed saccades, as well as the information whether there was text present at all. If there is no text present, such as when the user's gaze is directed on an image, we already know that no reading takes place and thus the following computation is skipped. Otherwise the following multi-stage process is being invoked.

- Filtering** The incoming gaze stream  $G = (g_1, g_2, \dots)$ , where  $g_i$  denotes the  $i$ th measured pixel position, first filtered by two independent 5-stage median filters. One is applied to the  $x$ -axis, one to the  $y$ -axis, and the resulting virtual median is considered as the current gaze position. On top of the gaze position a (100 ms, 50 px)-dispersion based fixation detection is being performed, resulting in a stream of fixations  $F' = (f'_1, f'_2 \dots)$ .
- Conversion** The fixations  $F'$ , which are being recorded in screen coordinate space, are converted to document coordinate space  $F = (f_1, f_2 \dots)$ . This has the advantage that all elements can later be directly matched to the nearest word bounding box. Also this measure implicitly encodes the user's scrolling behavior.
- Normalization** Based on  $F$  we compute the stream of saccades  $S = (s_1, s_2, \dots)$ , such that  $s_i$  equals the saccade from  $f_i$  to  $f_{i+1}$ . We consider each  $s_i$  in its polar coordinate form, such that  $s_i = (\theta_i, l_i)$ , where  $\theta_i$  equals the angle and  $l_i$  the length, expressed in *virtual character units* (vcu). A vcu, in turn, is the average pixel-width of a single character underneath the saccades of consideration.
- Windowing** Based on  $S$  we consider a window  $w \subset S$  of subsequent saccades, where the size  $|w|$  varies somewhat on the specific use case. According to our findings the most accurate results can be obtained with  $|w| = 3$  (compare Biedert et al. 2012 for details), which gives the reading detection a reaction time of approximately 750 ms.

**Extraction** With  $w$  we next compute the overall saccade shape that manifests itself as two parameters  $h$ , which is the window's angularity, and  $p$ , the average forward speed. For the saccades  $s \in w$ ,  $h$  is being computed as

$$h = \text{atan}2\left(\sum_{s \in w} \begin{pmatrix} |s_x| \\ |s_y| \end{pmatrix}\right)$$

while the forward speed is calculated as

$$p = \beta \cdot \varnothing \left\{ l_i \in w : |\theta_i| < \frac{\pi}{3} \right\}$$

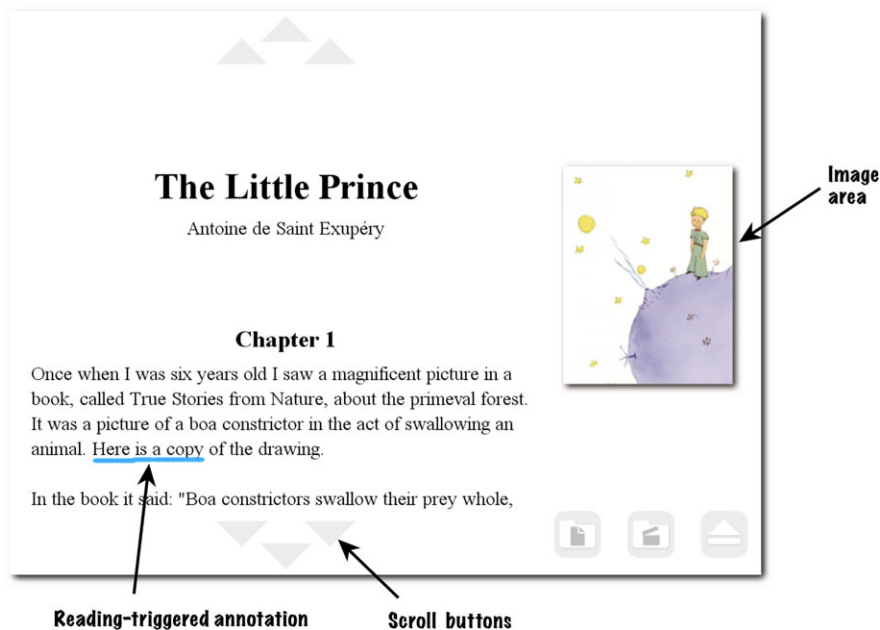
where  $\beta$  is a corrective factor to personalize the reading speed and  $\varnothing$  the *average* operator of the given set.

**Classification** The tuple  $(h, p)$  eventually serves as the input of a classification step. Trained on a set of six users and evaluated on a set of six different users, we found that the linear classifier  $-2.97 + 5.36h + 0.17p$  already gives reasonable results which we could also empirically verify in a number of demos. It returns values  $> 0$  for likely skimming patterns and values  $< 0$  for likely reading patterns. Again, see Biedert et al. (2012) for a more detailed discussion of the emerged classifier and its results.

With the technical measures presented in this section we have the tools to build gaze-responsive applications, which can likewise also store their usage traces for later analysis.

## 8.4 Uni-modal Gaze Interaction

In this section we will investigate how gaze input alone can be used to enhance the reading experience. We present how augmenting texts with manually authored effects, triggered through the user's reading progress, is being perceived by its users. The concept is centered around *Hollywood books*, implemented as the eyeBook (Biedert et al. 2010) prototype, a gaze aware e-book reader. In addition to the story's text it also contains `onRead` annotations that trigger effects such as background music, sound effects, themes or additional images exactly at the moment when the user is reading the corresponding passage, compare Fig. 8.3. The user therefore progresses through the story according to his speed, while at the same time he can enjoy multimedia effects that support the story's progress. Technically the eyeBook is implemented as full-screen Java/HTML book reader application. The main view is split into two areas, a content or text area, which also contains two scrolling buttons, as well as an image area on the right. After a book is loaded the content area renders the book's text like in a normal e-book reader. The visible key difference to *traditional* e-book readers is the presence of two gaze active scrolling buttons. If the user dwells upon them they are being executed within 500 to 1000 ms, depending



**Fig. 8.3** Interface of the eyeBook application. The left part of the screen is dominated by the HTML reading area. The right part contains the image area. The text area renders the story's text and contains the two dwell-based scrolling buttons. When reading is detected and the user gaze passes by an annotated word, its associated effect is executed, such as the playing music or sound effects, displaying an image or changing the theme

on how often they have been triggered before, thus exercising some sort of learning behavior and addressing the tradeoff between perceiving and interaction related to the Midas touch (Jacob 1995). When reading is detected the books also evaluates the annotations linked to the text to trigger certain effects. These include playing background music, sound effects, displaying an image in the right frame or changing the theme.

The key question in the creation of the eyeBook was how the technology is being perceived by its users. Since the application provides visual and acoustic feedback in parallel to the reading process there is likewise a huge potential for cognitive conflicts and distractions, especially if the augmented effects are not integrated well.

### 8.4.1 Experiment

For our experiment we therefore carefully annotated two book implementations according to what we considered esthetic. The source texts we selected for annotation were the chapters of *The Little Prince* as well as excerpts from Jonathan Harker's

diary of the novel *Dracula*. We considered them to be suitable since they contained only few philosophical passages (which are hard to augment with suitable effects) and in contrast more of the protagonist's action and environmental descriptions. The music composition for the Little Prince was selected to be unobtrusive, e.g., *easy listening*, for *Dracula* we also took parts of the cinema's musical theme. The sound effects used for annotation were partially ambient effects such as wind and rain that accompanied the reader during a whole passage, and partially singular effects such as a leaving carriage. Overall each text is about ten screen pages long and contains between one to five annotations per page.

The actual experiment was performed on a book fair. The scenario allowed us to access a broad audience with a wide variety of backgrounds. For three days we randomly asked visitors during *low-traffic* times if they were willing to participate anonymously in a usability study of a novel book concept. If they agreed we explained them the function of the eye tracking system, a Tobii 1750 device, and how the subsequent calibration and experiment was to take place. They were seated in front of the eye tracker facing towards a wall and given head phone, effectively preventing most visual and acoustical distractions. We instructed them to relax and read normally, as if they would read a book, without skipping or skimming, and that they should take as much or as little time as they wanted. The participants were told that they were under no time pressure and that the experimenter would leave them alone. A questionnaire was given with the request to fill it out and put it folded into a drop box. The questionnaire contained 22 items related to the eyeBook and general statistical information. We took care to balance the items as positive and negative statements<sup>7</sup> to counter possible answer biases. All items were rated on a five-point Likert scale ranging from  $-2$  (strong disagreement) to  $+2$  (strong agreement). Eventually the calibration was being performed and the application started.

## 8.4.2 Evaluation

Overall we had 17 visitors participating, while two declined to take part in the survey. Five of our participants were male, 12 female. Their average age was 35.6 years, ranging from 18 to 68, two participants refused to specify their age. Nine of them wore vision aids and all of them reported that the text was clearly visible to them. Also, all of them reported that they believed and were aware of that the survey was conducted anonymously. Regarding their background knowledge all of them reported to be German, with an average of 20.5 years of English experience. The eyeBook version *Dracula* was used by 15 of our readers, while 9 reported that they had read *The Little Prince*. We also investigated some aspects related to the entertainment value of the books, compare Table 8.1. There are a few aspects worth noticing. In general the application was seen very positively. While the majority of

---

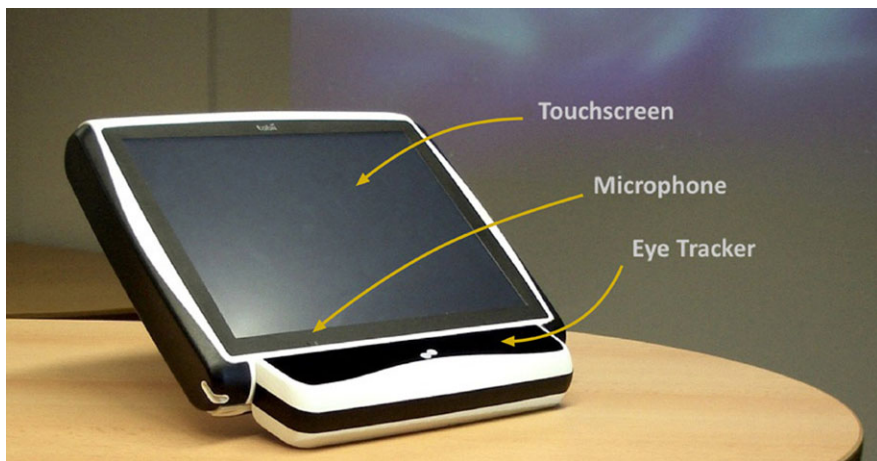
<sup>7</sup>For example "The effects happened at the right time." as a positively formulated statement and "The effects were distracting." as a negatively formulated statement, see Table 8.1.

**Table 8.1** Detailed list with answers of the eyeBook survey. The possible answers ranged from +2 (very strong agreement) over 0 (neutral) to -2 (very strong disagreement)

Question	Average rating
I can remember having read <i>The Little Prince</i> earlier.	1.09
I can remember having read <i>Dracula</i> earlier.	-0.57
The effects were appropriate in time.	1.13
The effects were appropriate in content.	1.38
The effects were distracting.	-0.57
I would like to have disabled the effects after some time.	-0.71
It was unclear to me why an effect was being played.	-1.21
I would like to have had other effects.	-0.71
In the future, I would refrain from buying books with such a technology.	-1.14
The acoustic effects became distracting after some time.	-0.93
The scrolling was pleasant.	1.29
After scrolling I often lost my line.	0.07

the effects was played at the right time (1.13) and also perceived as being semantically appropriate (1.38) apparently the visual effects were considered slightly distracting (there as a not very strong (-0.57) rejection of the statement that they were not). The acoustic effects on the other hand were perceived as better (-0.93) and overall there was a general agreement that the users would want to buy or use similar technology in the future (-1.14). Anecdotally one girl, which did not participate in the survey, started crying while reading *The Little Prince*. However, we did not ask whether it was mainly because of the story or due to our implementation. In general approximately 200 users took the demo during the event, several hundreds for the whole lifetime of the eyeBook application during numerous other events, and the survey results are in line with our general impression of these other demos. The majority of our users reported to enjoy the interaction and the most common cause for problems are inaccurate calibrations and excessive pupil-size-change-based drift.<sup>8</sup> Also, the placement of scroll buttons near the camera region of the used eye tracking device, where the detection of the eyes, precision and accuracy apparently drop notably, has caused scrolling problems for some. From the content perspective especially the music and ambient sounds are reported to be highly regarded, while some of the effects are reported to be somewhat intrusive. In general—not in particular—we got the impression that the more unobtrusive and subtle an effect was, the better it was being perceived.

<sup>8</sup>Especially the implementation of *Dracula* was prone to drift since it contained a theme change from day (white background) to night (black background) which caused the measured gaze position to go off for some of our participants.



**Fig. 8.4** The Tobii C12 device upon which we implemented the eyePad prototype. The *book* reacts to the reader’s gaze, e.g., by updating a visual map while reading the related text. Also it allows the user to write commands and perform ad-hoc speech interaction, like by asking “Who is *that* again?”

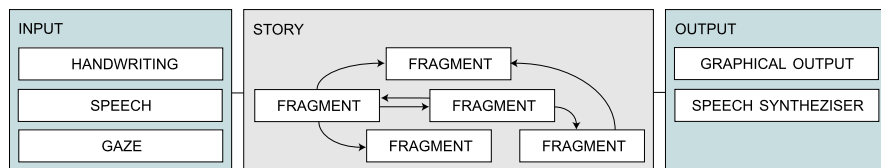
## 8.5 Multimodal Gaze Interaction

In the previous chapter we established that the uni-modal use of gaze can already enhance the reading experience. Motivated by the current trends in interaction techniques we can see a convergence of different input modalities into current application architectures. High precision touch sensitive devices, which can likewise be used for handwriting, are put into cell phones and e-book readers. At the same time speech recognition techniques built into platforms such as Android<sup>9</sup> or iOS<sup>10</sup> devices have seen notable interest, and eye tracking companies like Tobii and SMI are pushing into the market to provide low cost, embedded eye tracking systems. Thus given eye tracking devices become available for the mass market, we argue it will only be a matter of time until they will also be found in book readers. Therefore, in this chapter we will investigate how augmented text, using gaze, handwriting and speech interaction can be used to enhance the reading experience, as well as the obstacles portable devices might cause.

We present the application *eyePad* (see Fig. 8.4), which as a successor to the *eyeBook* also uses gaze as one of its inputs. In addition we added two new input modules interfacing with the touch screen and microphone. They allow the user, besides the implicit gaze interaction to explicitly write commands or commence in an ad-hoc conversation on a certain topic he is currently looking at.

<sup>9</sup>See <http://www.android.com/>.

<sup>10</sup>See <http://www.apple.com/ios/>.



**Fig. 8.5** Internal architecture of the eyePad. At each time a certain *story fragment* (similar to a chapter, topic or scenario in a book) is being displayed on the screen which with the reader can interact. Depending on the user's input the book will either change to another fragment or provide additional information about the current fragment

### 8.5.1 Architecture

From a hardware perspective the eyePad is implemented on a Tobii C12 eye tracking tablet.<sup>11</sup> The unit contains a 12" resistive touch screen with  $1024 \times 768$  pixel resolution, microphone and speakers and an Intel Core Duo U2500, 1.2 GHz CPU and 2 GB RAM for processing. The eye tracking module allows for 60 cm remote eye tracking with a  $40 \text{ cm} \times 30 \text{ cm} \times 20 \text{ cm}$  tracking box and a reported  $0.5^\circ$  accuracy. The overall device weight is approximately 3.3 kg.

From an architectural perspective there are two key differences between the eyeBook and the eyePad. The eyeBook is in principal a reader that displays a linear story at a given time. Although the user is free to read at any speed and skip or skim text this is not intended usage behavior and the reading experience is likely to degrade if the user does so since parts of the story will be missed. Also, the user is normally not actively interacting with the book, but rather progresses through the story while the book implicitly adds additional effects.

The eyePad on the other hand targets a scenario with a high degree of interaction, in which the user also possesses the freedom to progress throughout the story according to his interests. While there is a single starting point the actual passage through the book's story is nonlinear. Also the means by which the reader progresses is to some extent open. Parts of the story and interactions are reached or triggered by mere reading, while others are responses to handwritten input or spoken commands or questions. This is also reflected in the more complex detailed architecture of the application, see Fig. 8.5.

The application consists of three main modules, related to the story, the input and the output. Like the eyeBook it is written in Java and interfaces with the system IO facilities to perform low level input and output operations. Since the platform's performance is very limited home tricks had to be applied however. Instead of integrating the full-fledged HTML setup as described in Sect. 8.3.1 into the application we re-implemented a simpler renderer and layout module to provide us with the bounding boxes. While the gaze integration is straightforward and similar to the

<sup>11</sup>See <http://www.tobii.com/en/eye-tracking-integration/global/products-serviceshardware/tobii-c12-eye-tablet/>.



eyeBook implementation as described in Sect. 8.4, some of the modules required special consideration about which we will give an overview below.

### 8.5.1.1 Story Module

The application's core is the story module that *plays* an eyePad book and provides the high-level logic. A book is a bundle that consists of a number of *fragments*, each of which represents a part of the story the book is supposed to tell. Besides text, these fragments also contain fonts, images, sounds and dialog information, as well as pointers to other fragments which can be reached through some form of interaction. In addition, variables may be kept between fragments so that the interaction within one fragment might affect the status or behavior of others. In game development terms each book could be considered as a mixture between a (textual) game book<sup>12</sup> and visual point-and-click adventures lifted to multimodal reading interaction. The individual fragments can also be programmed with fragment-dependent logic, as well as there is a general database with background information about various topics which can be queried by the user. As with many dialog systems usually there are multiple ways the user can ask about a certain topic to receive an answer. For example “*I want to know what X did.*”, “*What did X do?*”, or sometimes even simply “*X*”, would be valid questions to address the topic X, and the book engine responds by picking a suitable answer for it. Similar to grammars used for speech recognition we also implemented an extension to allow for a JSGF<sup>13</sup>-based dialog output synthesis. With a very limited amount of work this allows for a huge amount of possible output combinations. For example a simple greeting encoded in the grammatical form [It is][nice|good] to see you [again|today|this day]. would already generate 12 possible responses the system could produce, dramatically increasing the diversity of answers.

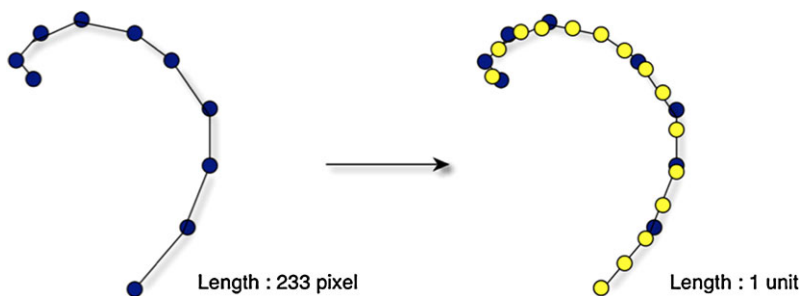
### 8.5.1.2 Handwriting Recognition

Besides gaze input the handwriting recognition is our second input modality. In contrast to touch-based input it allows for a greater freedom of input, at the expense of computational ability to reason on or react to the space of all possible inputs. As the processing power of the C12 unit is very limited we implemented a lightweight handwriting recognition that fulfills the needs of the fragments for the planned evaluation. It basically consists of a two-stage process with the phases *transcription* and *matching*. In the *transcription phase* we recognize strokes as a series of coherent pen measurements and extract raw gestures as a series of pixel coordinates  $r' = (p'_i, p'_{i+1}, \dots)$ . The stroke is successively normalized and a series of relative

---

<sup>12</sup>See <http://en.wikipedia.org/wiki/Gamebook> for an overview on the topic.

<sup>13</sup>Java Speech Grammar Format, see <http://www.w3.org/TR/jsgf/>.



