Proceedings of the

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction

in Conjunction with IUI 2011, the International Conference on Intelligent User Interfaces

February 13, 2011

Palo Alto, California, USA

Workshop Organizers:

Yukiko Nakano (Seikei University, Japan)
Cristina Conati (University of British Columbia, Canada)
Thomas Bader (Karlsruhe Institute of Technology, Germany)

Preface

In interactive systems, eye-gaze and attentional information have great potential in improving the communication between the user and the systems. For instance, by combining with situational and linguistic information, user's focus of attention is useful in interpreting the user's intentions, understanding of conversation, and attitude towards the conversation. In addition to such multimodal interpretation of users' inputs, generating proper gaze behaviors serving as nonverbal signals is also indispensable in mediated communication using avatars as well as during interaction with humanoid autonomous agents.

Following the first eye-gaze workshop held at IUI 2010, which addressed issues of eye-tracking technologies, analyses of human eye-gaze behaviors, multimodal interpretation, user interfaces using an eye-tracker, and presenting gaze behaviors in humanoid interfaces, this workshop continues exploring this important topic, and covers wider topics from eye-tracking technologies to building gaze-based interactive IUIs. Moreover, the workshop will bring together researchers including human sensing, intelligent user interface, multimodal processing, and communication science, with the long term goal of establishing a strong interdisciplinary research community in "attention aware interactive systems".

We would like to thank authors of each paper for their contributions to the workshop, and thank our Program Committee members for their precious time and great effort in reviewing papers. This workshop has received generous support from SMI SensoMotoric Instruments GmbH.

Yukiko Nakano, Cristina Conati, and Thomas Bader

Workshop Co-organizers

Program Committee

Elisabeth André (University of Augsburg, Germany)
Nikolaus Bee (Augsburg University, Germany)
Justine Cassell (Carnegie Mellon University, USA)
Joyce Chai (Michigan State University, USA)
Andrew Duchowski (Clemson University, USA)
Jürgen Geisler (Fraunhofer IOSB, Germany)
Patrick Jermann (Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland)
Yoshinori Kuno (Saitama University, Japan)
Kasia Muldner (Arizona State University, USA)
Toyoaki 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)
Candy Sidner (Worcester Polytechnic Institute, USA)

Supported by

Table of Contents

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

An Eye-tracking Intervention Study

Jia Guo
Duke University
Department of Psychology and Neuroscience
jia.guo@duke.edu

Gary Feng
University of Michigan
Educational Studies, School of Education
garyfeng@umich.edu

ABSTRACT
This paper studies how eye-tracking can be used to measure and facilitate joint attention in parent-child interaction. Joint attention is critical for social learning activities such as parent-child shared storybook reading. There is a 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-related skills. Traditional research paradigm does not measure joint attention in real-time during shared storybook reading. In the current study, we simultaneously tracked eye movements of a parent and his/her child with two eye-trackers. We also provided real-time feedback to the parent where the child was looking at, and vice versa. Changes of dyads' reading behaviors before and after the joint attention intervention were measured from both eye movements and video records. Baseline data show little joint attention in parent-child shared book reading. The real-time eye-gaze feedback significantly changes parent-child interaction and improves learning.

Author Keywords
Joint attention, eye-tracking, eye gaze, inter-person interactions, shared storybook reading.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces: Computer-supported cooperative work; H.1.2 User/Machine Systems: Human information processing.

INTRODUCTION
Joint attention is critical in social learning such as parent-child shared storybook reading. To acquire print-related skills, children and adults must coordinate their joint attention during dynamic interactions [6, 11, 14, 22]. This requires the partners to maintain a triangular attentional structure, i.e., attending simultaneously to the target of learning and among themselves. However, prior eye-tracking studies show that pre-reading children focus almost exclusively on illustrations while parents read from print texts [4, 5, 8, 9]. In fact, it is difficult for a child to know where on the page the adult reader is attending to, and vice versa. This limits the partners' ability to regulate their joint attention. We hypothesize that the lack of joint attention impedes the acquisition of print-related skills, and children will learn better if we can facilitate the regulation of joint attention.

The present study has two goals. First, we seek to objectively measure the joint attention in shared storybook reading, by simultaneously tracking the eye gaze of the parent and the child. 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 real-time 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 new sight words.