**Fig. 8.6** The raw pixel positions of a consecutive stroke are converted into a normalized form, resulting in 40 equidistant points in polar coordinate form. These are then classified as single characters using a trained SVM. The characters are concatenated and matched against a built-in library of expected commands

points is being extracted, similar to the normalization step described in Sect. 8.3.3 where absolute lengths are being converted to relative lengths and angles, resulting in a normalized gesture  $r$ , which consists of a fixed set of points in polar coordinates form  $(\theta_i, d_i, \theta_{i+1}, d_{i+1}, \dots)$ , as can be seen in Fig. 8.6. Each vector  $r$  is eventually classified with a trained RBF-SVM into a single letter. The *matching phase* then concatenates all detected letters into a string, which is eventually matched against a built-in database of all available commands. The matching is performed by computing the edit distance between all possible candidates. Edit distances above a certain threshold empirically determined in a pre-test are generally rejected, and the minimal edit distance still accepted is considered as the user's input.

### 8.5.1.3 Speech Recognition and Synthesis

As with the handwriting recognition, the speech recognition and synthesis showed to be somewhat problematic in terms of available performance, and available voices. Both problems could be addressed by outsourcing the speech synthesis to a remotely connected computer to which the speech module could submit phrases generated by the dialog system that were subsequently returned as synthesized audio files. These files then merely had to be played. The speech recognition was performed by Windows' built-in speech recognition system. For each state in the dialog the set of all expected user utterances were grouped and passed to the recognizer that, in turn, called back the application when something was detected. Since the device was hand-held we could also notice that there was quite some *interaction* noise, either from handling the device, and also the pen interaction often caused false positives for the speech recognizer. We therefore designated an embedded device button as a push-to-talk button.

### 8.5.2 Experiment

In the previous evaluation we could quantitatively show that using gaze to provide a rich multi multimedia experience is well perceived. While for the uni-modal eye-Book such an evaluation is already highly influenced by the specific story being used, this matter is even more complex in the multimodal setup of the eyePad. For the purpose of this evaluation we therefore focus primarily on a qualitative analysis of confluence of the presented methods—in other words: what inherent challenges would someone face (commercially) producing such a product. Our key question therefore are how the interaction with the device is being perceived, what issues arise in the fusion process and how the tracking performance is being affected due to the movement and interaction with the handwriting part.

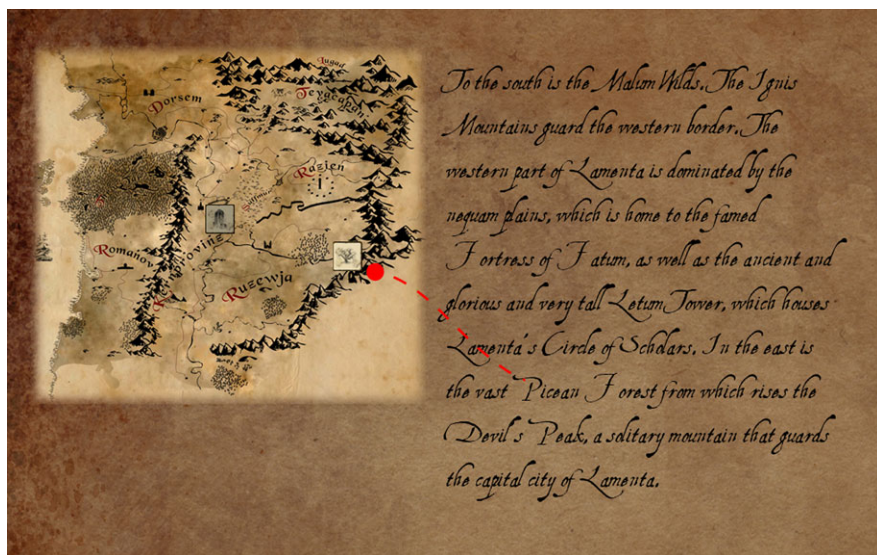
In lack of adequate source material for such a technology we created a demo book from scratch. Inspired by the works *Harry Potter*, *Into the Wild*, and *Dragon Age* we created a mashup with ideas from these works. It consists of a number of fragments which narrate the story of a character traveling into the wilderness whose companion the book has been. When being taken into the hand by the user and gaze is detected the book starts to initiate a conversation by fading in handwritten text. The texts and book's utterances were composed in a way that they revealed clues to the human on how to continue the conversation. After the introduction the book would provide the reader with:

I'm Aedan's memory, do you want to hear the story of Ferelden? Or perhaps I should tell you about the darkspawn . . . If this is your first time you should probably ask me how do I work.

To this the reader could then respond either verbally or in written form asking how the book worked, or about the places it mentioned. For our experiment our participants were also asked to particularly explore a number of special fragments which in addition provided novel interaction paradigms, most notably an *interactive map fragment*. In addition, the text as part of the diary also a map is presented on the left side of the screen, compare Fig. 8.7. When the user reads about a certain location, an icon on the map will slightly fade in and when the user shifts his attention to the map it is highlighted fully.

### 8.5.3 Evaluation

For the evaluation we invited eight students, four of them male, four of them female. Their average age was 21.5 years and all of them were engineering or computer science students. The device was presented and they were told to participate in a usability study. We asked them to interact with the device extensively and told them to perform a think-aloud evaluation. Afterwards we conducted a training run on the handwriting recognition and calibrated the tracking device. Eventually the book was handed over to them and they could interact with it freely. The overall time for the



**Fig. 8.7** One of the special fragments. While the user reads the text on the right side the map on the left is visually fades in small icons reflecting the currently considered location

completion of the experiment was 30 minutes, including calibration and training. The time spent on the interaction was approximately 15–20 minutes. Similar to the eyeBook evaluation we also asked a number of questions regarding the participant's general impression of the prototype.

The overall results of the survey can be seen in Table 8.2. Most participants agreed that in this scenario eye tracking improved the experience of interaction (1.38); as well as it helped to visualize and imagine locations read in the text (1.88). Handwriting and speech input were also perceived as favorable but with less strength (0.75). When investigating how well the participants thought the individual technologies were integrated, eye tracking again was rated highest in terms of how often errors or glitches in the interaction were being perceived (1.13) with some distance to the other two modalities (0.86). We believe this is mainly an indicator that in the specific scenarios we implemented eye tracking was the best suited interaction method, especially with respect to the method's accessibility. While the gaze-reliant parts were mostly straightforward to use, handwriting and speech allowed for a much greater space of possible inputs, and in return disappointment when certain questions or commands were not understood or could not properly be answered.

While most of the participants enjoyed the interaction in general we also received a number of other remarks. One participant said he was *annoyed* by the handwriting interaction and would have preferred to tap or click on keywords in the text instead of having to write something. Also, another participant would have liked the handwriting to be faster in terms of interaction speed, since typing versus talking was considerably slower. We could also notice that handwriting requires the user to hold

**Table 8.2** Detailed list with answers of the eyePad survey for our eight participants. The possible answers ranged from +2 (very strong agreement) over 0 (neutral) to -2 (very strong disagreement)

Question	Average ( $\mu$ )	$\sigma^2$
Integrating the eye tracker enhances the reading experience.	1.38	1.41
Integrating handwriting enhances the reading experience.	0.75	1.04
Integrating the speech interaction enhances the reading experience.	0.75	0.46
Overall the eye tracking quality was very good.	1.13	0.83
Overall the handwriting recognition was very good.	0.86	1.36
Overall the speech recognition was very good.	0.86	0.99
The quality of the speech synthesis was very good.	0.5	0.76
I had the feeling the diary was able to respond to what I say.	0.75	0.89
The map application helps to visualize the geographical subcontext.	1.88	0.35

the unit differently in order to free one hand for a pen which makes it, in contrast to speech or gaze, uncomfortable to use for spontaneous interaction. While the speech synthesis was explicitly criticized by only one participant as *hard to understand* there was a number of issues with the speech recognition overall (0.5). A few participants also forgot to press the push-to-talk key or expected casual utterances to be recognized as well. The gaze interaction was most problematic for one participant with high prescription numbers, which caused problems during calibration and tracking. It was also remarked by one user that had liked to move more during the interaction and change their body posture.

## 8.6 Considering Emotions

In the previous two chapters we presented prototypes which target at enhancing the reading *experience*, i.e., making reading and text interaction more entertaining. To evaluate these prototypes we mostly made use of explicit user feedback questionnaires and our own observations. This is even ahead of the process how texts are being produced in today, where often only anecdotal usage feedback is being provided to the creators of a work. Obviously there are a number of problems with these methods. Through mere study observations tend to be biased and users tend to give socially acceptable (*social desirability bias*) or anticipated answers (*experimenter effect*) when reporting back. Also, the presentation of new concepts, media or technologies is known to elicit a positive feedback initially, even though in the long term other results would be obtained (*novelty effect*). In this section we therefore investigate how live manifested usage experience—emotions—during reading can be acquired objectively, how they can be matched to the text and stored for later analysis, as well as being used in real time.



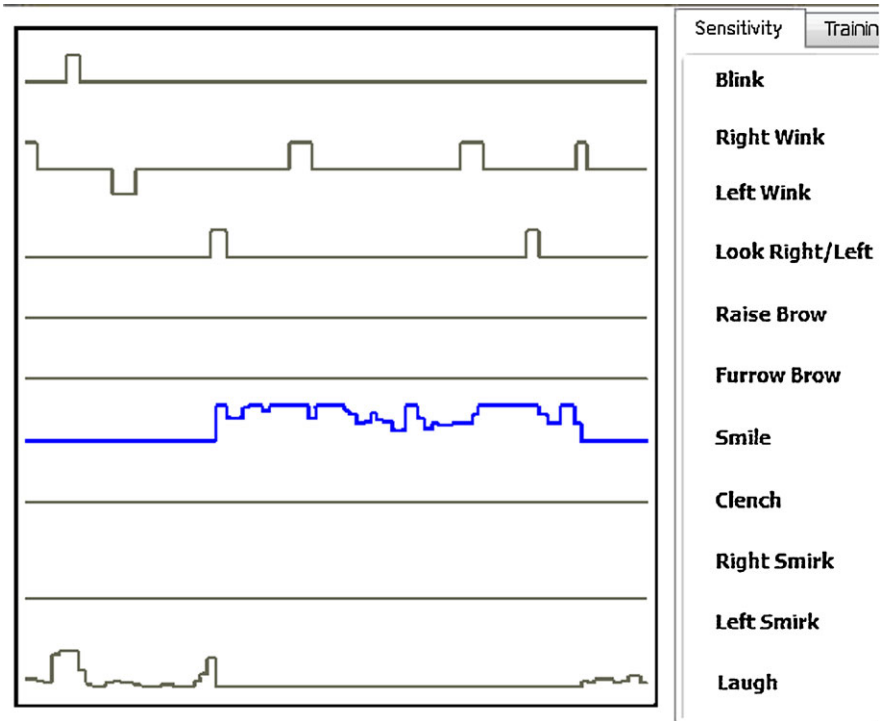
**Fig. 8.8** For the emotion measurement we make use of a low cost EEG device. While the user works on the eye tracker muscular activity and brain waves are also recorded and stored with the gaze data for real-time reaction and later analysis

### 8.6.1 Emotion Measurement

There are many definitions of what emotions are (Kleinginna and Kleinginna 1981), and in this work we will focus only on a small subset. Since our main task is not proposing novel ways to detect emotions but rather how they can be integrated we limit ourselves to four emotions we considered pivotal in human-text interaction. These are *joy*, *boredom*, *interest* and *doubt*, in addition to a neutral state. We consider joy and doubt to manifest primarily through muscular activity, such as smiling and frowning. Interest and boredom, while they also should be perceivable externally, are likely to be measured best cognitively. Hence, as our primary source of emotion detection we make use of a low cost EEG headset produced by Emotiv.<sup>14</sup> It is a low cost EEG measurement device,<sup>15</sup> that features 14 saline sensors at the regions AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, compare Fig. 8.8. Internally the device samples with 2048 Hz and provides data with 128 Hz resolution from the given locations. While the device's capabilities of measuring actual brain signals are somewhat limited compared to full scale EEGs (Ekanayake 2010; Campbell et al. 2010; Hoffmann 2010) it is capable of measuring facial muscular activity. For the purpose of measuring emotions on text we rely on two of Emotiv's APIs which deliver low level muscular activity (*expressiv[sic] suite*) and

<sup>14</sup>See <http://www.emotiv.com>.

<sup>15</sup>For specification details see <http://emotiv.com/upload/manual/sdk/Research%20Edition%20SDK.pdf>.



**Fig. 8.9** The *expressiv suite* channels provided by the Emotiv API. For our muscular based detection of emotions we rely on the channels smile, furrow and laugh

engagement levels (*affectiv suite*, based on the intensity of alpha and beta band) upon which we built an additional classification and mapping scheme.

Since our targeted emotions do not directly map to the device's output (compare Fig. 8.9) we introduce a mapping which allows us to convert the device's output into our four required classes. The subject first needs to train her emotions using the provided Emotiv tools, afterwards we perform a subsequent internal training run on prepared texts that were previously classified as falling into one of the emotional classes. Eventually we acquire individual thresholds for each subject and each of the emotions to discriminate when they were actually evoked.

### 8.6.2 Tagging and Interaction

To be able to use the detected emotions in real time or for later analysis they need to be matched to the text and stored. Since due to eye tracking inaccuracies and user behavior such as saccades and re-reading we cannot just assign measured emotions to the word the user is fixating. Instead we employ a multi-stage process to filter and



match the reading data accordingly. Since we are interested in the evoked emotions during the first reading pass reading incoming gaze data is filtered for reading. During phases when no reading is detected also no emotions will be assigned to the text, for example when the user is skimming or jumping back and forth. The next issue are saccades and the fact that even during reading not every word is in fact fixated. Hence, for every saccade we extract a slice of words between the current fixation and the previous fixation and assign all emotions measured during the time the slice. Technically the assignment is being performed by updating the words' DOM elements with information about the most prevalent emotion measured. At the same time this information is also written into the meta information layer of the recording stream (Sect. 8.3.1) for later analysis.

In addition the measured emotions are made available to the JavaScript layer to support new emotional event handlers. In the same way as the handling facilitates described previously (Sect. 8.3.2) we introduce an emotional set of handlers which not only take into account the gaze data, but also the emotional state of the user.

- `onSmile`      When the user is looking at a certain element and the emotion *joy* is being detected for the first time the attribute's associated script is being executed.
- `onInterest`   Similar to `onSmile` the script is being executed when interest was measured. In contrast to muscular activity like smiling the cognitive engagement level associated to interest usually takes longer to rise or decline.
- `onFurrow`     Also mostly triggered by a muscular reaction the associated script is evaluated when furrowing was detected in the consideration area annotated with this attribute.
- `onBoredom`    Like the `onInterest` rather the results of a cognitive measurement, triggered when *boredom* was measured through the *affectiv API*.

### 8.6.3 Experiment

In order to evaluate the overall performance of the system with respect to how closely the tagged text parts match the user's evoked emotions we prepared a reading experiment. We prepared a number of articles from web sites such as [reuters.com](http://reuters.com), [slashdot.org](http://slashdot.org) and [dailyme.com](http://dailyme.com), which mostly were already pre-rated into categories such as mostly *funny* or *interesting* through the readers of the according web site. For the experiment we invited nine users, four female, five of them male. They were introduced to the setup and told to participate in a reading experiment. We then calibrated them with the Emotiv headset on training texts and performed the eye tracking calibration on a Tobii X120 unit. Each of them was presented five documents which they were supposed to read. After they completed all documents the texts were presented to them again and they could manually tag the texts with four different markers related to the actually evoked emotions, which then served as the ground truth for our subsequent evaluation.



**Table 8.3** Precision, recall and F-measures for the four emotions and the neutral state for the tagging experiment

Emotion	Precision	Recall	F-measure
Joy	0.74	0.93	0.82
Doubt	0.54	0.93	0.68
Interest	0.85	0.72	0.78
Boredom	0.67	0.13	0.22
Neutral	0.56	0.41	0.47

### 8.6.4 Evaluation

After an initial observation of the eye tracking data we noticed we had to discard 13 out of 45 read documents due to missing or bad eye tracking data, resulting in 32 tagged texts left for evaluation. The emotions joy and doubt were evaluated on a sentence level, i.e. if a word was tagged with joy, although this particular word did not evoke this feeling, but another one in the same or adjacent sentence did, then it was considered as correctly classified and tagged. The neighboring sentences were allowed due to the fact that the emotions might be shifted because of the skipping or skimming of words.

The emotions interest, boredom and neutral on the other hand were evaluated on a paragraph level. This distinction was made because of the difference in the nature of the neuroheadset's signals. Joy and doubt depend on muscular movements represented by pulses and are usually instantaneously detected. The detection of interest and boredom is based on a continuous EEG data signal and it needs time to rise and fall with the reader's mood. Thus, since changes are not instantaneously detected, we agreed on a range of a paragraph which would provide enough time for the signal to stabilize itself and give correct feedback about the current emotional state of the user.

The results of this evaluation can be seen in Table 8.3. While the rather expressive emotions joy and doubt were often detected when they occurred, boredom was almost imperceptible. The most common cause for misclassification of joy and doubt were unintentional facial movements by the readers. This included moving lips while reading or furrowing the forehead when being highly concentrated.

In addition to these measured numbers we also performed a survey of how the system was perceived. Although all said that the presented choice of emotions was useful to start with, most of them commented that it was difficult for themselves to differentiate between *doubt* and *interest*. Either because the difference between them was too small, or because they found that the emotional state *doubt* was already included in *interest*. Interestingly, all male students suggested further to add *anger* or *frustration* to the emotion selection. Regarding the necessity of a neutral emotional state, two students claimed that neutrality did not exist while reading.

According to some of our participants' further comments, text should always be assigned to an explicit emotion, i.e. it is either boring or interesting but never

the only person who uses his computer mainly for the purpose of diddling with his computer. Dave Barry \*\*\* I do not fear computers. I fear the lack of them. Isaac Asimov \*\*\* I think computer viruses should count as life. I think it says something about human nature that the only form of life we have created so far is purely destructive. We've created life in our own image. Stephen Hawking \*\*\* I think it's fair to say that personal computers have become the most empowering tool we've ever created. They're tools of communication, they're tools of creativity, and they can be shaped by their user. Bill Gates

**Fig. 8.10** Sample output of automatically tagged text by using the emotions measured by the Emotiv device

neutral for example. The remaining five students on the other hand indicated that the neutral state did exist, that in situations where they were just reading to complete certain text parts in order to continue no explicit emotions were evoked, thus, were said to be neutral. But it was also mentioned that sometimes the neutral state could be interpreted as boredom. When asked about the emotions they thought would be useful in applications such as searching all agreed on the practicality of the emotions *interest* and *joy* and the redundancy of having *doubt* and *boredom*, but the need for all of them in other applications such as article rating. Finally, it was stated that in general the combination of emotions and text in real time had augmented the reading experience and that it had provided the users with a new understanding, reflecting their interaction with the text. In addition, they remarked that this idea could be used in future implementations of web applications for on-the-fly rating and advertising. A sample output of an automatically tagged text can be seen in Fig. 8.10.

## 8.7 Conclusion and Outlook

In this paper we presented various ways how eye tracking can be used to enhance the reading experience, and how the readers' emotions and interactions can be acquired for later analysis. We started by outlining and evaluating the eyeBook, a gaze aware e-book reader which plays music, sounds and images according to the user's progress throughout the story. We could show that augmenting the text with additional, especially ambient, effects was in general well received and that there was a significant user interest in the technology. We continued our investigation with a multimodal prototype, that in addition to gaze also included a speech- and handwriting recognition. While in this scenario eye tracking again blended in nicely, there were some issues with the more *open* input technologies. Although both were well received likewise, it appears that they suffered somewhat from the fact that some of the things said or entered could not be answered or responded to. We think integrating complex dialog management (Holzapfel 2008) and the general strategies involved to overcome communication problems in human-robot interaction (Holzapfel and Gieselmann 2004) could make at least the verbal book interaction more dynamic and responsive. Handwriting, in contrast, we found was by far the slowest interaction and it took a considerable amount of time to enter commands instead of saying them (or, preferably clicking as it was reported). Eventually we

presented a way to record and use emotions elicited during human text interaction. By using a low cost EEG we are able to measure muscular and to some extent brain activity to identify the emotions interest, boredom, joy and doubt. We provide an architecture to annotate the text with these emotions and introduce a set of four handlers to enable the text to react to them in real time.

**Acknowledgements** The authors like to thank Farida Ismail, who implemented many parts of the Emotional Text Tagging prototype and supervised the experiment. She also contributed to parts of Sect. 8.6.4. Our gratitude also goes to Mostfa El Hosseiny who put lots of effort into programming the eyePad demo. He also came up with the book's multimodal story and put lots of energy into designing the demo and bringing the modern Tom Riddle to life.

## References

- Anonymous (2007) To read or not to read: a question of national consequence. Tech. rep., National Endowment for the Arts, Washington. <http://www.arts.gov>
- Biedert R, Buscher G, Dengel A (2010) The eyeBook—using eye tracking to enhance the reading experience. *Inform-Spektrum* 33(3):272–281. doi:10.1007/s00287-009-0381-2
- Biedert R, Buscher G, Schwarz S, Hees J, Dengel A (2010) Text 2.0. In: Extended abstracts of the proceedings of the 28th international conference on human factors in computing systems (CHI EA'10), pp 4003–4008. <http://portal.acm.org/citation.cfm?id=1753846.1754093>
- Biedert R, Buscher G, Hees J, Dengel A (2012) A robust realtime reading-skimming classifier. In: Proceedings of the 2012 symposium on eye-tracking research & applications
- Buscher G (2010) Attention-based information retrieval. PhD thesis, University Kaiserslautern, Kaiserslautern
- Campbell A, Choudhury T, Hu S, Lu H, Mukerjee M, Rabbi M, Raizada R (2010) NeuroPhone: brain-mobile phone interface using a wireless EEG headset. In: Proceedings of the second ACM SIGCOMM workshop on networking, systems, and applications on mobile handhelds (MobiHeld'10). ACM, New York
- Eisenberg A (2011) Pointing with your eyes, to give the mouse a break. <http://www.nytimes.com/2011/03/27/business/27novel.html>
- Ekanayake H (2010) P300 and Emotiv EPOC: does Emotiv EPOC capture real EEG?, pp 1–16. <http://www.ucsc.cmb.ac.lk/People/hbe/eeg/P300nEmotiv.pdf>
- Hoffmann A (2010) EEG signal processing and Emotiv's neuro headset. Technische Universität Darmstadt. <http://data.text20.net/documentation/thesis.emotivsp.pdf>
- Holzapfel H (2008) A dialogue manager for multimodal human-robot interaction and learning of a humanoid robot. *Ind Robot* 35(6):528–535. doi:10.1108/01439910810909529
- Holzapfel H, Gieselmann P (2004) A way out of dead end situations in dialogue systems for human-robot interaction. In: 4th IEEE/RAS international conference on humanoid robots, vol 1, pp 184–195
- Hyrskykari A (2006) Eyes in attentive interfaces: experiences from creating iDict, a gaze-aware reading aid. <http://acta.uta.fi/pdf/951-44-6643-8.pdf>
- Hyrskykari A, Majoranta P, Aaltonen A, Riih   KJ (2000) Design issues of iDICT: a gaze-assisted translation aid. In: Proceedings of the 2000 symposium on eye tracking research & applications (ETRA'00). doi:10.1145/355017.355019
- Jacob RJK (1995) Eye tracking in advanced interface design. In: Barfield W, Furness TA (eds) Virtual environments and advanced interface design. Oxford University Press, New York, pp 258–288
- Kleinginna P, Kleinginna A (1981) A categorized list of emotion definitions, with suggestions for a consensual definition. *Motiv Emot* 5(4):345–379

- Miller CC, Bosman J (2011) E-books outsell print books at Amazon. New York Times, p B2. <http://www.nytimes.com/2011/05/20/technology/20amazon.html>
- Milliot J (2011) Fiction rules e-books. <http://www.publishersweekly.com/pw/by-topic/digital/content-and-e-books/article/47660-fiction-rules-e-books.html>
- Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychol Bull* 124(3):372–422
- Starker I, Bolt RA (1990) A gaze-responsive self-disclosing display. In: Proceedings of the SIGCHI conference on human factors in computing systems: empowering people, pp 3–10
- Wade NJ, Tatler BW (2009) Did Javal measure eye movements during reading? *J Eye Movement Res* 5:5–7

# Chapter 9

## Natural Gaze Behavior as Input Modality for Human-Computer Interaction

Thomas Bader and Jürgen Beyerer

**Abstract** Natural gaze behavior during human-computer interaction provides valuable information about user's cognitive processes and intentions. Including it as an additional input modality therefore provides great potential to improve human-computer interaction. However, the relation between natural gaze behavior and underlying cognitive processes still is unexplored to a large extent. Additionally, most interaction techniques proposed in recent years which incorporate eye gaze as input modality require the user to consciously diverge from natural gaze behavior in order to trigger certain events. In this paper we present results from two user studies. The first one aims at identifying and characterizing major factors which influence natural gaze behavior during human-computer interaction with a focus on the role of user's mental model about the interactive system. We investigate how natural gaze behavior can be influenced by interaction design and point out implications for usage of gaze as additional modality in gaze-based interfaces. With the second user study we demonstrate how gaze-based intention estimation based on analysis of natural gaze behavior can be used to improve interaction in multi-display environments.

### 9.1 Introduction

In general there are two ways to incorporate eye gaze as an input modality into multimodal human-computer interfaces. The first way forces users to consciously look at certain locations in order to trigger actions. One example for such approaches is

---

The research described in this chapter was conducted during the first author's employment at the Vision and Fusion Laboratory, Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Adenauerring 4, 76131 Karlsruhe, Germany.

T. Bader (✉)  
AGT Group (R&D) GmbH, Hilpertstrasse 35, 64295 Darmstadt, Germany  
e-mail: [thbader@agtinternational.com](mailto:thbader@agtinternational.com)

J. Beyerer  
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB,  
Fraunhoferstrasse 1, 76131 Karlsruhe, Germany  
e-mail: [juergen.beyerer@iosb.fraunhofer.de](mailto:juergen.beyerer@iosb.fraunhofer.de)

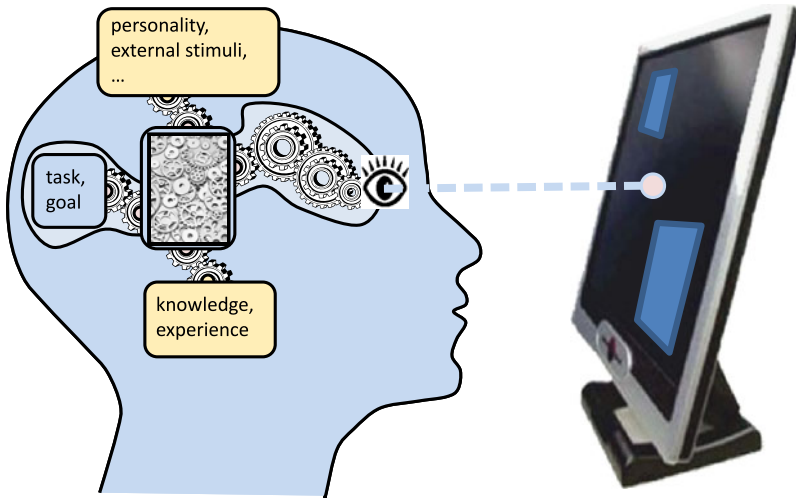
eye typing, which has been studied for decades (Majaranta and Riih  2002). Eye gaze is used directly as pointing device and actions are mostly triggered by dwell times, which determine how long a certain object needs to be looked at until it is activated (e.g., a key on a virtual keyboard). The biggest advantages of such approaches are, that they are easy and straightforward to implement and do not require analysis of complex gaze behavior. Especially for people with severe disabilities such input techniques often provide the only way for interacting with visual interfaces. However, for most people conscious and direct usage of gaze as input modality is very unnatural and hence requires training and/or induces cognitive workload (Jacob and Karn 2003).

The second way to use eye gaze as input modality is to interpret natural gaze behavior during human-computer interaction in the sense of non-command interfaces (Nielsen 1993; Jacob 1993). Also other modalities can be incorporated as primary input modality. Promising examples for such interaction techniques are presented in Hyrskykari et al. (2003) and Zhai et al. (1999). In both approaches *natural gaze* behavior is analyzed and the user is not forced to diverge from that natural behavior for interaction purposes. *iDict* (Hyrskykari et al. 2003) analyzes the duration of fixations while the user reads a text in a foreign language and automatically provides a translation of the fixated word if a longer fixation is detected. In the approach “Manual And Gaze Input Cascaded (MAGIC) Pointing” (Zhai et al. 1999) the mouse pointer is placed close to the currently fixated object in order to eliminate a large portion of the cursor movement. Both approaches do not use gaze directly as pointing or input device, but interpret gaze data in the context of the task (reading, pointing).

In general, the second approach has the advantage that valuable information contained in natural gaze behavior can be used for improving human-computer interaction. Additionally, the user has not to consciously diverge from natural gaze behavior.

However, natural gaze behavior is highly complex and many different influencing factors have to be considered for appropriate interpretation (see Fig. 9.1). Therefore, a thorough understanding of natural gaze behavior during human-computer interaction is necessary in order to incorporate it as input modality in intelligent user interfaces. It has been shown that the task and the experience of users are key factors influencing natural gaze behavior (e.g., in Johansson et al. 2001; Land and McLeod 2000).

Numerous studies of natural gaze behavior and hand-eye coordination during manipulative activities in natural environments like block-copying (Pelz et al. 2001), basic object manipulation (Johansson et al. 2001), driving (Land and Lee 1994) and playing cricket (Land and McLeod 2000) revealed gaze shifts and fixations to be commonly proactive (eye-movements occurred previous to movements of the manipulated object or the manipulator). In addition, a detailed study on hand-eye coordination during an object manipulation task (Johansson et al. 2001) revealed, that subjects almost exclusively fixated landmarks critical for the control of the task and never the moving object or hand. Such landmarks could be obstacles or objects



**Fig. 9.1** Dependency of natural gaze behavior from various factors

in general that are critical for the completion of the task, like in Land and McLeod (2000), where batsmen concentrated on the ball, and not on their hands or the bat. These studies show, that natural gaze behavior is complex and determined by many different parameters (e.g., position of obstacles in Johansson et al. 2001 or previous experience of a person, Land and McLeod 2000).

Gaze behavior was also studied in various tasks related to HCI. In Smith et al. (2000) results of a study on hand-eye coordination during a pointing task with different indirect input devices are described. The main finding of the study is that users used a variety of different hand-eye coordination patterns while moving the cursor to a target on the screen. Also in Bader et al. (2009), where natural gaze behavior was investigated during a direct manipulation task at a large tabletop display, many different gaze behaviors were observed. Other studies from the field of psychology and physiology, e.g. Gesierich et al. (2008), Flanagan and Johansson (2003) investigated differences in gaze behavior during action execution and observation. They distinguished three different gaze behaviors during object manipulation, namely proactive (gaze between object position and target), reactive (gaze between object position and its starting position) and tracking gaze behavior (smooth pursuits) (Gesierich et al. 2008).

In all of the above studies on natural gaze behavior, numerous different gaze patterns were observed during task execution and were described informally. However, an understanding of the reasons why a person looks at a certain location in a certain situation is necessary to judge the usefulness of natural gaze behavior for HCI and to integrate gaze with other modalities, respectively.

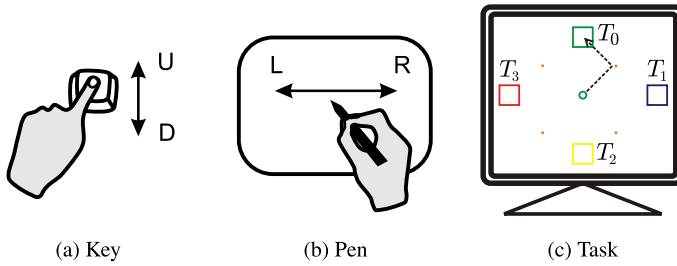


Fig. 9.2 Input devices and task

## 9.2 Influence of User's Mental Model on Natural Gaze Behavior

In this section we report about a study in which we tried to characterize different influences on natural gaze behavior during an object manipulation task. Additionally, we point out their implications for designing gaze-based multimodal interaction techniques for future intelligent user interfaces.

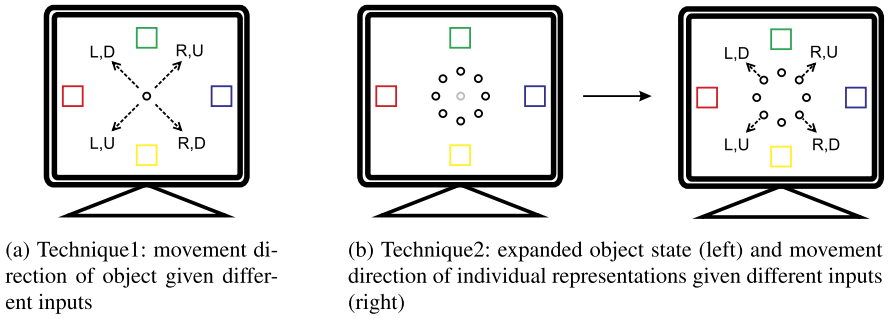
### 9.2.1 Task and Apparatus

The task to be solved by participants is designed based upon a basic object manipulation task as it commonly appears during work with graphical user interfaces. The visual representation of an object has to be moved from one location to another on a display. However, in order to be able to investigate effects of user's mental model on natural gaze behavior in a controlled way, we designed the mapping between input and system reaction in an unusual way not expected by the users. This ensures that all users have the same level of knowledge about the system at the beginning of the experiment and can be considered as novice users. Additionally, we are able to monitor changes in natural gaze behavior with increasing knowledge about the system.

As input devices we use one single key of a keyboard (Fig. 9.2a) and a pen tablet, while only horizontal movements of the pen on the tablet are interpreted by the system (Fig. 9.2b). The task is illustrated in Fig. 9.2c. A colored point which initially is displayed at the center of the display is to be moved to one of the four squares  $T_0, \dots, T_3$  with the same color. Note that the labels  $T_0, \dots, T_3$  shown in Fig. 9.2c were not displayed to the user during the experiment and only serve as reference for the respective target areas within this section.

For manipulating the object position we implemented two different interaction techniques. The mapping between inputs and system state transitions (position of the point) is graphically illustrated for the first technique in Fig. 9.3a. For example, a horizontal movement of the pen to the right (R) causes a movement of the point to the upper right if the key is not pressed (U) and to the lower right if the key is pressed (D). In principle the mapping for the second technique is the same. However, before





**Fig. 9.3** Mapping of input to system actions for different interaction techniques

the object is moved from its initial position, as soon as the pen touches the tablet, its visual representation is split into eight objects arranged on a circle around the initial position, representing possible future object positions (see Fig. 9.3b, left). This representation in the following is denoted as *expanded state* of the object. In order to avoid hints about the true mapping of inputs to movement directions by this representation, objects are also displayed along directions the object cannot be moved to directly (e.g., to the right). However, all eight representations have the same color, namely the color of the target area the object is to be moved into. As soon as the object is in expanded state, a movement of the pen on the tablet leads to a movement of one of the eight object representations into the respective direction, while all other representations disappear. For example, if the pen is moved horizontally to the right (R) and the key is not pressed (U), the object representation in upper right direction is moved to the upper right, while all other objects are faded out (Fig. 9.3b, right).

In order to move the object from its initial position to the (green) target area  $T_0$  at the top of the display along the path illustrated in Fig. 9.2c, for both techniques users first would have to move the pen to the right (R) while leaving the key unpressed (U) and, as soon as the little orange help point is reached, press the key (D) and move the pen to the left. An alternative way to solve the task would be to first move the point to the upper left (input: L, D) and then to the upper right (input: R, U). Users were free to choose the way to the respective target areas during the experiment.

In preliminary experiments with Technique1 we observed that experience of users seems to have significant influence on proactivity of gaze behavior. Novice users, for example, mainly directed visual attention towards the initial object position at the beginning of the task. In contrast, expert users predominantly anticipated future object positions. With Technique2 we wanted to investigate whether it is possible to induce more proactive gaze behavior, especially for novice users, by avoiding visual feedback in proximity to the initial object position right before the first object movement. By explicitly presenting possible future object positions to the user we expected gaze movements to be directed more towards those visual targets than towards the initial object position. As an example this would allow for robust estimation of users' intention from gaze data.

The size of the display is  $33.7 \times 27$  cm with a resolution of  $1280 \times 1024$  pixels. Eye-gaze of the users was captured during task execution by a Tobii 1750 tracking device.

## 9.2.2 Participants

Since we want to investigate effects of mental model building on natural gaze behavior we chose a between-subjects design to avoid any prior knowledge of participants about the task or interaction techniques. We had two groups with 10 participants each. Participants were between 21 and 32 years old and did not know anything about the experiment, except that their gaze is being measured.

## 9.2.3 Procedure

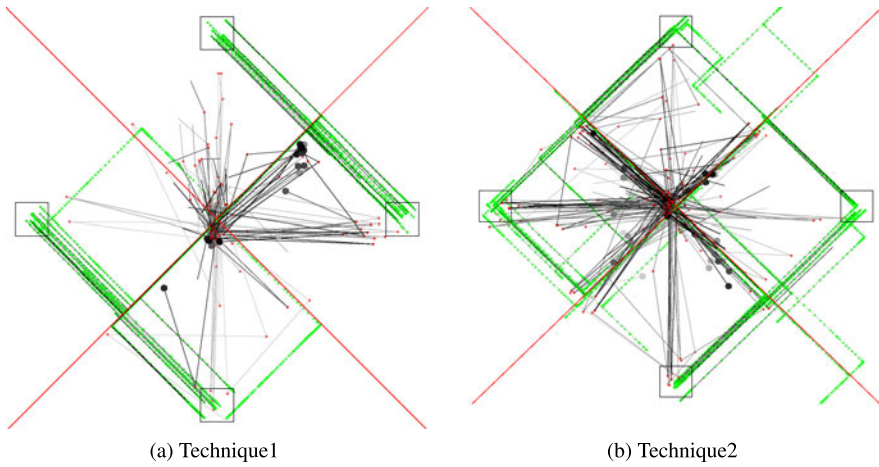
The experiment was organized in two phases A1 and A2 with 40 runs each. Every run consists of moving an object from its initial position at the center of the screen to the respective target area. During both phases of the experiment every color of the object and hence every task occurred 10 times. The order of tasks was randomized to avoid consecutive repetition of the same task. For all participants the task order was the same to allow for direct comparison of performance and gaze behavior and to investigate influence of certain tasks and task order. Except the order of tasks there was no difference between phase A1 and A2.

In order to allow for a more detailed analysis of the temporal development of objective measures in subsequent sections the two phases are further divided into A1/1, A1/2, A2/1 and A2/2 with 20 runs each.

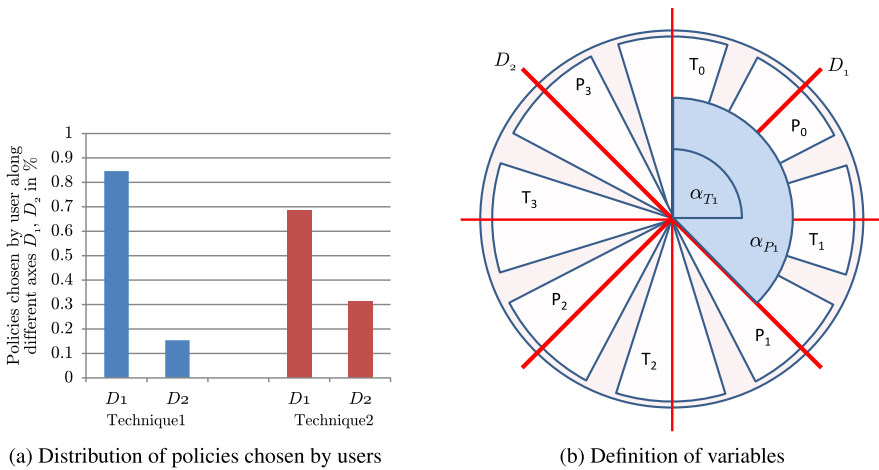
## 9.2.4 Results

Most interesting from the interaction design perspective are gaze movements which occur before any object movement. In the following we denote such gaze data as *pre-object* gaze data and *pre-object fixations*, respectively. Such data allows for estimating users intentions previous to any input made by the user. Therefore in this work we mainly focus on the analysis of such data.