RELATED WORK
Print-related skills refer to children's knowledge of the rules for translating the particular writing they are trying to read into sounds, including letter knowledge, phonological awareness, knowledge of letter-sound correspondence, print knowledge, and sight word recognition [18, 20, 24]. There is converging evidence that a key to develop print-related skills is to engage children in a joint attention on print words. This can be achieved by pointing to words while reading and by having print-focused conversations [8, 10], both of which are key elements in a joint attentional

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

interaction. However, traditional reading strategies have some inherent limitations. One problem is that the two partners do not have accurate knowledge of where the other person is attending to at any moment, which makes attention regulation difficult and ineffective. In addition, most existing reading strategies are adult-centered - i.e., the parent regulates the child's attention without much information about the child's interest at the moment.

We conjecture the ideal solution to the lack of joint attention in the naturalistic shared reading is to provide partners consistent, individualized, and real-time feedbacks during reading. By externalizing adults' reading processes, the pre-reading children will have a model which they can mimic and internalize. And with a projection of children's attention and thinking process, adults will be able to scaffold and strategize pedagogical goals accordingly.

Short of a magic window into each other's mind, the state of eye-tracking technologies allows us to show where the partner is looking at 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. This process of helping people understand real-time eye movements is called the eye-gaze awareness training. Eye movements provide 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 [15, 16]. As previous studies showed that the coordination of joint attention is essential for communication, collaborative visual search and problem solving tasks [1, 12, 17], we expect that discovering children's real-time attention state may also 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. We expect that children will be more likely to repair their own flawed mental models if they recognize that theirs conflict with adults' reading model. Children may even start to follow adults' scanning patterns and pay more attention to texts, which in turn provides more teachable moments for parents to introduce print-related knowledge and skills.

Second, the scanning pattern of the eyes can be used to gauge children's cognitive processes. For example, recent psycholinguistic studies have shown that the eye typically follows the order of words in sentence comprehension [2, 7, 21]. It is intuitive that a child is not paying attention or is having difficulty comprehending a story if his or her eyes do not keep up with things mentioned by the adults. Adults may slow down the pace of reading and help children catch up in this situation.

Lastly, 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 two partners spend on reengaging joint attention. Gone is the need to ask "Are you looking at here?" because the answer is on the screen. The success or failure of a pedagogical attempt is immediately seen on the screen as well. Adults can give children more prompt and precise feedback when they watch children's real-time eye movements. If a parent has the goal of teaching print but notices that his/her child's eye movements only focus on pictures, the parent probably would not ignore this information and keep talking about the print by himself/herself. Instead, the parent may first follow the child's interest on pictures and then utilize verbal or nonverbal strategies to direct the child's attention to texts.

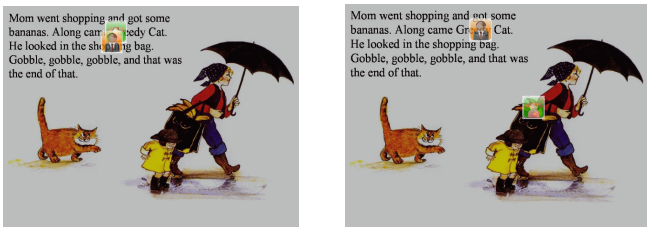
METHODOLOGICAL INNOVATIONS

The current study involves a series of methodological innovations. While there are a handful of published studies looking at children's eye gaze during shared book reading, none has investigated the correlation and contingency between eye movements of children and parents. Using two eye trackers simultaneously, 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.

Furthermore, we sought to intervene in the dynamic interactions by showing dyads the other partner's eye gaze positions. To improve pre-readers' understanding of reading, we presented a moving cursor on the computer screen which shows where adults are looking at any moment. We also investigated how parents use the real-time eye gaze information to regulate joint attention.

EXPERIMENT

There are three experiments in the current study. Experiment 1 is the baseline in which we tracked the eye gaze of both the parent and the child during the naturalistic shared reading. We define joint attention as when the partners look simultaneously at (or near) the same visual object on a page (see the examples in Figure 1). Children's



A. The dyad has join attention on texts. B. The dyad does not have join attention.

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

Figure 1: Examples of parent-child joint attention on one book page during the shared storybook reading.

sight word recognition was measured before and after the reading. We anticipate little joint attention on print and

therefore limited print learning. This serves as the control condition for subsequent intervention experiments.