In Fig. 9.4 a plot of object- and gaze-data during the first 40 runs is shown for each of the two interaction techniques for one user. Green dots represent object positions, small red dots connected by gray lines are pre-object fixations and larger dots, colored from gray to black, indicate the last pre-object fixation for each run. The red diagonal lines indicate possible movement directions of the object from its initial position and were not shown to the users during the experiments.

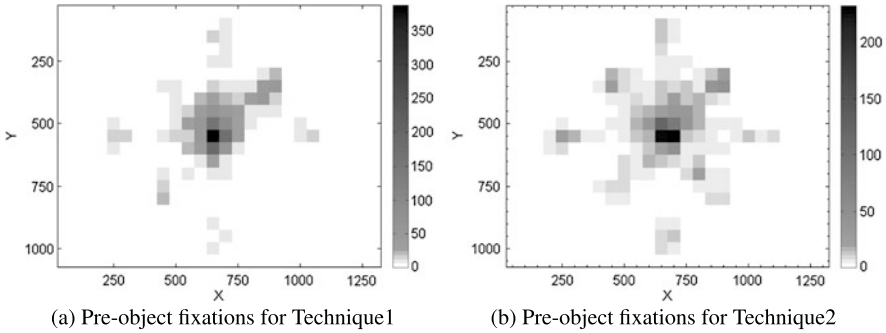


**Fig. 9.4** Data captured for different interaction techniques during phase A1 from one user for each technique



**Fig. 9.5** Task solution strategy of users and definition of variables

For the first interaction technique two things can be easily seen from Fig. 9.4. First, the preferred policy for solving the task seems to be first moving the object along the diagonal line reaching from the lower left to the upper right ( $D_1$ , see Fig. 9.5b for definition). This corresponds to an input sequence where the key is not pressed (U) during the first phase. Second, fixations are mainly located at three different positions on the screen. While the last pre-object fixation is either located at the initial position of the object or along the preferred diagonal axis  $D_1$ , other fixations also can be observed towards or at the target areas.



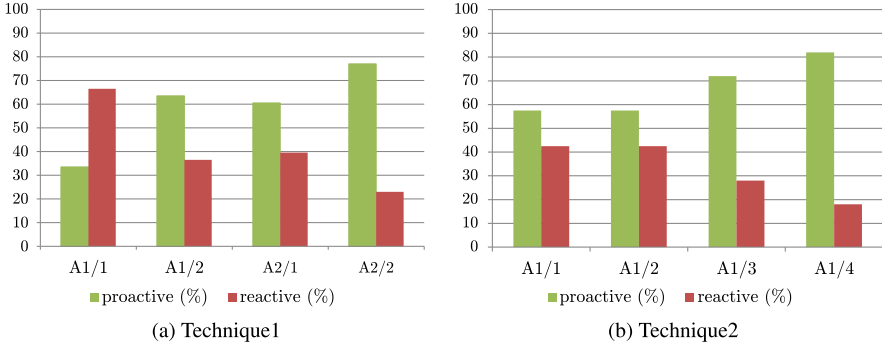
**Fig. 9.6** Distribution of pre-object fixations

Both observations in average can be confirmed for all participants. In Fig. 9.5a the distribution of tasks which were solved by moving the object first along the different axes  $D_1$  and  $D_2$  is shown for both interaction techniques. A clear majority of the users first moved the object along  $D_1$  for both interaction techniques. However, the policies with first movement direction along axis  $D_2$  was used more often for Technique2 (31.5 %) compared to Technique1 (15.38 %).

This difference in interaction behavior also shows an effect on pre-object gaze behavior. Figure 9.6 shows the distribution of positions of all pre-object fixations for all users and tasks for the two interaction techniques. Note that the color scale at the lower end is not linear in order to improve the visibility of the plot. Both plots show that most pre-object fixations are centered around the initial position of the object. However, also a significant amount of fixations can be observed at different locations on the screen which are related to the task. Except from the initial object position for Technique1 fixations are mainly distributed along axis  $D_1$  or at the target areas. The plot for Technique2 in Fig. 9.6b shows also fixations along axis  $D_2$  and in general more proactive fixations. For further task related characterization of fixations we use two features:

- *Distance*  $d$  of a fixation from initial object position
- *Direction*  $\alpha$  of the vector between fixation and object position

Along  $d$ , fixations are classified in *proactive fixations* ( $d > r_p$ ) and *reactive fixations* ( $d \leq r_p$ ). The threshold  $r_p$  defines when a fixation is considered to be on the object (reactive) or not (proactive). While reactive fixations indicate attention allocation towards the current state of the object, proactive fixations are induced by mental planing activity for solving the task or anticipation of future system states. In our experiment we defined  $r_p = 100$  based on the observed distribution of gaze positions when focusing on a certain object. The design of the task additionally allows for distinguishing between fixations which are directed towards one of the target areas and fixation induced by anticipation of the first movement direction of the object by evaluating  $\alpha$ . We further denote the different target areas as  $T_0, \dots, T_3$  in clockwise direction, starting from the top. The different policies users can choose to solve a task are denoted by  $P_0, \dots, P_3$  in clockwise direction according to the



**Fig. 9.7** Development of ratio between proactive and reactive fixations with increasing knowledge about the system

first primary movement direction starting from the top. The definition of  $T_i$  and  $P_i$  is also illustrated in Fig. 9.5b. For example, if the task of moving the object to the upper target area is solved by moving the object first to the upper right (R, U) and then to the upper left (L, D), this corresponds to  $P_0$ . First moving the object to the upper left and then to the upper right for the same task would be  $P_3$ .

Based on these definitions the target of visual attention  $A$  indicated by a fixation can be categorized as follows:

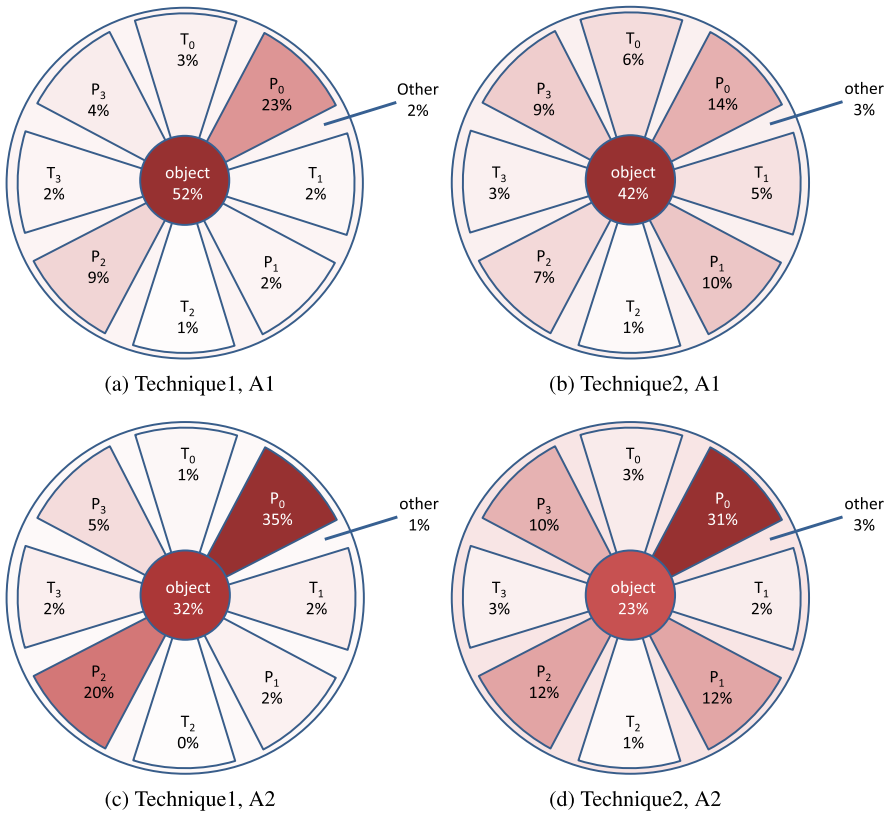
$$A = \begin{cases} T_i & \text{if } |\alpha_{T_i} - \alpha| < \alpha_{max} \\ P_i & \text{if } |\alpha_{P_i} - \alpha| < \alpha_{max} \\ \text{other} & \text{otherwise} \end{cases}$$

where  $\alpha_{T_i}$  and  $\alpha_{P_i}$  denote directions of vectors between the initial object position and the corresponding target  $T_i$  or first movement direction of policy  $P_i$  (see Fig. 9.5b).

The threshold  $\alpha_{max} = 20^\circ$  was chosen based on the analysis of gaze data captured during the experiments.

The development of the ratio of proactive and reactive last pre-object fixations over all phases of the experiment is shown in Fig. 9.7. In average the ratio for Technique1 is about 58/41 (proactive/reactive) and 67/32 for Technique2. For phase A1/1 (first 20 runs) with Technique1 66.5 % of all last pre-object fixations are reactive and 33.5 % are proactive. In contrast, during phase A1/1 with Technique2 57.5 % of the fixations are proactive and 42.5 % reactive. The plots show both, significant influence of growing experience on the location of the last pre-object fixation and significant differences between the two interaction techniques.

As already mentioned above, we further analyze pre-object proactive fixations regarding the underlying target of visual attention  $A$ . Figure 9.8 shows the distribution of  $A$  over all possible targets  $T_0, \dots, T_3, P_0, \dots, P_3$  for all last pre-object fixations. The different areas represent the categories as defined above by  $r_p$  and  $\alpha_{max}$  and are colored according to the occurrence of fixations within the corresponding area on the screen.



**Fig. 9.8** Distribution of target of visual attention of last fixation before first object movement for all users and tasks

For both techniques the number of last pre-object fixations which occur on the object are reduced from phase A1 to phase A2 of the experiment almost to the half. For Technique2 approximately 10 % less fixations are made on the object for both of the two phases compared to Technique1. In all plots among all policies  $P_0, \dots, P_3$  a clear majority of fixations can be found along policy  $P_0$ . While for Technique1 proactive fixations are mainly distributed along axis  $D_1$  (policies  $P_0$  and  $P_2$ ), for Technique2 an almost equal distribution over policies  $P_1$ ,  $P_2$  and  $P_3$  can be observed. This corresponds to findings illustrated in Fig. 9.5a, where similar differences in policies chosen by the users for solving the task are depicted.

## 9.2.5 Discussion

The results in the previous section show that both independent variables we used in the experiment, namely the interaction technique and the experience of users, have significant influence on natural gaze behavior during human-computer interaction.

For both interaction techniques, increasing experience of the user with the system resulted in a highly increased number of proactive fixations with increasing orientation towards policies at the expense of decreasing orientation towards target areas. This development can be explained from an information theoretical perspective. The more knowledge the user has about the dynamics of the system the less new information can be acquired by reactive fixations on the initial object position and by observing the first object movement, respectively. If future expected object positions can be accurately predicted by acquired knowledge, it is more efficient to directly draw visual attention towards expected future object states, e.g., in order to support accurate positioning of the object at the intended target location. The decreasing orientation of visual attention towards target areas can be explained by the same effect. Increasing knowledge of the location of certain target areas decreases the value of directing visual attention towards the target areas.

When comparing gaze data for the different interaction techniques a significantly increased number of proactive fixations and a slight increase in fixations directed towards the target areas can be observed for Technique2. Additionally, while for Technique1 the policies along axis  $D_1$  are predominantly chosen by the users and proactive fixations are mainly distributed along this axis, with Technique2 the policies along axis  $D_2$  are chosen significantly more often and fixations along  $P_1, \dots, P_3$  are almost equally distributed. Obviously, the different ways how visual feedback is organized for the different interaction techniques not only influences natural gaze behavior, but also human decision processes and task solution strategies.

For both interaction techniques and independent from experience of users, by far most of the proactive fixations are made along  $P_0$ . Participants' gaze behavior seems to be more proactive when moving the object from the left to the right than into the opposite direction. Possible explanations for that bias could be found by further examination of influence of writing direction, handedness or other cultural and individual factors.

For designing interaction based on natural gaze behavior the observations above have different implications. Natural gaze behavior is influenced by many different factors. These factors can either be used for adapting human-computer interaction or they prevent the development of consistent interaction techniques due to their dependency from uncontrollable and varying environmental conditions (e.g., experience of users, different cultural background).

In this user study we identified 4 classes of major factors influencing natural gaze behavior during object manipulation and characterized their influence in proactivity and direction of visual attention:

1. task
2. policy
3. experience of users/state of mental model
4. visual feedback/interaction technique

We further identified phenomenons which probably could be explained by individual differences among users and/or cultural factors (e.g., increased proactivity for  $P_0$ ).

The first two factors can be used for estimating user's intention from gaze data. Either the goal of the task or the policy chosen by the user to solve the task can be estimated previous to the first object movement and user input, respectively. However, their visibility in gaze data in the form of proactive fixations towards a certain task related location on the display depends to a large extent on the third factor, namely the state of user's mental model. This fact in principal can be used for estimating user's experience and adaptation of interaction (see, e.g., Bader 2011). However, if the main goal is to design a consistent gaze-based interaction technique for novice and experienced users the goal would be to minimize the influence of experience on natural gaze behavior. According to the results of our study one option would be to use the fourth factor and to design interaction techniques which reduce this influence as we demonstrated it with Technique2. However, as we showed in the results section there still remain variances in natural gaze behavior which probably can be explained by individual differences among users or cultural factors. These factors have also to be considered when interpreting natural gaze behavior and designing appropriate system reactions.

### **9.3 Multimodal Interaction Using Gaze-Based Intention Estimation**

In this section we illustrate by example how the identified factors influencing natural gaze behavior can be used to incorporate eye gaze as an additional modality in multimodal interaction techniques. We especially use proactive gaze behavior and its dependency on the task and goal the user has in mind to estimate user's intention from observed gaze patterns.

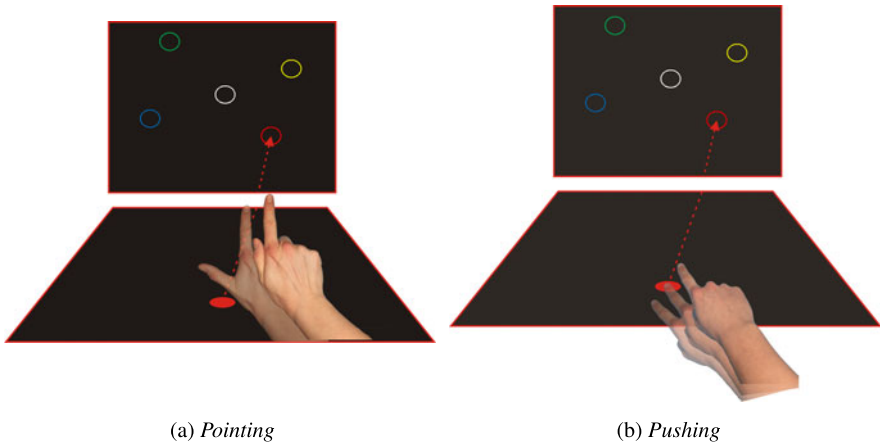
The following section specifies the task considered for the user study which is described in Sect. 9.3.2. The results of the study are discussed in Sect. 9.3.3.

#### ***9.3.1 Task and Interaction Design***

As application domain for our investigations we selected interaction in multi-display environments. Due to increasing number and variety of available displays and devices such environments have gained increasing attention by the international research community in recent years. Early work on that field such as Rekimoto and Saitoh (1999), Streitz et al. (1999) focused on software infrastructure and interaction techniques spanning multiple displays for supporting collaborative co-located teamwork. More recent work with similar research goals can be found, e.g., in Borning et al. (2010), Johanson and Fox (2002), Johanson et al. (2002), Nacenta et al. (2006).

Especially interaction with distant displays could benefit from taking eye gaze as additional input modality into account. Pointing gestures, which are often used





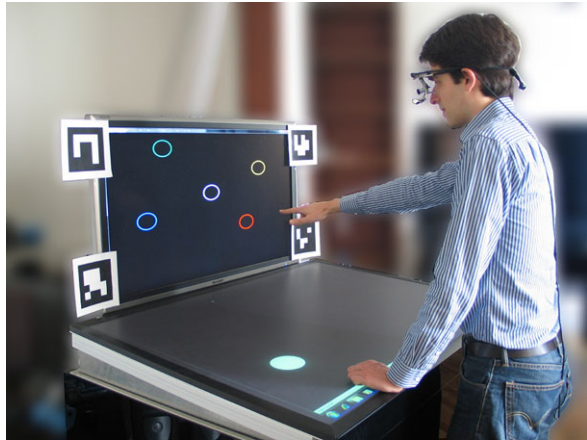
**Fig. 9.9** Two interaction techniques for moving objects from within grasping range to a distant display

for interaction with distant or large displays (see, e.g., Schick et al. 2009), are fatiguing and inaccurate (Nickel and Stiefelhagen 2003). This disadvantage could be compensated for by taking into account gaze as additional input modality.

In the following we consider an object manipulation task spanning multiple displays for further investigations. Objects need to be moved by users from one display to another. In particular, a colored object displayed on a tabletop display in front of the user needs to be moved into the corresponding target area on a vertical display which is mounted behind the tabletop (see Fig. 9.9).

For this task two interaction techniques are investigated. They are illustrated in Fig. 9.9. For the interaction technique *Pointing* the object first needs to be selected on the tabletop display by touching it. The success of the selection is indicated by a slight fade of the object. Afterwards the object can be moved to the vertical display by means of a pointing gesture. The movement of the object on a straight line towards an indicated target area is triggered by changing the hand pose as illustrated in Fig. 9.9a. Previous to the execution of this gesture the object remains at its initial position on the tabletop. The selection for the interaction technique *Pushing* is identical to selection for *Pointing*. The movement of the object towards the target area, however, is triggered by a small pushing movement towards the respective target area. This requires significantly less physical movement of the hand compared to pointing. Hence, we expect less physical fatigue for *Pushing* compared to *Pointing*. However, since the accuracy of the pushing movement is very limited, the target area cannot be indicated through pushing accurately. Also for pointing gestures it is known that they do not always indicate the target position pointed to by the user very accurately (Nickel and Stiefelhagen 2003).

For both techniques in the following we investigate to which extend natural gaze behavior during interaction can be used to accurately determine the desired target position of the object on the vertical display and hence to compensate for disad-

**Fig. 9.10** Experimental setup

vantages of gesture-based interaction (physical fatigue, inaccuracy of pointing and pushing gesture). For recognizing the individual gestures video-based gesture recognition systems like Bader et al. (2009), Schick et al. (2009) or commercial systems like Microsoft Kinect<sup>1</sup> can be used.

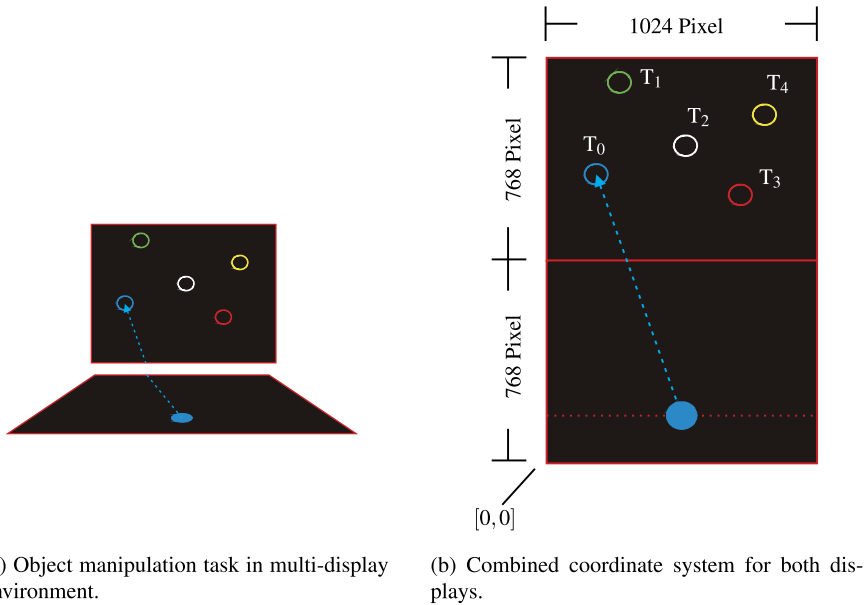
### 9.3.2 User Study

In the following we describe apparatus, task, procedure and results of the user study which was conducted in order to further investigate the role of eye gaze in the scenarios described above.

#### 9.3.2.1 Apparatus

As multi-display environment we used a setup consisting of two displays as illustrated in Fig. 9.10. It consists of a horizontal rear projected display ( $124 \times 89$  cm) and a large scale vertical LCD display ( $102.5 \times 57.5$  cm,  $1024 \times 768$  pixels resolution). For capturing gaze data the mobile eye tracking system Dikablis from Ergoneers is used. Through markers attached to the vertical display the gaze position can be transformed to display coordinates. The distance of participants from the vertical display is approximately 1 m. With a good calibration the accuracy of measurement of gaze position on the vertical display is approximately  $\pm 10$  pixel for a certain object fixated. However, it largely depends on the quality of the calibration of the eye tracker which highly varies for different users. Since the system only can be calibrated to one single surface, the gaze position on the horizontal display can be determined only at a very coarse level. Hence, detailed analysis of gaze data in

<sup>1</sup><http://www.xbox.com/kinect>.



**Fig. 9.11** Task and coordinate system spanning multiple displays

this experiment is limited to the vertical display. For the horizontal display we only distinguish whether eye gaze is located on the display or not.

The eye tracking system delivers data with 25 Hz. We empirically determined a latency of approximately 160 ms. For offline analysis of the captured eye tracking data we compensated this latency by synchronization with other system events.

### 9.3.2.2 Task

For every task in the beginning a colored circle is displayed on the horizontal display which is to be moved to the corresponding target area on the vertical display. In Fig. 9.11a this task is schematically displayed for a blue object ( $T_0$ ). The labels  $T_0, \dots, T_4$  in Fig. 9.11b were not displayed to the users. The center of the target was  $[155, 440]$  for  $T_0$ ,  $[246, 88]$  for  $T_1$ ,  $[498, 330]$  for  $T_2$ ,  $[707, 517]$  for  $T_3$  and  $[799, 210]$  for  $T_4$  target in pixels of the coordinate system of the vertical display.

Some of the target areas on purpose are arranged in a row along the movement direction of the object in order to investigate the influence of this arrangement on natural gaze behavior (see, e.g., target  $T_3$  and  $T_4$ ). The initial object position is varied along the dotted red line in Fig. 9.11b.

### 9.3.2.3 Procedure

The user study is conducted with 16 participants (2 female, 14 male) and is based on within-subject-design. Participants are divided into two groups and the order of

the two interaction techniques is varied among them to balance potential learning effects regarding the task. Every participant executes 80 tasks as described above, 40 task with each interaction technique. Instructions for the different interaction techniques are given prior to every sequence of tasks with the respective techniques through a short video showing the technique. No participant practiced any of the interaction techniques before.

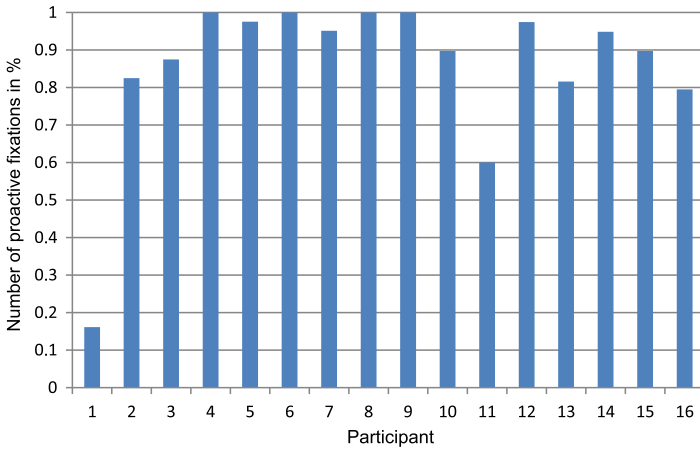
The study is conducted as so-called Wizard-of-Oz experiment. The triggering input (change of hand pose or pushing gesture) is not recognized by the system directly but by the experimenter who manually triggers the respective action (movement of the object into the target area as soon as he observes it). This has the advantage that errors at recognition level of the triggering event are mostly excluded and do not lead to wrong system reactions. For 36 of the 40 tasks for each interaction technique the object moves to the correct target area. In order to investigate the influence of wrong system reaction to users' natural gaze behavior for task 31, 32, 34 and 36 wrong system reactions are triggered and the object moves to a wrong target area.

Between the two interaction techniques and after the experiment users were asked to fill in a questionnaire to capture if system reaction was according to their expectations, if the Wizard-of-Oz experiment was recognized by the users and subjective ratings for accuracy, physical fatigue and satisfaction.

### 9.3.2.4 Results

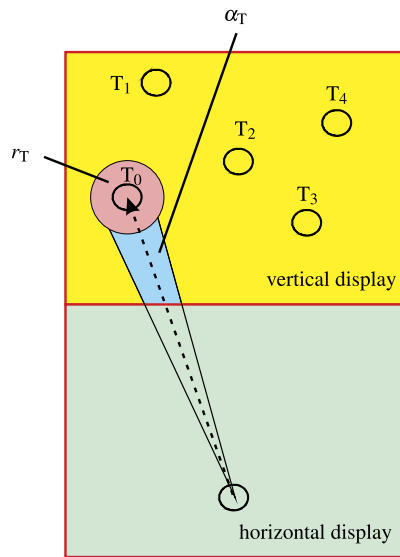
Over all 16 participants and 40 tasks with interaction technique *Pushing* the last fixation before the object movement was triggered was located on the vertical display in 86 % of the cases. For *Pointing* with 98 % this number is even higher. The difference can be explained by the design of the interaction techniques. For *Pointing* the object movement is triggered by the change of the hand pose. The visual attention obviously is directed towards the target area while the pointing gesture is executed by the user and does not have to be turned away for the execution of the triggering hand movement. In contrast, for *Pushing* the triggering gesture is to be performed on the horizontal display. The visual focus of most users is directed towards the target area during this movement as well, however, in the user study especially two users diverged from that behavior significantly. This can be seen from Fig. 9.12, where the number of last fixations before the triggering event which are directed towards the vertical display are shown for all tasks and individual participants. Especially for participants 1 and 11 in the videos recorded by the eye tracking system it can be observed that their visual focus of attention remains on the horizontal display until the object starts moving and then jumps towards the respective target area.

Intention estimates are based on the last fixation before the respective triggering event is detected. They consist of a target area to which the current object most likely is moved according to user's gaze behavior. In order to determine this target area the last fixation before the triggering event is assigned according to its position on the vertical display with respect to the following criteria:



**Fig. 9.12** Proactivity of last pre-input fixation for interaction technique *Pushing*. Participants 1 and 11 significantly diverge from proactive gaze behavior of other participants

**Fig. 9.13** Areas defining the assignment of fixations to target area  $T_0$  according to gaze position on the vertical display. Different assignment areas resulting from the two assignment criteria (distance to center of target area and movement direction of object) are colored differently



- *distance to center of target area*  $r_T$  is less than 140 pixels
- *angle between movement direction* of object and direction of gaze with respect to the initial position of the object  $\alpha_T$  is less than  $5^\circ$

See Fig. 9.13 for an illustration of these assignment criteria. The threshold for the first criterion is chosen so that no double assignment to different target areas is possible based on this criterion. The second criterion, however, allows for assignment of one fixation to multiple target areas which are located along the same line

**Table 9.1** Confusion matrices showing the results of gaze-based intention estimation for interaction techniques *Pushing* (a) and *Pointing* (b)

True target	Estimated target					True target	Estimated target				
	T <sub>0</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	T <sub>4</sub>		T <sub>0</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	T <sub>4</sub>
T <sub>0</sub>	92	7	1	0	0	T <sub>0</sub>	109	7	0	0	0
T <sub>1</sub>	14	71	17	1	1	T <sub>1</sub>	12	100	9	0	0
T <sub>2</sub>	0	0	85	16	3	T <sub>2</sub>	0	1	109	10	1
T <sub>3</sub>	0	0	0	78	16	T <sub>3</sub>	0	0	2	105	14
T <sub>4</sub>	0	3	1	19	62	T <sub>4</sub>	0	3	0	14	76

(a)

(b)

of movement (e.g., for T<sub>3</sub> and T<sub>4</sub>). An estimation is considered as correct if the last fixation is assigned to the correct target area corresponding to the current object based on one of the criteria described above. In the case of an assignment based on the second criterion the correct assignment needs to be among all possible assignments in order to be considered as correct. A correct assignment based on the first criterion superimposes a different assignment according to the second criterion.

In Table 9.1 the confusion matrices for intention estimates for both of the two different interaction techniques are shown. For both techniques the matrices show good results. Precision and recall both are around 80 % for *Pushing* (average over individual measures for T<sub>i</sub>). For *Pointing* 87 % is achieved for both measures. These results can be further improved, especially for *Pushing*, by also taking fixations into account for intention estimation which occur during the movement phase of the object and not only right before the triggering event.

Table 9.1 shows that most of the wrong estimates occur for neighboring target areas or targets lying along the same movement direction (e.g., T<sub>3</sub> and T<sub>4</sub>). Such wrong estimates are mainly caused by inaccuracy of measurements by the eye tracking system and the transformation into display coordinates, respectively. With more accurate measurements the intention estimation is expected to be improved significantly. Additionally, the results in Table 9.1 indicate that intention estimates for targets lying along the same movement direction such as T<sub>3</sub> and T<sub>4</sub> are less accurate compared to others.

### 9.3.2.5 Subjective Data

In addition to the eye tracking data subjective impressions of participants were collected via questionnaire. Results of general questions are shown in Fig. 9.14.

For both interaction techniques in average four participants indicated a system reaction which does not meet their expectations. As reasons all of them mentioned the wrong system reaction which was implemented on purpose at the end of the trial in order to investigate users' behavior in such cases. Hence, these indications of unexpected system behavior is not related to the interaction techniques investigated

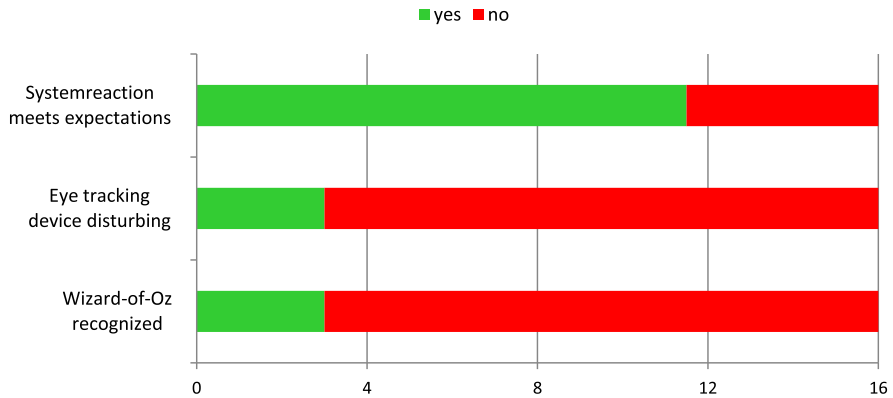


Fig. 9.14 Results of general questions in questionnaire

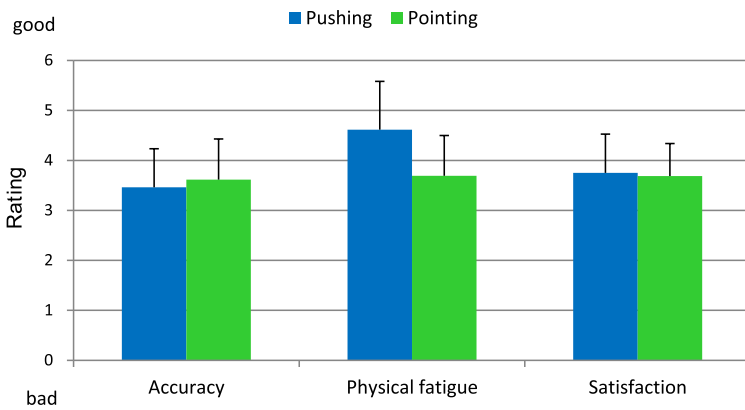
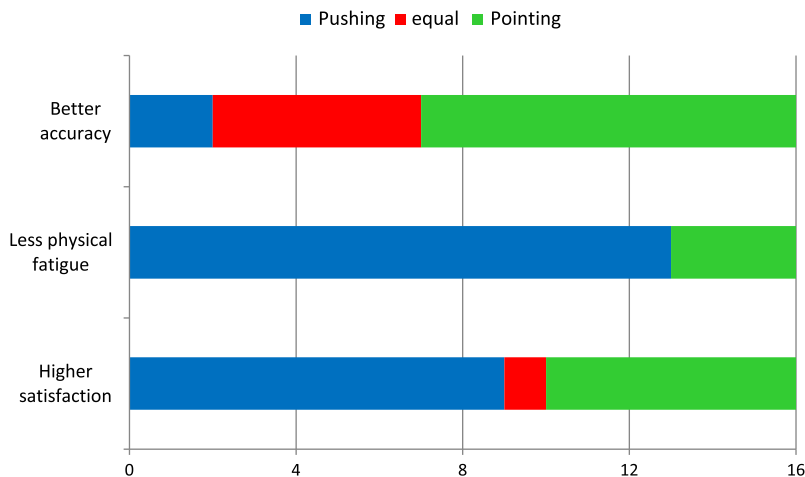


Fig. 9.15 Subjective rating of both interaction techniques regarding different criteria (individually)

in this study. Three of 16 participants state that they have recognized that the trial was conducted as a Wizard-of-Oz experiment, in particular, that system reactions were triggered by the experimenter. However, due to comments made by the participants while answering this question it can be assumed that this was only due to the perfect system reaction (movement of the object to the correct target area) during the first phase of the experiment. Participants wondered why the objects moved to the correct target even if their input was very vague (e.g., through the pushing movement).

In Fig. 9.15 the subjective ratings of both interaction techniques regarding different criteria is shown. *Pointing* is rated slightly better regarding accuracy than *Pushing*. However, it needs to be mentioned that the accuracy of both interaction techniques due to the Wizard-of-Oz experiment was identical. The difference



**Fig. 9.16** Subjective rating of both interaction techniques regarding different criteria (direct comparison)

between mean values of both techniques is statistical not significant ( $t(24) = 0.548$ ;  $p = 0.589$ ). In contrast, for the rating of physical fatigue a statistical significant difference of mean values can be observed ( $t(19) = 2.467$ ;  $p = 0.024$ ). Regarding user satisfaction again no significant difference is measured ( $t(30) = 0.255$ ;  $p = 0.801$ ).

If participants are asked to directly compare both techniques and rate them as better or worse compared to each other the results shown in Fig. 9.16 are obtained. They show the same tendency as the isolated rating of each technique.

### 9.3.3 Discussion

The subjective ratings of both techniques confirm the expected advantage of *Pushing* compared to *Pointing* regarding physical fatigue. *Pushing* is rated by 81 % of participants as less fatiguing compared to *Pointing*. It can be expected that this rating further develops in favor of *Pushing* if the interaction techniques are used over a longer period of time.

In order to allow for accurate interaction with techniques like *Pushing*, where user input only provides very vague information for target oriented object manipulation, an additional input modality may be included. The results of the experiment regarding proactivity of gaze behavior around the triggering pushing movement and gaze based estimation of user's intention are positive indications for using eye gaze as additional modality in this context. In 88 % user's intention can be estimated correctly for the setting described above. Errors almost exclusively occur for target areas which are located close to each other or for targets located along the same



movement direction of the object in the form of a straight line. One option to improve intention estimation further would be to trigger the object movement based on intention estimated previous to the triggering event and to adjust the movement path according to refined estimations based on gaze data analysis during the movement phase. Additionally, for calculation of the initial movement direction of objects the direction of the pushing movement could be combined with gaze-based intention estimation.

For *Pushing* a significant smaller amount of proactive fixations can be detected before the object movement is triggered compared to *Pointing*. One reason could be the subjective perception of *Pushing* as inaccurate, as indicated by the results obtained from the questionnaires. As illustrated in Sect. 9.2 the uncertainty of user's mental model about system reactions has significant influence on natural gaze behavior during interaction and in particular on its proactivity. Therefore this could be one reason for reduced proactivity of gaze behavior compared to *Pointing*.

The results also indicate the potential of improving accuracy of pointing gestures through multimodal combination with gaze-based intention estimation. For *Pointing* user's intention is estimated correctly with 92 % before the triggering input is detected. Additionally, gaze behavior is proactive in 97 % before the object movement is triggered.

One further interesting aspect regarding usage of natural gaze behavior as input modality can be derived from the rating of *Pushing* as inaccurate input and the comments given by participants on the Wizard-of-Oz experiment. Some participants mentioned that they were surprised that objects always moved to the correct target area during the first phase of the experiment. The same effect could happen if the system really reacts on natural gaze behavior according to the interaction technique *Pushing*. Ideally the object always moves to the correct target area as it was the case in the first phase of the experiment. For some participants it can be observed that they start to provoke wrong system reactions if, despite of very imprecise input via pushing gestures, the object always moves to the correct target area. Hence, it could be useful to help users to become aware of the functional principle of the interaction technique (e.g., through appropriate visual feedback) in order to allow for generation of an adequate mental model.

In the experiment target areas are located relatively far away from each other. How close target areas can be located to each other for still being able to estimate user's intention robustly mainly depends on the accuracy of the eye tracking system. Therefore, with increasing accuracy of eye trackers applications for graphical user interfaces with finer structure can be considered.

Finally, based on the results obtained from the user study natural gaze behavior can be considered as a valuable input modality for the application scenario outlined within this section. Based on gaze-based intention estimation accuracy of imprecise inputs like pushing or pointing gestures can be improved by combining them into multimodal interaction techniques. With gaze-based multimodal interaction techniques like *Pushing* the subjectively sensed physical fatigue can be reduced compared to alternative solutions such as *Pointing*.

## 9.4 Conclusion

In this chapter we presented results of two experiments which were conducted to investigate the use of natural gaze behavior as input modality for human-computer interaction. In the first experiment we were able to identify different factors influencing natural gaze behavior during an object manipulation task and to characterize their influence on proactivity and direction of fixations towards different task-related targets. Additionally, we demonstrated that the influence of individual factors can be changed by interaction design and adjusted visual feedback, respectively. The results reported regarding this experiment show the variety of information contained in natural gaze behavior. By analyzing natural gaze behavior during human-computer interaction information like user's intention or experience can be inferred which can be used for designing proactive or adaptive intelligent user interfaces.

In the second study we investigated natural gaze behavior as additional input modality for interaction in multi-display environments and especially in combination with gesture based interaction. By means of two interaction techniques for moving objects from one display within grasping range to another distant display new opportunities for combining gesture and gaze in multimodal interaction techniques were explored and evaluated. The results show that gaze-based intention estimation is valuable for compensation of inaccuracy of imprecise hand gestures (e.g., pushing or pointing gestures) and hence for designing new multimodal interaction techniques which cause less physical fatigue.