Experiment 2 and 3 are two intervention experiments in which we investigated whether the real-time feedback of eye movements enhances parent-child joint attention and children's print-related 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 reading. The research methods of the three experiments are illustrated as follows.

Participants

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.

Materials

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. We used stuffed animals as props to ask children where adults look on a page (pictures vs. texts) and in which direction they read (left to right vs. right to left).

Apparatus

Two contact-free eye trackers, a Tobii X50 (see http://www.tobii.se) and an Eyelink 1000 system (www.eyelinkinfo.com), were used in the study. The Tobii X50 system is a video-based remote eye tracking system that makes no contact with the participant. It samples at 50Hz, and has a typical accuracy of approximately 1 visual degree (measured by repeated calibrations). The system uses infrared cameras to automatically capture eye images from a reading distance. Eyelink 1000 is also an infrared-based system, but with much higher accuracy (0.5 visual degree) and sampling rate (500 Hz). As a remote system, it also allows contact-less operations. 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 verbal and non-verbal 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 (1280x1024 pixels resolution) and Eyelink 1000 were put approximately 60cm away from the child; while another LCD monitor (1280x1024 pixels resolution) and Tobii X50 were put approximately 60cm away from the parent (see Figure 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.

Figure 2: The apparatus and experiment set-up.

The child and the parent were individually calibrated on the Eyelink and Tobii eye trackers. They were monitored throughout the study and recalibrated as necessary. Both remote eye trackers use the corneal reflection to compensate for head movements. Our past experience shows that neither exhibits substantial drifts that affect results in our experimental paradigms.

To the extent both eye trackers are accurately calibrated, aligning the gaze positions is straightforward. Because identical images are shown on both monitors, we took the gaze position in terms of screen coordinates and mapped to the other screen. This is essentially the same as the standard technique of displaying a gaze cursor on the same monitor as the participant is viewing.

Procedures

For all three experiments, each parent-child dyad read 3 books in four reading trials (order counter-balanced among participants). They read one same storybook in the first and fourth trial, and the other two 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 based on the pretest.

In Experiment 1 parents read children storybooks on screens but there was no eye gaze feedback for both partners. Experiments 2 and 3 involved showing a moving cursor on one of the participants' screen; the moving cursor indicated the location of the other person's eye gaze in real-time. We ensured children and parents understood the gaze indicator using an iSpy-like game. Even the youngest children had no problem understanding the correspondence. In Experiment 2, we showed the child where the parent was looking. In Experiment 3 the parent saw where the child was looking. In both conditions the other participant looked at a normal, static display of the page. To tease apart the

impact of the instructions from the moving cursor, we asked children in all three experiments to follow the parent's eye gazes while listening to stories, even though they could not actually see the eye gazes in Experiment 1 and 3.

Data Transcription and Coding

For the eye movement data, areas of interest (AOIs) were defined for pictures and texts on each page of the three books. An AOI includes a margin of approximately 1.3cm on each side of object it encloses. Eye movements fell outside of any AOIs (e.g., on the white background) were excluded from analyses. Data were also discarded when the eye trackers lost track, which could occur when the participant looked away from the monitor, moved the head rapidly, or closed his/her eyes. The average percentage of children's fixations on text AOIs in the first reading trial was compared with that in the fourth reading trial.

To measure the real-time joint attention during the reading session, we compared the distance of two partners' eye gaze locations with a cut-off value of 201.18 pixels which was determined for three reasons. First, 201.18 pixels correspond to the 80th percentile of the effective eye movement data points in the distance distribution of joint attention trials in the pretest. 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 shared book reading interactions were transcribed and coded with the InqScribe software. Adapting from the coding systems in previous studies [3, 13, 19, 23], we have developed a coding system to analyze 11 types of parent-child joint attention interactions, including parents looking at children, children looking at parents, parents' verbal attention regulation, parents asking children to look at specific words, parents teaching specific words, parents providing children specific feedback, parents pointing to words on the screen, children reading texts along with parents, children pointing to texts on the screen, children completing a sentence with parental prompt, and children talking about print.

RESULTS AND DISCUSSION

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 acquisition of print skills. Our data support both predictions.