In general both of the two experiments show that natural gaze behavior contains valuable information about user's cognitive processes which can be used to improve human-computer interaction for conventional workspaces but also for novel interactive multi-display environments. Further research needs to be done in order to extend the knowledge about natural gaze behavior, especially for more complex tasks and larger user groups. The Wizard-of-Oz experiment used in the second study revealed important insights regarding usage of natural gaze behavior for object manipulation tasks in multi-display environments. These results need to be extended by further studies where input data is captured by technical input devices (e.g., video-based gesture recognition system and real-time eye tracking) causing uncertainty due to misrecognition and wrong system reactions. Interesting research questions which could be addressed by further studies concern the influence of such uncertain input channels on natural gaze behavior as well as effects related to long term usage of gaze-based interaction. In the latter case changes in natural gaze behavior may occur due to training effects and changing mental models.

## References

- Bader T (2011) *Multimodale Interaktion in Multi-Display-Umgebungen*. PhD thesis
- Bader T, Vogelgesang M, Klaus E (2009) Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In: *Proceedings of the 2009 international conference on multimodal interfaces (ICMI-MLMI)*. ACM, New York, pp 199–206

- Boring S, Baur D, Butz A, Gustafson S, Baudisch P (2010) Touch projector: mobile interaction through video. In: Proceedings of the 28th international conference on human factors in computing systems, pp 2287–2296
- Flanagan JR, Johansson RS (2003) Action plans used in action observation. *Nature* 424(6950):769–771
- Gesierich B, Bruzzo A, Ottoboni G, Finos L (2008) Human gaze behaviour during action execution and observation. *Acta Psychol* 128(2):324–330
- Hyrskykari A, Majoranta P, Riih  K (2003) Proactive response to eye movements. In: Rauterberg M (ed) Human computer interaction, INTERACT 2003. IOS Press, Amsterdam, pp 129–136
- Jacob RJK (1993) Eye movement-based human-computer interaction techniques: toward non-command interfaces. In: Advances in human-computer interaction. Ablex, Norwood, pp 151–190
- Jacob R, Karn K (2003) Eye tracking in human-computer interaction and usability research: ready to deliver the promises. Elsevier, Amsterdam, pp 573–605
- Johanson B, Fox A (2002) The Event Heap: a coordination infrastructure for interactive workspaces. In: Proceedings of the fourth IEEE workshop on mobile computing systems and applications, pp 83–93
- Johanson B, Fox A, Winograd T (2002) The Interactive Workspaces project: experiences with ubiquitous computing rooms. *IEEE Pervasive Comput* 1(2):67–74
- Johansson RS, Westling G, B ckstr m A, Flanagan JR (2001) Eye-hand coordination in object manipulation. *J Neurosci* 21(17):6917–6932
- Land M, Lee D (1994) Where we look when we steer. *Nature* 369:742–744
- Land MF, McLeod P (2000) From eye movements to actions: how batsmen hit the ball. *Nat Neurosci* 3:1340–1345
- Majoranta P, Riih  KJ (2002) Twenty years of eye typing: systems and design issues. In: Proceedings of the 2002 symposium on eye tracking research & applications, ETRA'02. ACM, New York, pp 15–22
- Nacenta MA, Sallam S, Champoux B, Subramanian S, Gutwin C (2006) Perspective cursor: perspective-based interaction for multi-display environments. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 289–298
- Nickel K, Stiefelhagen R (2003) Pointing gesture recognition based on 3D-tracking of face, hands and head orientation. In: Proceedings of the 5th international conference on multimodal interfaces, pp 140–146
- Nielsen J (1993) Noncommand user interfaces. *Commun ACM* 36(4):83–99
- Pelz J, Hayhoe M, Loeber R (2001) The coordination of eye, head, and hand movements in a natural task. *Exp Brain Res* 139:266–277. doi:[10.1007/s002210100745](https://doi.org/10.1007/s002210100745)
- Rekimoto J, Saitoh M (1999) Augmented surfaces: a spatially continuous work space for hybrid computing environments. In: Proceedings of the SIGCHI conference on human factors in computing systems: the CHI is the limit, pp 378–385
- Schick A, van de Camp F, Ijsselmuiden J, Stiefelhagen R (2009) Extending touch: towards interaction with large-scale surfaces. In: Proceedings of the ACM international conference on interactive tabletops and surfaces
- Smith BA, Ho J, Ark W, Zhai S (2000) Hand eye coordination patterns in target selection. In: Proceedings of the 2000 symposium on eye tracking research & applications, ETRA'00. ACM, New York, pp 117–122
- Streitz NA, Gei ler J, Holmer T, Konomi S, M ller-Tomfelde C, Reischl W, Rexroth P, Seitz P, Steinmetz R (1999) i-LAND: an interactive landscape for creativity and innovation. In: Proceedings of the SIGCHI conference on human factors in computing systems: the CHI is the limit, pp 120–127
- Zhai S, Morimoto C, Ihde S (1999) Manual and gaze input cascaded (magic) pointing. In: CHI99. ACM, New York, pp 246–253

# Chapter 10

## Co-present or Not?

### Embodiment, Situatedness and the Mona Lisa Gaze Effect

Jens Edlund, Samer Al Moubayed, and Jonas Beskow

**Abstract** The interest in *embodying* and *situating* computer programmes took off in the autonomous agents community in the 90s. Today, researchers and designers of programmes that interact with people on human terms endow their systems with humanoid physiognomies for a variety of reasons. In most cases, attempts at achieving this embodiment and situatedness has taken one of two directions: virtual characters and actual physical robots. In addition, a technique that is far from new is gaining ground rapidly: projection of animated faces on head-shaped 3D surfaces. In this chapter, we provide a history of this technique; an overview of its pros and cons; and an in-depth description of the cause and mechanics of the main drawback of 2D displays of 3D faces (and objects): the Mona Lisa gaze effect. We conclude with a description of an experimental paradigm that measures perceived directionality in general and the Mona Lisa gaze effect in particular.

#### 10.1 Introduction

The interest in *embodying* and *situating* computer programmes took off in the autonomous agents community in the 90s (e.g. Steels and Brooks 1995). Today, researchers and designers of programmes that interact with people on human terms—most notably using speech in human-machine dialogue and computer-mediated human-human dialogue—endow their systems with humanoid physiognomies for a variety of reasons, ranging from a hope to exploit the purported benefits of humanlike dialogue as a human-machine interface—people know how to speak and many of us are most comfortable communicating face—to a desire to use speech technology and working models of human dialogue to gain deeper understanding of how people communicate (Traum 2008; Edlund 2011).

In most cases, attempts at achieving this embodiment and situatedness has taken one of two directions. The first is to implement virtual characters, often referred to as virtual humans (e.g. Traum and Rickel 2002) or embodied conversational agents

---

J. Edlund (✉) · S. Al Moubayed · J. Beskow  
KTH Speech, Music and Hearing, Lindstedtsvägen 24, 100 44 Stockholm, Sweden  
e-mail: [edlund@speech.kth.se](mailto:edlund@speech.kth.se)

(ECAs; e.g. Cassel et al. 2000). We will adhere to the latter terminology and use ECA here. In principal, an ECA is a 2D or 3D model of a character in virtual space, that is displayed as a 2D rendition on a monitor in physical space. The relationship between the ECA and its virtual space, the monitor, and the humans watching the ECA can be portrayed in several ways by the ECA designer. Common images include the ECA living its life in another world, which is displayed to the onlookers as if it were a movie, as exemplified by Cloddy Hans and Karen in the NICE project (Boye and Gustafson 2005); the ECA again living in its virtual reality but peering out through a window (the monitor) through which the onlookers peer back in, as exemplified by Ville in the DEAL system (Hjalmarsson et al. 2007); and the ECA not living in a virtual world at all, but rather sharing the same physical world as the onlookers, as exemplified by MACK (Cassell et al. 2002) and the characters of the Gunslinger project (Hartholt et al. 2009).

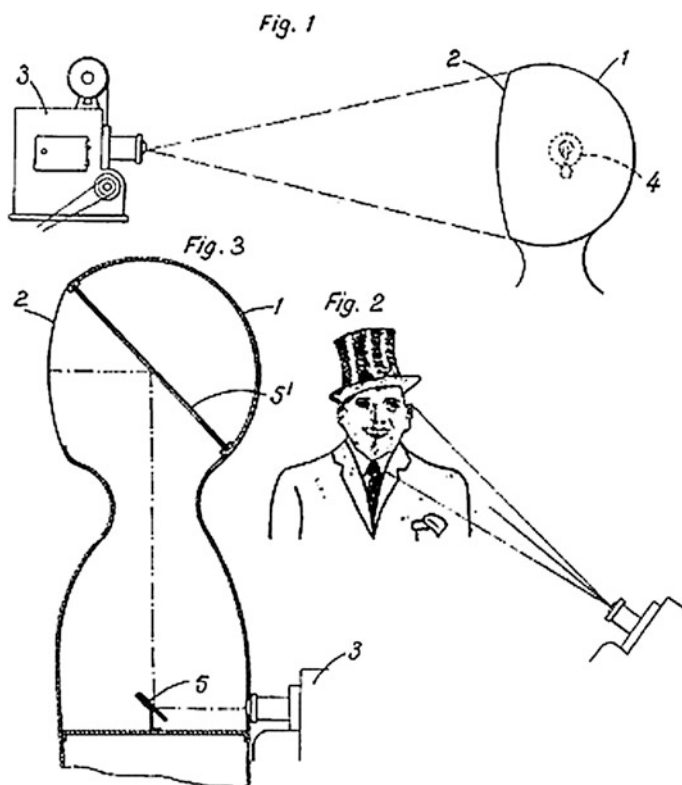
The other main direction is to implement actual physical robots which imitate people, such as MIT's Cog and Kismet (Breazeal and Scassellati 2001). Honda Research's robot Asimo is also merits mention in this context. Although not chiefly developed for communication studies, recent work on retargeting of motion captured from humans allows Asimo to reproduce human gestures quite closely in real-time, which opens up new possibilities for investigations into human communicative gesture (Dariush et al. 2008).

In studies of face-to-face communication, the head and face are often given centre stage. This is particularly true for head pose and gaze, as these are associated with a number of important communicative functions such as turn-taking and grounding. In addition, a number of other features of the head and face—for example facial expressions, eye brow movements, and lip synchronization—frequently receive special attention. In this chapter, we focus on a particular type of face and head embodiment which is becoming increasingly popular—a combination of a virtual talking head and a physical robotic head: projection of animated faces on head-shaped 3D surfaces.

The following section provides a history of the technique. After that, the next section holds a brief overview of the pros and cons of the technique compared to other methods of embodiment, followed by a section providing an in-depth description of the main drawback of 2D displays of 3D faces (and objects): the Mona Lisa gaze effect. This section includes a proposed explanation of the cause and mechanics of the effect; an examination of its consequences for face-to-face communication; a description of an experimental paradigm that measures perceived directionality in general and the Mona Lisa gaze effect in particular. Finally, the chapter is summed up with an account of how projection on head-shaped 3D surfaces completely cancels the effect.

## 10.2 Face Projection on 3D Surfaces

The method of embodiment that is our focus here is sometimes called *relief projection*, and the result is sometimes called a *projection augmented model*. In short,



**Fig. 10.1** Drawings of front and back projected faces taken from US Patent 1653180 of 1925. The drawing is in the public domain and copyright free, like all patents issued in the United States

a photographic image of an object is projected on a physical, three-dimensional model of the same shape as the object with the aim of creating a realistic-looking 3D object with properties that can be changed by manipulating the image. In the cases we are interested in, the image is a moving image, either a film of a person's face or a generated face such as those used for ECAs.

The earliest attestation of this technique is a patent application submitted by Georges Jalbert in December 1924 (France) and May 1925 (US; Jalbert 1925). The application describes both front projected faces and faces projected from the inside of a translucent bust, as seen in Fig. 10.1. Another patent (Liljegren and Foster 1989) specifically adds fibre optics as the means of transferring the images to within the bust.

The first modern well-documented implementation of face projection on 3D surfaces is the ghosts performing Grim Grinning Ghosts at the Disneyland Haunted Mansion ride. The ride was built in the 60s and opened in 1969, and the technology was described in a 1970 behind-the-scenes TV feature called Disneyland

Showtime,<sup>1</sup> which incidentally also features facial animatronics that seemingly measures up to MIT's Kismet. The ghosts are created by projecting films of strongly lit, highly contrasted faces against a black background onto relatively featureless white busts with shapes matching the faces in the films. Disney's ghosts, as well as another Haunted Mansion character produced with a similar technique, Madame Leota, are popular projections in private Halloween installations featuring face projection on busts. Films showing such installations dating from some time into the 2000s and onwards are easily found on the Internet, and a number of amateur special effects makers claim having produced them as early as the early 80s. Although there are several claims of proof in the form of footage and films, we have not been able to find any of these materials.

Another early and well-attested creation is the talking head projection of MIT's Architecture Machine Group in the early 80s. Inspired by Disney's Haunted Mansion creations. One of the creators of the MIT talking head projection, Michael Naimark, observed visitors at one of the talking heads in the Haunted Mansion at length. He concludes: "It was clear that as the woman spoke, the image of her moving lips would mis-register from the mask-shaped screen, but to most everyone viewing it briefly from their dark-ride car, this anomaly went unnoticed. Most people seemed convinced that they had just seen a full color, moving hologram (which, of course, is nonsense)" (Naimark 2005). The experience led Naimark and colleagues to develop the MIT talking head projection, an elaborate contraption which recorded not only image and sound, but also motion. The film was back projected to a head shaped mask moving in sync with the recorded person (Naimark 2005).<sup>2</sup> MIT Media Lab presented similar display in a tribute to the original experiments at their Defy Gravity exhibition in 2010.

In more recent years, a number of groups have put together 3D projection surfaces that are intended to be used with computer animated faces. From the more obscure prototype system HyperMask, which aims to project a face onto a mask worn by an actor who is free to move around in a room (Morishima et al. 2002) to more mundane ideas of embodying computer persons or improving telepresence. Hashimoto and Morooka (2006) use a spherical translucent projection surface and a back projected image of a humanoid face in combination with a robot neck. Light-Head of University of Plymouth (Delaunay et al. 2009) is more elaborately shaped, but maintains a stylized quality, while Technische Universität München's Mask-bot (Kuratate et al. 2011) and KTH Royal Technical Institute's Furhat (Al Moubayed et al. 2011) aim for higher degrees of human-likeness and project realistic faces into masks more closely resembling the human anatomy. It is worth mentioning that there are other ways to go as well, such as simplistic robot heads with small monitors for eyes and lips or the life-like mechatronic design of Hanson Robotics. And for the future, flexible and curved displays such as the spherical OLED displayed at

---

<sup>1</sup>Walt Disney's Wonderful World of Color, Season 16, Episode 20, Walt Disney Productions.

<sup>2</sup>Michael Naimark has made a film showing the talking head projection in action available at <http://www.naimark.net/projects/head.html>.

the Museum of Science and Education in Tokyo, and display techniques that utilize reflected light only, such as the Kindle, hold promise for development.

### 10.3 Pros and Cons of Face Projection

This section provides an overview of salient differences between face projection on 3D surfaces and the two main alternatives, physical robotic faces and 2D displays of 3D models.

#### *10.3.1 Compared to Physical Robotic Faces*

Compared to physical robot heads with moving parts for lips, eye brows, eyes, and other facial features, a projected face has a several advantages. To begin with, it is considerably cheaper to develop, and even more so to modify, as long as the modification can be done using the projection, rather than modifying the actual mask and other hardware. Development and in particular modification and adaptation is not only cheaper, but much faster, making face projection a much more feasible alternative for rapid development and experimentation, regardless of budget.

Another advantage is that projected movements—eye gaze shifts, brow raises, lip movements, and so on—are soundless in the projected face, whereas the hydraulics usually used in robotic components make noises that risks countering the humanlike impression of the robot. The projection is also able to make these movements more rapidly than robot actuators, at a speed that can easily match that with which a human produces them. An example of this is the SynFace lip synchronization (Beskow et al. 2009) that can be used with Furhat, allowing it to function as a remote representation of a human using the original voice of the human, but its own lip synchronization based on analysis of the acoustic signal, eliminating the need for a video stream in order to acquire lip movements.

As for disadvantages, there are two major drawbacks compared to robotic faces. The first one is that the light conditions needed for the back projection to be efficient are restrictive—even though the type of pico projector used in Furhat and other back projected talking heads are rapidly getting stronger at ever-better prices, with the current technology it is unlikely that the face will ever work well in direct sunlight. The second has to do with the inflexibility of the projection surface. Although the small misalignments caused by for example speaking are barely noticeable, as noted by Naimark (2005), larger jaw movements such as wide yawning will cause the face to look clearly out of order, with a projected jaw line clearly missing the jaw line of the projection surface.



### 10.3.2 Compared to 2D Displays of 3D Models

The most salient difference is the manner in which eye gaze is perceived, which we go through in detail in the next section. Besides that, the most obvious difference is how the projection is interpreted compared to a monitor. Whereas an ECA displayed on a monitor requires interpretation—it is not obvious whether it should best be interpreted as something in another world seen through a window or peeking out of a window or if it is supposed to be viewed as sharing the same space as its onlooker, it is immediately clear that the latter is the case for the chase minor advantages in being even more inexpensive and adaptable than the other method, and its gaze characteristics can be utilized as an advantage for specific purposes.

## 10.4 The Mona Lisa Gaze Effect

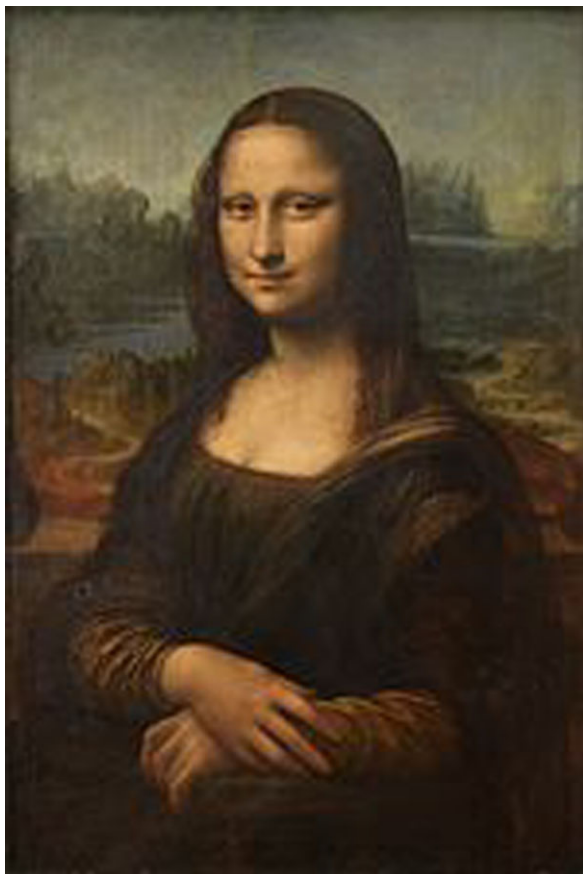
The perception of 2D renditions of 3D scenes is notoriously riddled with artefacts and illusions—for an overview, see Gregory (1997). The most important of these for embodiment is *the Mona Lisa gaze effect*, commonly described as an effect that makes it appear as if the Mona Lisa's gaze rests steadily on the viewer as the viewer moves through the room (Fig. 10.2). Although the reference to the Mona Lisa is a modern invention, documentation of the effect dates back at least as far as Ptolemy in around 100AD “[...] the image of a face painted on panels follows the gaze of moving viewers to some extent even though there is no motion in the image itself” (Smith 1996).

### 10.4.1 Mechanics of the Mona Lisa Gaze Effect

The Mona Lisa gaze effect has earned frequent enough mention, and a number of more or less detailed explanations have been presented from Ptolemy and onwards (e.g. Smith 1996; Cuijpers et al. 2010), but these do not provide an explanation that satisfies the requirements of a designer of embodied computer programmes. In Al Moubayed et al. (2012a, 2012b), we propose a model that explains Mona Lisa stare effects as well as other observations with a minimum of complexity, and verified its predictions experimentally. The model is based a number of observations, which are described in the following, before the model in itself is presented.

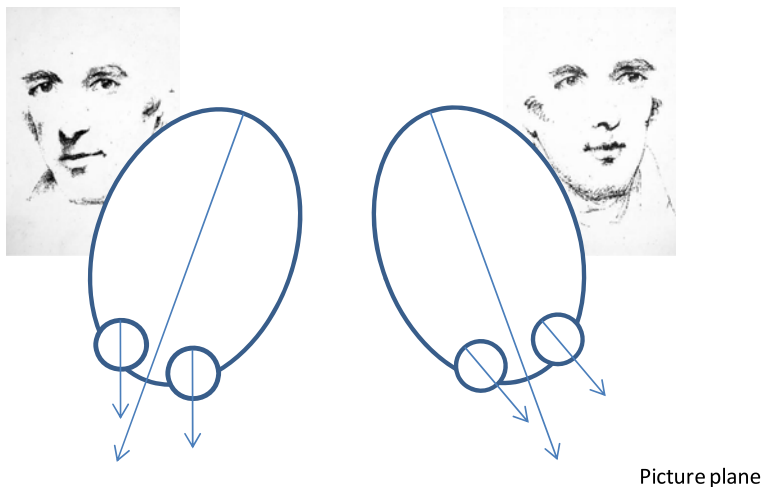
Our first, seemingly trivial observation is that *in order to judge gaze direction, it is not sufficient to know the angle of the eyes relative to the head*—which can be estimated for example by means of relative pupil position within the sclera (e.g. Cuijpers et al. 2010). An estimation of *the position and angle of the head is also required*. The background of Fig. 10.3 shows the Wollaston effect (Wollaston 1824), in which two pair of identical eyes appear to gaze at different points when drawn in to heads that have different angles. This “effect” seems to result from an insistence to view our interpretation of depicted eyes as somehow isolated from the head

**Fig. 10.2** Leonardo da Vinci's *Mona Lisa*. Mona Lisa appears to be looking straight at the viewer, regardless of viewing angle. The painting is the public domain and copyright free

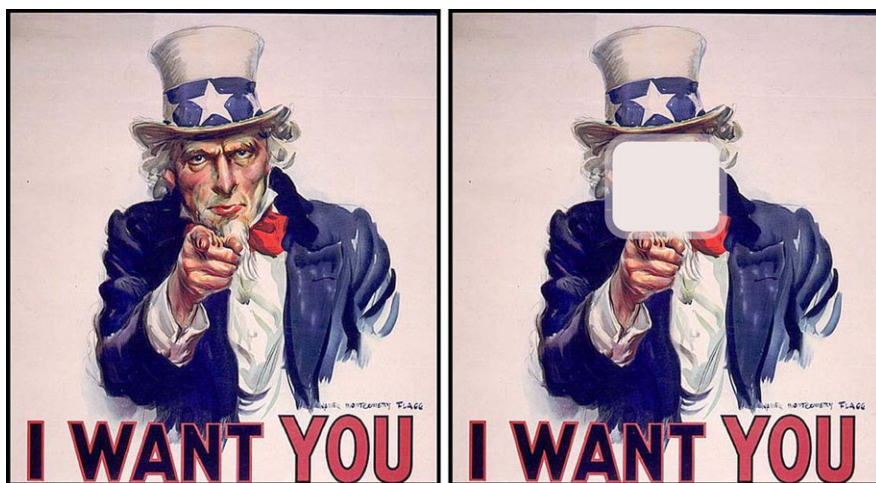


in which they are lodged. If we, like Todorović (2006), instead assume that head and eyes are interpreted in relation to each other and to the space they are depicted in, the Wollaston effect is not only accounted for, but rather ceases being an effect, as illustrated in the foreground of the figure. Todorović's account relates eyes and head pose in virtual space directly to perceived gaze direction in physical space. We generalize this account by means of simplification, and speak exclusively of gaze direction within the same (virtual or physical) space: *the perceived gaze direction within a space, virtual or physical, of a creature within that same space, is a function of the perceived angle of the gazing creature's head within that space, and the perceived angle of her eyes, relative her head.*

The second observation, illustrated in Fig. 10.4, is that *the Mona Lisa gaze effect is not restricted to eye gaze*, but generalizes to anything pointing out from a picture, such as an outstretched index finger. Most viewers feel that complete Uncle Sam in the left pane of the figure and faceless Uncle Sam in its right pane both point straight at them, regardless of viewing angle. This means that although eye and pupil position clearly affects how we perceive gaze direction, they cannot hold the



**Fig. 10.3** The Wollaston effect is seen in the two drawings: gaze direction is perceived differently although the eyes are identical, and only the head shape differs. The *ovals* with two *circles* represent a possible interpretation of the drawings as seen from above in virtual space. The two drawings are from Wollaston (1824), and are in the public domain and are copyright free



**Fig. 10.4** *I want you for the U.S. Army nearest recruiting station*, commissioned by the US federal government and painted by James Montgomery Flagg (cropped in the *left pane*, and cropped and edited in the *right pane*). The painting is in the public domain and copyright free

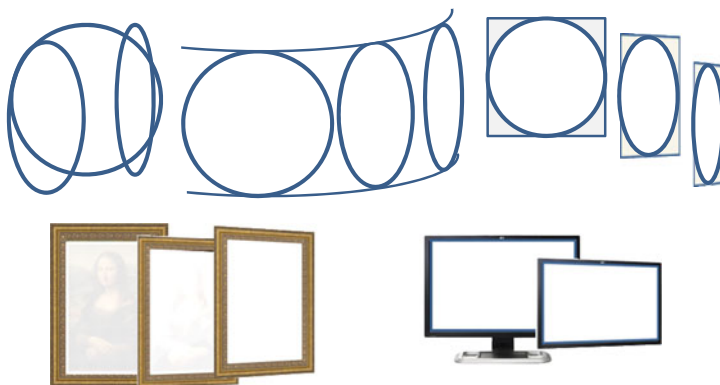
key to the effect, as the effect is present also when eyes and pupils are not. This observation also allows to generalize our statement from the last paragraph to not concern not only eyes and heads, but any object with a perceived direction contained within another.



**Fig. 10.5** *Interiors of the Winter Palace. The Throne Room of Empress Maria Fiodorovna.* Painting by Yevgraf Fyodorovich Krendovskiy. The picture is in the public domain and copyright free

The third observation is that *2D images representing 3D objects or scenes are interpreted as having their own virtual 3D space, distinct from physical space*. The axes of this virtual space are oriented along the horizontal and vertical edges of the image (perceived as width and height, respectively), with the third axis perpendicular to its surface (perceived as depth). This is particularly clear when we watch photos or paintings of large rooms with walls, ceiling and floors at right angles, as in Fig. 10.5. The painting in the figure gives a clear impression of a large three-dimensional space with a throne located at the far back. The location of the throne in relation to physical space is ambiguous: if our viewing angle and distance to the painting is varied, the throne's position in the portrayed virtual space is maintained, and its position in physical space remains unclear.

Our fourth observation has to do with the high degree of interpretation that goes in to the shapes we perceive in images. The phenomenon, known as *shape constancy*, is well-documented and was described early on. Descartes states in his *Dioptrics* of 1637: “[...] shape is judged by the knowledge, or opinion, that we have of the position of various parts of the objects, and not by the resemblance of the pictures in the eye; for these pictures usually contain only ovals and diamond shapes, yet they cause us to see circles and squares” (Descartes 1637, p. 107). Phrased differently, *viewers of 2D images perceives the shapes in the images as invariant, even when the viewing angle changes*, as exemplified by the two top right groups of circles in Fig. 10.6. Although the top left group contains exactly the same shapes, it is not necessarily perceived as three circles but rather as a circle and two ovals. This indicates that this perception is indeed dependent on interpretation. Note that when the circular shapes are viewed at a steep angle, most viewers still perceive a circle,



**Fig. 10.6** Three groups of three rounded shapes, a group of picture frames, and two computer monitors

although the shape is in fact distorted to one of Descartes' ovals. The figure also illustrates that shape constancy holds true for other shapes, such as the rectangular shape of the frame of the Mona Lisa or the edge of a standard monitor, *both of which are perceived as perfectly rectangular regardless of viewing angle*. For a more detailed account of shape constancy, see Gregory (1997).

We now have all the pieces we need for our explanation of the Mona Lisa gaze effect except one: a description of how viewers of 3D virtual space align that virtual space with their own, physical space. To solve this, we assume that a mechanism similar to shape constancy is at play: *viewers of 2D images depicting 3D objects interpret their position in relation to the virtual 3D space as head-on, perpendicular to the surface plane of the image*.

In addition to the support provided by what is known about shape constancy, this is intuitively pleasing as well. 2D images, at least those that use perspective to depict a 3D space, are created as seen from some vantage point in front of the objects seen in the picture. In the case of photographs or paintings created using camera obscura, this vantage point can be calculated exactly from the geometry of the image and the characteristics of the lens. Paintings allow for artistic license and may leave more ambiguity, but are still generally interpreted as seen head-on. This is again an observation that may seem trivial, but it has bearing as to how we may connect the virtual 3D space depicted in an image to the physical space of our surroundings.

It is worth pointing out that provided that we are standing in front of a picture, interpreting the general orientation and left-right position of the objects depicted in it is straightforward, whereas deciding the distance to the objects from the imagined vantage point of the creator can pose more of a problem, as illustrated by Fig. 10.7. It is trivial in both panes to see that all of the animals in the sculpture face left and that the rooster is on top and the horse at the bottom. Size variation, however, is ambiguous in 2D depictions and can be interpreted as deriving from at least three sources: the size of the depicted object, the distance and projection from the object to the position from which it is captured, and the size of the actual 2D image. The

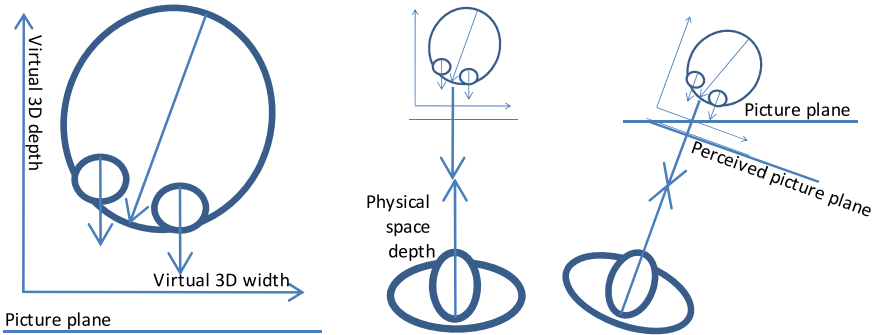




**Fig. 10.7** *Bremen Town Band*, Bremen, Germany. The picture was taken in 1990 by Adrian Pingstone and released into the public domain

image in the figure's left pane has been edited to remove the sculpture's surroundings. Without references, it is difficult to judge the size of the sculpture. The right pane contains more clues for the viewer to get a fair sense of distance and size, but the distance from the vantage point of the camera to the sculpture and from the sculpture to the people in the back are difficult to guess, so the size of the sculpture is elusive in that pane too.

We now have all the observations and assumptions we need for our proposed interpretation of the Mona Lisa gaze effect. Combining the assumptions we propose that the directionality of objects in 2D images are interpreted in relation to a virtual 3D space with axes oriented along the horizontal and vertical edges of the image and the third axis perpendicular to it, and that this space is aligned to the physical space of the viewer as if the image were viewed head-on. Shape constancy further allows us to make this interpretation regardless of the actual viewing angle, so that when observed, anything pointing straight out of the picture is perceived as pointing directly at the viewer, regardless of viewing angle. Figure 10.8 illustrates the model. The leftmost pane relates the head and eye of the gazing creature to the 3D space of the virtual space created by the picture, and to the picture plane. The illustration represents a head in virtual 3D space at a  $20^\circ$  angle relative to the depth axis, and eyes at the same but opposite angle relative to the head. The resulting eye direction is parallel with the virtual 3D depth axis and perpendicular to the picture plane. Virtual space is then aligned to physical space along their respective depth axes, as illustrated in the centre pane. Finally, shape constancy allows the viewer to view the picture *as if* facing it head-on, regardless of the viewer's position in the room, as



**Fig. 10.8** Our observations and assumptions combined into a model of how gaze direction (and the directionality of other objects) in 2D pictures are perceived

in the rightmost pane, causing the Mona Lisa gaze effect to occur. In other words: if something points straight out of a two-dimensional picture, it will be perceived by each on-looker, regardless of position in the room, to point straight at said on-looker. The model predicts that people viewing 2D images of gaze should have no problem judging gaze relative the virtual space coordinates, and should judge gaze direction in physical space *as if they were standing directly in front of the picture*. In other words, any gaze directed straight out of the picture would be perceived as looking straight at an on-looker, as is the case with Mona Lisa. Furthermore, gaze to the left or to the right the depth of the virtual space should always to the left or right, respectively, of an on-looker, and by a constant angle. As it turns out, all of these predictions bear out (Al Moubayed et al. 2012a, 2012b).

The model is very similar to that suggested by Todorović (2006), with the addition of shape constancy to account for the fact that most viewers do not perceive drawings viewed at an angle as distorted. The processing model, in which differences caused by viewing angle are removed initially, has the further advantage that the actual recognition becomes simpler, as there is less variability left to account for.

#### ***10.4.2 Impact on Human-Human and Human-Machine Communication***

The importance of gaze in social interaction is well-established. From a human communication perspective, Kendon’s work on gaze direction in conversation (Kendon 1967) is particularly important in inspiring a wealth of studies that singled out gaze as one of the strongest non-vocal cues in human face-to-face interaction (see e.g. Argyle and Cook 1976; Bavelas et al. 2002). Gaze has been associated with a variety of functions within social interaction—Kleinke’s review article from 1986, for example, contains the following list: “(a) provide information, (b) regulate interaction,

(c) express intimacy, (d) exercise social control, and (e) facilitate service and task goals” (Kleinke 1986).

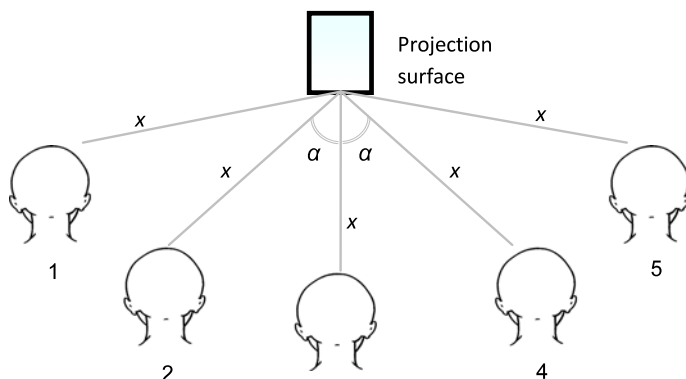
These efforts and findings, in turn, were and are shadowed by an increasing effort in the human-computer interaction community, which recognized the importance of modelling gaze and its social functions such as expressing and communicating attitudes and emotions in embodied conversational agents (ECAs). Examples include Takeuchi and Nagao (1993), Poggi and Pelachaud (2000), Bilvi and Pelachaud (2003), and Lance and Marsella (2008). As multimodal and facial communication with communication devices become more advanced and more popular, the demand for ECAs in control of their gaze behaviour increases. Multimodal interfaces are now able to provide testing and manipulation frameworks for behavioural models of gaze and other non-vocal signals. Such systems have recently been effectively used to investigate and quantify the effects of gaze using controlled experiments (Edlund and Nordstrand 2002; Lance and Marsella 2008; Gu and Badler 2006; Edlund and Beskow 2009; Nordenberg et al. 2005).

Given the importance of gaze, the effects of presenting an ECA which displays a perceivable gaze direction without being able to control this direction are potentially devastating for the communication and for how the ECA is perceived. Oddly enough, there is one clear example when the Mona Lisa gaze effect does not cause this to happen, but rather presents us with the remedy: when our ECA communicates with one single person whose head and face we have no ability to track. Incidentally, this is historically the most common setup for interactional experiments with spoken dialogue systems represented by an ECA.

The way this works is as follows. A key problem with using ECA gaze for communicative purposes is that unless we have access to sensors and head tracking equipment, which were expensive and hard-to-get until rather recently, the system does not know where its human interlocutor’s head and eyes are, which makes directing the ECAs gaze at them a feat of magic. In many cases, experimenters have simply hoped that the human interlocutor will stay relatively immobile in front of the mobile, and used a gaze straight out from the monitor as an approximation of “looking at the interlocutor”. Whether a case of insight or sheer luck, this method is quite reliable—more so than one would think. As movements by the human interlocutor are negated by the Mona Lisa gaze effect, the system is always perceived as gazing at the interlocutor when it attempt to do so, and never when it attempts to look away. Under these quite restricted but rather common circumstances, harnessing the Mona Lisa gaze effect is the only way to achieve gaze reliably towards the interlocutor, as access to head pose information does not improve the situation—on the contrary, attempting to gaze at the real position of the interlocutors head would have the opposite effect in all cases except when the interlocutors sits straight in front of the monitor.

As soon as there is more than one person in the room, the Mona Lisa gaze effect becomes a very real problem. The system designer has a choice of having the ECA look straight out from the monitor, thus being perceived as looking straight at each person in the room simultaneously (and meeting their gaze of they look back), or look away from every person in the room. Note that having access to head pose data still will not help.





**Fig. 10.9** Schematic layout of a subject/target experiment with 5 targets at  $x$  meters distance from the stimuli—a projection surface—and with equal distances between adjacent subject/targets

### 10.4.3 Measuring Perceived Direction

There are many cases where we would want to test how directionality in face-to-face situations is perceived, for example to verify a model such as the one just proposed; to investigate the accuracy of perception, perhaps under adverse conditions; or to train targets of a robot so that they match human perception. There are also ways to present ECAs that make it less than obvious whether the Mona Lisa gaze effect is in place or not, or that are perhaps only partially susceptible to the effect, such as the illusionistic 3D ECA presented by Kipp and Gebhard (2008).

Beskow and Al Moubayed (2010) pioneered an experimental paradigm that was developed to allow experimenters to quickly investigate and gather large amounts of data on human perception of gaze targets/direction. The paradigm is described here in generalized form, allowing it to function as a means of comparing not only gaze targets but arbitrary directional stimuli such as directional audio or verbal descriptions. In recognition of the fact that the key to the effectiveness of the paradigm is to utilize the same people as subjects and targets for the directional stimuli, we will call the method the subject/target paradigm here. The practice of using subjects as targets also adds to the ecological validity of the paradigm, as the distinction between being or not being the person gazed at or spoken to is a salient distinction in face-to-face interaction.

In the subject/target paradigm, a group of  $N$  subjects are placed in a circle or semi-circle, so that there is one point at their centre which is equidistant to each subject, from which all stimuli are presented (the *centre*). Subjects positions are numbered  $P_1$  to  $P_N$ , and the angle between each subject's position, that of the centre, and that of the subject's closes neighbouring subjects ( $A(P_1 P_2) \dots A(P_N P_1)$ ) is calculated. Subjects may or may not be equidistant from their closest neighbours. Figure 10.9 shows a subject/target setup with five subjects and stimuli presented on a projection surface.

All subjects double as targets for the directional stimuli. During an experiment, directional stimuli are aimed at each of the subjects. The order is varied systematically, and the number of stimuli is such that each subject is targeted as many times as the others in one set of stimuli. A set of stimuli, then, contains a multiple  $R$  of  $N$  for a total of  $R*N$  stimuli. Once one set is completed, the subjects rotate—they shift their positions by one step and the process of presenting a set of  $N*R$  stimuli is repeated. The rotation is repeated  $N$  times, until each subject has been in each position once, making the total number of stimuli presented in an experiment  $N*R*N$ .