Specifically, Experiment 1 showed that the average percentage of time children had joint attention with parents on texts in reading trial 1 was very small (2.91%). In the reading trial 4, when adults were asked to teach children three keywords, children significantly increased their joint attention on texts to 6.41% (t (36) = 2.48, p=.018). Children learned an average of 0.38 words as measured by pre- and post-test of word recognition.

Both interventions significantly increased parent-child joint attention on texts. In Experiment 2, where children saw parents' eye gaze, the joint attention increased from 5.35 % in reading trial 1 (no-intervention trial) to 22.7 % in reading trial 4 (intervention trial). The increase of 17.35% is significantly higher (p=.000) than the increase of 3.5% in Experiment 1 (see Figure 3).

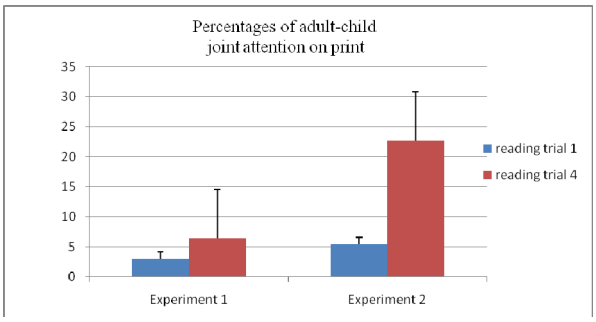


Figure 3: Percentages of parent-child joint attention on print from reading trial 1 to 4 between Experiment 1 and 2

More print-directed joint attention resulted more word learning: children learned on average 1.0 word, compared to 0.38 words in Experiment 1, p=0.02. This result indicates that children who received the eye-gaze awareness training learned more keywords from the pretest to the posttest than children who did not receive this training.

Moreover, seeing parents' eye movements also induced conceptual changes. Children were at chance in the pretest answering questions on where adults look in reading (62.96%) and which way they scan the text (51.85%). After the intervention the correct rate increased to 88.89% and 81.48%, respectively, and both are significantly above chance (p=.000 and p=.001 respectively).

Positive effects are also found in Experiment 3, in which the parent received real-time feedback on the child's visual attention but the child did not see the parent's eye gaze. Joint attention on texts increased from 3.48 % in reading trial 1 (no-intervention trial) to 12.87 % in reading trial 4 (intervention trial). The increase of 9.39% was significantly

higher than the increase of 3.5% in Experiment 1 (p=.012, see Figure 4).

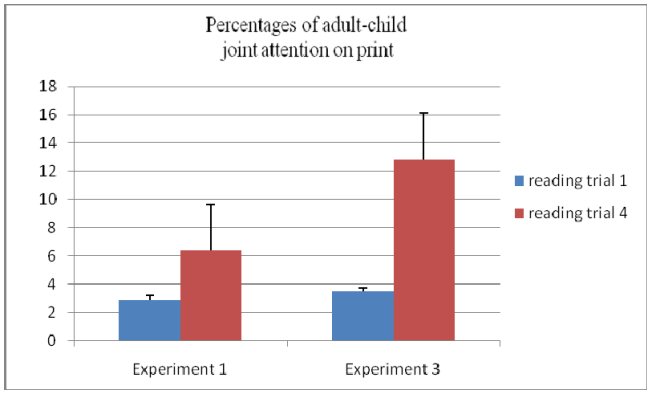


Figure 4: Percentages of parent-child joint attention on print from reading trial 1 to 4 between Experiment 1 and 3

Parents became more effective facilitating children's sight word learning. Children learned 1.25 words, significantly higher than that in Experiment 1(0.38 words, p=.000). This result further confirmed that children learned more sight words from the pretest to the posttest when parents used the eye-gaze feedback to effectively direct children's attention to words.

For the behavior analysis, the frequencies (times per minute) of these 11 parent-child interaction behaviors in reading trial 4 for all the three experiments are shown in Table 1.

Types of Behaviors	**Experiment 1 (baseline)**	**Experiment 2**	**Experiment 3**
Parents look at children	1.46(2.35)	.69(1.22)	1.56(1.82)
Children look at parents	2.48(2.86)	1.33(2.18)	1.97(2.24)
Parents' verbal attention	.26(.38)	.17(.34)	.78(1.1) *
Parents ask children to look at specific words	.53(.76)	.24(.54)	1.56(1.32) **
Parents point to words on the screen	.15(.53)	.01(.06)	.22(.61)
Parents teach specific words	.17(.44)	.08(.18)	.74(.88) **
Parents provide specific feedback	.01(.04)	0(0)	.62(.72) **
Children read along with parents	.25(.42)	.61(.93) *	.27(.55)
Children complete a sentence with parents' prompt	.32(1.15)	1.14(2.21) *	.27(.67)
Children point to texts on the screen	.13(.36)	.13(.33)	.39(.61)
Children talk about print	.27(.5)	.37(.67)	1.26(1.1) **

* P<.05, ** p<.01

Table 1: Mean (standard deviation) of the behavior occurrences per minute in reading trial 4 for all 3 experiments

To compare the changes of the frequencies of the behaviors from reading trial 1 to reading trial 4 between Experiment 1(control group) and Experiment 2 (child eye-gaze awareness training group), we did repeated measures ANOVAs using the average frequency of each behavior as the within-subjects variable (reading trial 1 vs. reading trial 4) and whether children received the eye-gaze awareness training as the between-subjects factor (child eye-gaze awareness training group vs. control group).

The results suggested that children in both groups significantly increased the occurrences of the behavior of *reading texts along with parents* from the first to fourth reading trial, but children who received the eye-gaze awareness training showed a significantly larger increase (p=.041). For children's *completing a sentence with parental prompt* type of behavior (e.g., children finished reading the last word of a sentence after parents stopped reading and waited for children's responses), children in both groups increased the occurrences of this behavior from the first to fourth reading trial, but children who received the eye-gaze awareness training showed a significantly larger increase (p=.046).

Overall, the comparisons between Experiment 1 and 2 indicated that with the eye gaze direction more tightly tied to the focus of join attention, children saw an external representation of reading processes unfolding in real-time and therefore responded more to parents' word teaching attempts, such as reading texts along with parents or reading the last word of a sentence with parental prompt. Children's increased responses to parents' reading

strategies further improved children's visual attention to print as well as their print learning outcomes.

We did the similar comparisons for the changes of the frequencies of the behaviors from reading trial 1 to reading trial 4 between Experiment 1 (control group) and Experiment 3 (parent eye-gaze awareness training group). For parents' *verbal attention regulation* type of behavior (e.g., adults saying "Look at the screen/words."), parents in both groups increased the occurrences of the behavior of verbally asking children to attend to print from the first to fourth reading trial, but parents who received the eye-gaze awareness training showed a significantly larger increase (p=.029).

For the *parents asking children to look at specific words* type of behavior (e.g., "Can you help me find the word 'cat' on the screen?"), parents on average increased the occurrences of this behavior from the first to fourth reading trial, but parents who received the eye-gaze awareness training showed a significantly larger increase (p<.001).

For parents' *teaching specific words* type of behavior (e.g., adults saying "The word 'book' starts with a letter 'B' and ends with a letter 'K'"; "The first word on the second line starts with a 'buh' sound, can you try to sound it out?"), parents on average increased the occurrences of this behavior from the first to fourth reading trial, but parents who received the eye-gaze awareness training showed a significantly larger increase (p=.002).

For parents' *providing specific feedback* type of behavior (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."), parents in the eye-gaze awareness training group significantly increased the occurrences of the behavior of providing children the specific feedback from the first to fourth reading trial (p=.000), but parents in the control group did not change much of this behavior.

For *children's talking about print-related things* type of behavior, children in both groups increased the occurrences of the behavior of asking or answering print-related questions from the first to fourth reading trial, but children whose parents received the eye-gaze awareness training showed a significantly larger increase (p=.000).

The above comparisons between Experiment 1 and Experiment 3 suggested when parents received the real-time feedback of their children's visual attention during shared book reading, they were better at regulating children's attention and at providing specific feedback following children's responses. Many parents spontaneously adjusted their reading strategies in response to children's real-time attention states. There were a lot more attention regulation types of interactions in the parent eye-gaze awareness training group, such that parents verbally asked children to pay attention to texts and constantly pointed to some specific words on the screen. Since parents could see where children looked at after a request or question, the redundancy of frequently asking children "Are you following mom and looking at the place I talked about?" was significantly reduced. Instead, parents directly provided children with more prompt and precise feedback such as "That's great, I can see you are looking at the right word." These reading behaviors are more efficient for accelerating communications between parents and children.