Each time a stimulus has been presented, each subject is asked to point out the intended target in such a manner that the other subjects cannot see it. This result in  $N$  judgements for each stimulus, for a total of  $N*R*N*N$  data points in one experiment. If more than one experiment condition is to be tested, the entire process is repeated from the beginning. The manner in which the subject/targets point out the intended target is not prescribed by the paradigm. Methods that have been used to date include jotting the result down on a predesigned form, which requires the full use of hands and eyes (Beskow and Al Moubayed 2010); simply asking the addressee to respond in an interactive test, which yields considerably fewer data points, but may increase ecological validity (Al Moubayed and Skantze 2011); and marking the target with through manual signing, for example by showing different numbers of fingers, which obviates the need for eye sight and so can be used for tests of acoustic directionality—the use of blindfolds would render regular form filling impractical, as the subjects cannot see to write (Edlund et al. 2012).

The experiment results can be analyzed in a number of ways. Subject performance measures such as inter-subject agreement, target accuracy, and average error in degrees are obvious examples, which can be analyzed more finely to show whether the average error is, for example, larger when the target is far away from the subject. Another use of the paradigm that is useful when the exact relation between system internal controls for pointing and perceived reality of the pointing is unknown. By pointing (gazing) at systematically varied spots along the circle of subject/targets and analyzing the resulting judgements, we can find a function connecting the system controls to perceived target angles. The experiments showed clearly that subjects are very good at reliably estimating gaze targets from 3D projected talking heads, but considerably less so from 2D displays.

A final example of analysis takes us back to the Mona Lisa gaze effect. We have stated that this effect ought not ruin a person's ability to judge directions altogether, but merely change the way these directions are mapped into the physical world of the person. This allows us to remap the subjects responses from an absolute target to a target relative to the subject's position. If the Mona Lisa gaze effect is in place, the re-mapped responses should be as accurate, or almost as accurate if we allow for some loss in translation, as the absolute targets when the Mona Lisa gaze effect is not in place. Al Moubayed et al. (2012a, 2012b) show in that this is indeed the case: the original mappings (as stated above) yield good results for the 3D projected talking heads and less so from 2D displays, while the remapped responses yield the opposite result.

We can take this reasoning one step further. The Mona Lisa gaze effect is in place when the gazing creature is perceived as being present in a separate space, such as a painting or the virtual reality inhibited by ECAs. And when the Mona Lisa gaze effect is in place, we get high accuracy of subject/target experiments once the results are re-mapped into subject-relative terms. On the other hand, the Mona Lisa gaze effect is not present when the gazing creature is perceived as sharing the same space as the subject—that is when it is *co-present* with the subject. In these cases, we get high accuracy of subject/target experiments with the original results. This suggests that by comparing the score of re-mapped, relative accuracy with original, absolute accuracy, we may be able to get a bearing on to what extent the subjects perceive an embodiment as co-present, as suggested by preliminary results presented in Edlund et al. (2011).

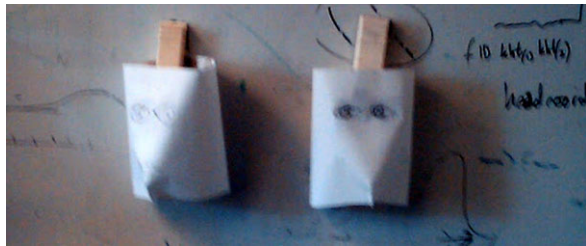
## 10.5 Summary

Hashimoto and Morooka (2006) state that “a curved surface image has a dependable direction of observation and presence in actual space”. To quantify this, the results of several studies of the accuracy and inter-subject agreement of perceived gaze targets of Furhat show unequivocally that the use of a front or back projected talking head onto a surface of similar shape completely cancels the Mona Lisa gaze effect (Beskow and Al Moubayed 2010; Al Moubayed and Skantze 2011; Al Moubayed et al. 2012a, 2012b). Preliminary results from multi-party conversations with Furhat also suggest that its gaze characteristics are suitable for turn-taking and addressee selection (Al Moubayed et al., in press).

What, exactly, is it that causes this? We have established that the Mona Lisa gaze effect is derived from human interpretation and is a result of a person aligning the coordinate system of a perceived virtual space with that of the physical space in which the human resides in such a manner that the human places herself in a position straight ahead of the image or movie that portrays the virtual space. We have even suggested that a measure based on a comparison of absolute (non-Mona Lisa) direction accuracy versus relative (Mona Lisa) gaze accuracy may give us some insight as to the extent to which an embodied computer programme is perceived as co-present with the viewer.

We suggest that in the end, it boils down to a simple matter on whether the viewer interprets the person gazing or the finger pointing as being present in the same room, or as being portrayed through a “window” onto another space. We leave this narration with an image—two pen drawings of two pairs of eyes on two plain sheets of A4 paper. When the drawings are viewed flat—as images on a rectangular piece of paper—they display the Mona Lisa gaze effect to its fullest. When, on the other hand, they are curved into cylinders, as in Fig. 10.10, their gaze is easily perceived as having an absolute target in the room—even though the area behind the eyes is forced flat by a piece of cardboard behind the eyes. Clearly, more study is needed to learn exactly what is needed to turn our perception from through-the-looking-glass to co-present mode. As it stands, it may well be a question of

**Fig. 10.10** Two pen drawings on rolled-up A4 sheets of paper



whether the features of a face appear to be drawn *inside* their own space on a piece of paper, or *outside* the object boundaries outlined by the same piece of curved paper.

## References

- Al Moubayed S, Skantze G (2011) Effects of 2D and 3D displays on turn-taking behavior in multiparty human-computer dialog. In: Proceedings of SemDial, Los Angeles, pp 192–193
- Al Moubayed S, Alexanderson S, Beskow J, Granström B (2011) A robotic head using projected animated faces. In: Salvi G, Beskow J, Engwall O, Al Moubayed S (eds) Proceedings of AVSP2011, p 69
- Al Moubayed S, Beskow J, Granström B, Gustafson J, Mirning N, Skantze G, Tscheligi M (2012a) Furhat goes to Robotville: a large-scale multiparty human-robot interaction data collection in a public space. In: Proceedings of LREC workshop on multimodal corpora, Istanbul, Turkey
- Al Moubayed S, Edlund J, Beskow J (2012b) Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. *ACM Trans Interact Intell Syst* 1(2):25
- Argyle M, Cook M (1976) Gaze and mutual gaze. *Science* 194(4260):54–55
- Bavelas J, Coates L, Johnson T (2002) Listener responses as a collaborative process: the role of gaze. *J Commun* 52(3):566–580
- Beskow J, Al Moubayed S (2010) Perception of gaze direction in 2D and 3D facial projections. In: The ACM/SSPNET 2nd international symposium on facial analysis and animation, Edinburgh, UK
- Beskow J, Salvi G, Al Moubayed S (2009) SynFace—verbal and non-verbal face animation from audio. In: Proceedings of the international conference on auditory-visual speech processing, AVSP’09, Norwich, England
- Bilvi M, Pelachaud C (2003) Communicative and statistical eye gaze predictions. In: Proceedings of international conference on autonomous agents and multi-agent systems (AAMAS), Melbourne, Australia
- Boye J, Gustafson J (2005) How to do dialogue in a fairy-tale world. In: 6th SIGdial workshop on discourse and discourse
- Breazeal C, Scassellati B (2001) Challenges in building robots that imitate people. In: Dautenhahn K, Nehaniv CL (eds) Imitation in animals and artifacts. MIT Press, Boston, pp 363–390
- Cassel J, Sullivan J, Prevost S, Churchill EE (2000) Embodied conversational agents. MIT Press, Cambridge
- Cassell J, Stocky T, Bickmore T, Gao Y, Nakano Y, Ryokai K (2002) MACK: media lab autonomous conversational kiosk. In: Proceedings of Imagina02, Monte Carlo
- Cuijpers RH, van der Pol D, Meesters LMJ (2010) Mediated eye-contact is determined by relative pupil position within the sclera. In: Perception ECVF abstract supplement, p 129
- Dariush B, Gienger M, Arumbakkam A, Goerick C, Zhu Y, Fujimura K (2008) Online and markerless motion retargeting with kinematic constraints. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS 2008), pp 191–198

- Delaunay F, de Greeff J, Belpaeme T (2009) Towards retro-projected robot faces: an alternative to mechatronic and android faces. In: Proceedings of the international symposium on robot and human interactive communication (RO-MAN), Toyama, Japan
- Descartes R (1637) *Dioptrics*. In: Discourse on method, optics, geometry, and meteorology. Hackett, Indianapolis, pp 65–162
- Edlund J (2011) In search of the conversational homunculus—serving to understand spoken human face-to-face interaction. Doctoral dissertation, KTH
- Edlund J, Beskow J (2009) MushyPeek—a framework for online investigation of audiovisual dialogue phenomena. *Lang Speech* 52(2–3):351–367
- Edlund J, Nordstrand M (2002) Turn-taking gestures and hour-glasses in a multi-modal dialogue system. In: Proceedings of ISCA workshop on multi-modal dialogue in mobile environments, Kloster Irsee, Germany
- Edlund J, Al Moubayed S, Beskow J (2011) The Mona Lisa gaze effect as an objective metric for perceived cospatiality. In: Vilhjálmsson HH, Kopp S, Marsella S, Thórisson KR (eds) Proceedings of the 10th international conference on intelligent virtual agents (IVA 2011), Reykjavík. Springer, Berlin, pp 439–440
- Edlund J, Heldner M, Gustafson J (2012) Who am I speaking at?—perceiving the head orientation of speakers from acoustic cues alone. In: Proceedings of LREC workshop on multimodal corpora 2012, Istanbul, Turkey
- Gregory R (1997) *Eye and brain: the psychology of seeing*. Princeton University Press, Princeton
- Gu E, Badler N (2006) Visual attention and eye gaze during multiparty conversations with distractions. In: Proceedings of the international conference on intelligent virtual agents
- Hartholt A, Gratch J, Weiss L, Leuski A, Morency L-P, Marsella S, Liewer M, Thiebaut M, Doraiswamy P, Tsiartas A (2009) At the virtual frontier: introducing Gunslinger, a multi-character, mixed-reality, story-driven experience. In: Proceedings of the 9th international conference on intelligent virtual agents (IVA'09). Springer, Berlin, pp 500–501
- Hashimoto M, Morooka D (2006) Robotic facial expression using a curved surface display. *J Robot Mechatron* 18(4):504–505
- Hjalmarsson A, Wik P, Brusik J (2007) Dealing with DEAL: a dialogue system for conversation training. In: Proceedings of SIGdial, Antwerp, Belgium, pp 132–135
- Jalbert G (1925) Lay figure. Technical report, US Patent 1653180
- Kendon A (1967) Some functions of gaze direction in social interaction. *Acta Psychol* 26:22–63
- Kipp M, Gebhard P (2008) IGaze: studying reactive gaze behavior in semi-immersive human-avatar interactions. In: Proceedings of the 8th international conference on intelligent virtual agents (IVA'08), Tokyo, Japan
- Kleinke CL (1986) Gaze and eye contact: a research review. *Psychol Bull* 100:78–100
- Kuratate T, Matsusaka Y, Pierce B, Cheng G (2011) Mask-bot: a life-size robot head using talking head animation for human-robot communication. In: Proceedings of the 11th IEEE-RAS international conference on humanoid robots (humanoids), pp 99–104
- Lance B, Marsella S (2008) A model of gaze for the purpose of emotional expression in virtual embodied agents. In: Proceedings of the 7th international conference on autonomous agents and multiagent systems, pp 199–206
- Liljégren GE, Foster EL (1989) Figure with back projected image using fiber optics. Technical report, US Patent 4978216
- Morishima S, Yotsukura T, Binsted K, Nielsen F, Pinhanez C (2002) HyperMask: talking head projected onto real objects. *Vis Comput* 18(2):111–120
- Naimark M (2005) Two unusual projection spaces. *Presence* 14(5):597–605
- Nordenberg M, Svanfeldt G, Wik P (2005) Artificial gaze—perception experiment of eye gaze in synthetic faces. In: Proceedings from the second Nordic conference on multimodal communication
- Poggi I, Pelachaud C (2000) Emotional meaning and expression in performative faces. In: Paiva A (ed) *Affective interactions: towards a new generation of computer interfaces*, pp 182–195
- Smith AM (1996) Ptolemy's theory of visual perception: an English translation of the "Optics" with introduction and commentary. Am. Philos. Soc., Philadelphia

- Steels L, Brooks R (eds) (1995) *The artificial life route to artificial intelligence: building embodied, situated agents*. Lawrence Erlbaum Associates, Hillsdale
- Takeuchi A, Nagao K (1993) Communicative facial displays as a new conversational modality. In: *Proceedings of the INTERACT'93 and CHI'93 conference on human factors in computing systems*
- Todorović D (2006) Geometrical basis of perception of gaze direction. *Vis Res* 45(21):3549–3562
- Traum D (2008) Talking to virtual humans: dialogue models and methodologies for embodied conversational agent. In: Wachsmuth I, Knoblich G (eds) *Modeling communication with robots and virtual humans*. Springer, Berlin, pp 296–309
- Traum D, Rickett J (2002) Embodied agents for multi-party dialogue in immersive virtual worlds. In: *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS 02)*. ACM, New York
- Wollaston WH (1824) On the apparent direction of eyes in a portrait. *Philos Trans R Soc Lond B* 114:247–260

# Index

## A

Ablated Wizard of Oz, 27  
ANOVA, 15, 17, 31–33  
Anvil, 90  
Areas of interest (AOI), 119  
Attentional behavior, 1–3, 5

## B

Boredom, 154, 156–159

## C

Calibration, 46, 53, 174  
Classification, 142, 143, 155  
Co-present, 185, 200  
Cognition, 111–115, 117, 118, 121, 122, 125, 132  
Collaboration, 63, 64, 66  
Collaborative model of referring, 25  
Common-word, 74, 78  
Concept-map, 66–69, 73, 80, 81, 83  
Confusion matrices, 178  
Conversation, 43–48, 52–58  
Cross validation, 122, 131  
Cursor, 10, 13, 14, 17, 20

## D

Decision tree, 85, 94–96, 109  
Disengagement, 85, 86, 90, 92–96, 98–102, 104–106, 108, 109  
DOM, 140, 141, 156  
Doubt, 154, 157–159  
Dracula, 145, 146

## E

E-book, 137, 143, 147, 158  
EEG, 137, 140, 141, 154, 157, 159

Embodying, 185, 188

Emotions, 137, 138, 153–159

Emotiv, 154–156, 158

Engagement, 85–90, 93–102, 104–109

Engagement-sensitive conversational agent, 97

Equal Error, 122

Equal Error Rate (EER), 122

Evaluation, 113, 118, 120, 122, 123, 125, 129

Event handling, 141

Experience, 137, 138, 143, 145, 147, 148, 151–153, 158

Experimenter effect, 153

Explicit reference, 83

Eye movement distance, 94–97, 100, 102, 103, 108

Eye movement features, 119, 120, 122

Eye movements, 9, 10, 14, 17, 138, 139

Eye tracker, 27, 28

Eye-gaze duration, 94, 96

Eye-tracking intervention, 9

Eye-voice span, 64, 70, 73

Eyelink, 11, 12

EyePad, 147–149, 151, 153, 159

## F

F-measure, 96, 109

Face projection, 186–189

Face tracking, 47, 51, 53, 54, 58

Face-directed gaze, 47, 52, 54–57

Face-to-face, 186, 196, 198

Face-to-face communication, 1, 2

Factorial experiment, 30

Feedback, 9–12, 15–20

Filtering, 142

**G**

Gaussian weighting function, 75  
 Gaze direction, 190–192, 196, 197  
 Gaze direction transition 3-gram, 90  
 Gaze modeling, 26, 27  
 Gaze tracking, 44–47, 52, 53, 58  
 Gaze-aware user interface, 1–3, 5  
 Gestures, 1–3, 5

**H**

Hand-eye coordination, 162, 163  
 Handwriting recognition, 149–151, 153, 158  
 Hollywood books, 143  
 Horizontal display, 174–176  
 HTML, 138–140, 143, 144, 148  
 Human robot dialogue, 23, 27, 36  
 Human-computer interaction, 161, 162, 170, 171, 182

**I**

Information state, 98  
 Intention, 114, 117, 118, 122, 125  
 Interaction effect, 32  
 Interest, 147, 148, 154, 156–159

**J**

Jerks, 139  
 Joint attention, 9–18, 20  
 Joy, 154, 156–159  
 JSGF, 149

**K**

Kappa score, 76, 78  
 Krohn-Rhodes theory, 51

**L**

Latin square, 30, 31  
 LibSVM, 120, 122  
 Likert scale, 100  
 Listener's fixation, 70, 74  
 Listener's span, 65, 70, 71, 73, 76

**M**

Machine learning pipeline, 131  
 Markov model, 41, 44, 48, 50, 57  
 Matching threshold, 75–77  
 Mental model, 161, 164, 166, 171, 172, 181  
 Midas, 144  
 Minimum gaze threshold, 74–78  
 Mismatched perceptions, 23–26, 28, 32, 34  
 Multi-display, 161, 172, 174, 182  
 Multi-party conversations, 200  
 Mutual gaze, 41–47, 51, 54–58, 87, 90

**N**

Nonlinear, 148  
 Normalization methods, 122, 123, 125, 126, 129, 131  
 Novelty effect, 153

**O**

OnBoredom, 156  
 OnFixation, 141  
 OnFurrow, 156  
 OnGazeOut, 141  
 OnGazeOver, 141  
 OnInterest, 156  
 OnRead, 141, 143  
 OnSmile, 156  
 Orientation, 171  
 Overall saccade shape, 142, 143

**P**

Participation attitudes, 86, 108  
 Performance task, 122  
 Physical space, 186, 191, 193–196, 200  
 Planning, 113, 117, 118, 122, 125  
 Pointing, 162, 163, 172–174, 176, 178–182  
 Pre-object fixations, 166, 168–170  
 Proactive fixations, 168–172, 181  
 Probing questions, 85, 86, 98–109  
 Problem-solving behavior, 111, 120, 122  
 Projection augmented model, 186  
 Psycholinguistic studies, 26  
 Pupil size, 85, 88, 94–97, 100, 103–105, 108, 109  
 Pushing, 173, 174, 176–182

**R**

Reactive fixations, 168, 169, 171  
 Reading detection, 138, 141, 142  
 Real-time, 9–11, 14–20  
 Reference-word, 78, 79  
 References, 63, 64, 66, 69–75, 79–83  
 Referential grounding, 23, 24, 36  
 REGARD, 63, 64, 70, 73, 74, 77–82  
 Relief projection, 186  
 Robots, 41–43, 48, 56, 57

**S**

Saccade, 114, 119, 121  
 Shape constancy, 193–196  
 Shared gaze, 23, 24, 27, 30, 32, 33  
 Shared storybook reading, 9, 10, 12  
 Shared workspace, 63, 64, 68, 74, 82  
 Situating, 185  
 Social dominance, 56, 57  
 Social interaction, 41, 42, 44  
 Spatial language, 34–36



Speaker's fixation, 70, 71, 73, 74  
Speaker's span, 65, 70, 71, 73, 76  
Speech, 41, 43–45, 48, 49, 51–54, 56–58,  
63–66, 68–73, 78, 80–82  
Speech recognition, 97, 99, 137, 138, 147, 149,  
150, 153  
Speech synthesis, 150, 153  
Support vector machine, 113, 121

**T**

Tabletop display, 163, 173  
The Little Prince, 144–146  
The Mona Liza gaze effect, 185, 186, 190  
Think-aloud protocol, 115  
Tobii, 12, 13, 28, 68, 89, 97, 116  
Tooltips, 139

**V**

Verbal data, 117  
Verbal reference, 63, 64, 69–75, 81  
Virtual agents, 42  
Virtual space, 186, 191–196, 200  
Voice-eye span, 64, 70, 73

**W**

Window size, 117, 120, 123–132  
Within-subject, 175  
Within-subject design, 99  
Wizard-of-Oz, 88, 89, 99, 109, 176, 179, 181,  
182

**Z**

Z-score, 122, 125, 127–130