In addition, parents in the eye-gaze awareness training group showed more frequent and effective print teaching behaviors compared to parents in the control group, such as asking questions about particular letters or sound within a word, helping children spell or sound out a word, and commenting on print-related contents. Parents in the control group also used some strategies to teach print, but since they did not know where children were looking at, their reading strategies were neither effective nor efficient. For example, the control group parents usually looked at their children and talked about one specific word for a long period of time while their children already got bored and did not really look at that particular word. As a consequence children could hardly recognize that same word in the posttest. In contrast, with the help of real-time visual attention feedback, parents in the eye-gaze awareness training group had better opportunities to observe their child's attention state and fine-tune their interactions to increase child interest and participation. Specifically, these parents could reduce the time and energy spent on looking at children or verbally checking children's responses. Instead, they frequently and naturally incorporated varies efficient and well-organized literacy activities within children's short attention span. And children more readily learned those words when they had more frequent but shorter durations of learning experiences.

Similarly, children who experienced positive direction, coaching, and correction more easily attended to and internalized the information and skills that 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.

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 book reading, and study shared reading as a joint attentional interaction involving dynamic transactions between partners and real-time cognitive strategies within individuals. The data and methodology of this study would also be useful to a wide array of research topics on collaborative learning and interaction situations.

The new technology enabled us to investigate a number of issues in shared storybook reading. Building on prior research, we found that pre-reading children have limited joint attention with their parents while listening to storybooks. More importantly, parents typically have little

information about where children are attending to, and children have even less idea about how adults actually read. This results in a poorly regulated joint attentional interaction when it comes to learning print-related skills.

An important contribution of the present study is the intervention experiments, in which we remedy the joint attentional structure by leveraging the eye-tracking technology. By providing real-time feedback of the partner's visual attention, we demonstrate significant improvements in the amount of joint attention on print texts 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. Our intervention targets limitations in joint attention regulation in the traditional shared reading practice, but it is not specific to reading. To the extent learning involves joint attention (e.g., in math tutoring), the eye-gaze feedback may be an effective aid for learning. More importantly, data 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 acquisition of print-related skills.

REFERENCES

1. Brennan, S. E., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*, (2007), 1465-1477.
2. Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47*,1 (2002), 30-49.
3. Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. Learning from human tutoring. *Cognitive Science, 25,*4 (2001), 471-533.
4. Evans, M. A., & Saint-Aubin, J. What children are looking at during shared storybook reading - Evidence from eye movement monitoring. *Psychological Science, 16,* 11 (2005), 913-920.
5. Feng, G., & Guo, J. From pictures to words: Young children's eye movements during shared storybook reading. *Journal of Educational Psychology,* (2010, submitted).
6. Fletcher, K. L., & Reese, E. Picture book reading with young children: A conceptual framework. *Developmental Review, 25,*1(2005), 64-103.
7. Griffin, Z. M., & Bock, K. What the eyes say about speaking. *Psychological Science, 11,*4 (2000), 274-279.
8. Justice, L. M., Pullen, P. C., & Pence, K. Influence of verbal and nonverbal references to print on preschoolers' visual attention to print during storybook reading. *Developmental Psychology, 44,*3(2008), 855-866.
9. Justice, L. M., Skibbe, L., Canning, A., & Lankford, C. Pre-schoolers, print and storybooks: an observational study using eye movement analysis. *Journal of Research in Reading, 28,* 3 (2005), 229-243.
10. Justice, L. M., Skibbe, L., & Ezell, H. K. Using print referencing to promote written language awareness. In T. A. Ukrainetz (Ed.), *Contextualized language intervention: Scaffolding preK-12 literacy achievement*. Greenville, SC: Thinking Publications University (2006), 389-428.
11. Mundy, P., & Newell, L. Attention, joint attention, and social cognition. *Current Directions in Psychological Science, 16*, 5 (2007), 269-274.
12. Nüssli, M.-A., Jermann, P., Sangin, M., and Dillenbourg, P. Collaboration and abstract representations: towards predictive models based on raw speech and eye-tracking data. In *CSCL 2009*: Proceedings of the 2009 conference on Computer support for collaborative learning. International Society of the Learning Sciences, 2009.
13. Ortiz, C., Stowe, R. M., & Arnold, D. H. Parental influence on child interest in shared picture book reading. *Early Childhood Research Quarterly, 16*, 2 (2001), 263-281.
14. Pellegrini, A. D., & Galda, L. Joint reading as a context: Explicating the ways context is created by participants. In A. v. Kleeck, S. A. Stahl & E. B. Bauer (Eds.), *On reading books to children: Parents and teachers*. Mahwah, NJ Erlbaum (2003), 321-335.
15. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 3 (1998), 372-422.
16. Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10*, 3 (2006), 241-255.
17. Richardson, D. C., Dale, R., & Kirkham, N. Z. The art of conversation is coordination. *Psychological Science, 18* (2007), 407–413.
18. Scarborough, H. S., & Dobrich, W. On the efficacy of reading to preschoolers. *Developmental Review, 14*, 3 (1994), 245-302.
19. Sulzby, E. Children's emergent reading of favorite storybooks: A developemtal study. *Reading Research Quarterly, 20,* 4 (1985), 458-481.
20. Sulzby, E. Assessment of emergent literacy - Storybook reading. *Reading Teacher, 44,* 7 (1991), 498-500.
21. Tanenhaus, M. K. Integration of visual and linguistic information in spoken language comprehension. *Science, 307,*5711 (2005), 851-851.
22. Tomasello, M., & Farrar, M. J. Joint attention and early language. *Child Development, 57,* 6 (1986), 1454-1463.

23. Whitehurst, G. J., Fischel, J. E., Lonigan, C. J., Valdezmenchaca, M. C., Debaryshe, B. D., & Caulfield, M. B. Verbal interaction in families of normal and expressive-language-delayed children. *Developmental Psychology, 24*, 5 (1988), 690-699.

24. Whitehurst, G. J., & Lonigan, C. J. Child development and emergent literacy. *Child Development, 69*, 3 (1998), 848-872.

Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data

Shahram Eivazi
School of Computing
University of Eastern Finland
seivazi@cs.joensuu.fi

Roman Bednarik
School of Computing
University of Eastern Finland
bednarik@cs.joensuu.fi

ABSTRACT
Inferring high-level cognitive states during interaction with a user interfaces is a fundamental task in building proactive intelligent systems that would allow effective offloading of mental operations to a computational architectures. In this paper, we propose a system that uses real-time eye-tracking to measure user's visual attention patterns and infers behavior during interaction with a problem solving interface. Using a combination of machine learning techniques, computational modeling and eye tracking, we investigate 1) the differences among good and poor performance groups, and 2) the possibility of inferring distinct cognitive states such as planning or cognition. We employ and train a support-vector machine (SVM) to perform a classification task on a set of features computed from eye movement data that are linked to concurrent high-level behavioral codes based on think aloud protocols. We contend that differences among cognitive states can be inferred from overt visual attention patterns with accuracy highly over chance levels. We observe that such a system can also classify and predict performance with up to 87% recognition rate for an unseen data vector for two classes. We suggest a prediction model as a universal model for understanding of 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.

Author Keywords
Eye-tracking, machine learning, proactive intelligent systems

ACM Classification Keywords
H1.2. Models and principles: User/machine systems.

INTRODUCTION
Modeling human behavior is one of the main challenges to create new adaptive interfaces that can understand user behaviors based on relevant user information record. The traditional data collection methods for the modeling task, such as logs or verbal data, are often not completely reliable or applicable. For instance, it has been frequently argued that tasks such as reading, mental computations, and problem solving are difficult to be assessed by traditional methods such as verbal protocol [9].

Eye tracking is considered as a technology that provides an unobtrusive, sensitive, and real time behavioral index of ongoing visual and cognitive processes. New, reliable and more comfortable eye trackers have become available. The availability of new eye-trackers motivates HCI researchers to employ them as input devices in real time interfaces [10]. For example, the technology has been applied in eye-typing [17], object pointing and selection [21], gaming [23], or interaction with problem solving [2].

Previous research also shows that eye movements during the observation of complex visual stimuli are regular and systematic (e.g. Yarbus [27] and Rayner [20]) which gives a motivation for modeling cognitive and perceptual processes based on eye-movement data. For example, differences between skilled and novice users have frequently been linked to the differences in the eye-movement patterns.

Modern eye-tracking research tends to rest on the eye-mind hypothesis [11]; eye-tracking data are commonly considered as a measure of overt visual attention and that is linked to the internal processing. Analysis 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 [13]. Loboda and Brusilovsky [16] and Bednarik [3] argued that eye tracking can be applied in the area of user modeling and adaptive tools for improving the accuracy of prediction models. Loboda and Brusilovsky pointed to the advantages of eye movement data for on-line assessment of user meta-cognitive behavior. Conati and Mertena [4] showed that eye-tracking data improves the performance of probabilistic models in online assessment.

In this paper we describe the design and components of a system that employs eye-tracking data to model user performance and cognitive activities in an interactive problem solving task. The system consists of two prediction

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

models to provide a comprehensive recognition and unambiguous interpretation of eye gaze pattern in order to feed new intelligent user interfaces with behavioral predictions.

RELATED WORK

People apply a range of different 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. Understanding user's plans and goals in real time would enable us to significantly improve interactive systems. Therefore, in order to create interfaces that are more sensitive to user's needs, the user's cognitive states must first be invariably recognized.

Ericsson and Simon [6] supported the idea of applying think aloud data to understand cognitive processes. They assumed that think aloud reports are a reflection of the cognitive processes that generate user's behavior and action. So far it is not however clear whether we can model aspects of the human mind only with verbal protocol. In real-time systems data collection with verbal protocols methods is problematic, because think aloud utterances are often incoherent [6] and verbalizing thought is not natural in everyday situations. Van Someren et al. [25] argued that in many cases it is possible to combine think aloud method with other data collection methods. Think aloud method is used to report data. Later this data can be used to support and promote analysis of other methods.

Another data collection tool frequently applied to get insights into cognition is eye tracking. Analyses based on eye-tracking have several advantages over other protocols. Glöckner & Herbold [7] argued that in a problem solving experiment recorded data from users with eye tracking methods decrease the chance of influence on decision process.

Goldberg and Kotval [8] argued that eye tracking is one of the particularly strong methods in the assessment of users' strategies. They consider 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. [1]) it is generally accepted that eye movements, eye-fixations and the derived measures provide information about cognitive processes. For instance, Velichkovsky et al. [26] 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 (e.g. deliberate consideration of information and planning) [7].

Recent empirical data obtained from eye movement models (e.g. EMMA Salvucci [21]) provide a good motivation for building systems that can adapt to users interaction with the environment and learn from eye movement data.

Both user expertise and cognitive states have been previously modeled using eye-tracking data. Based on machine learning classification Liu et al. [15] explained the differences between experts and novices in building concept maps. Participants constructed collaboratively concept-maps of the content in the text for 20 minutes as their eye-movement data were recorded. Results showed 96% recognition rate for two distinct clusters (experts and novices). The authors reported that while higher-skilled participants concentrated on specific concepts longer, lower-skill participants had shorter attention spans and scattered gazes. In another experiment Liang et al. [14] claimed that a general Support Vector Machine (SVM) is a strong machine learning method for classification of human behavior, especially for detecting cognitive states via eye movement data. Authors demonstrate that driver distraction can be detected using driver performance measures and eye movement measures in real time.

In another study, Simola et al. [22] 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 most interesting title from a list. The model was trained on eye-tracking data and achieved an accuracy of 60.2%.

MAPPING GAZE DATA TO SEMANTIC CONCEPTS

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

In this study we thus employ eye tracking to reveal such relevant information from user's ocular behavior. Gaze data are associated with human cognition states by using think aloud protocol as a ground truth. The presented method and system includes the following main parts: verbal protocol analysis of the cognition, feature extraction and mapping to the verbal protocols, and machine learning method for building associations between the two. First, we code all think-aloud data by listening to the user's speech during interaction. We applied a coding scheme based on O'Hara & Payne [19] method that is based on Ericsson's approach [12] and has also been applied with modifications in other studies (e.g. Morgan et al. [18]). In the second phase, we propose a novel way of mapping gaze-based data to qualitative differences in the corresponding think-aloud protocols. We compute a 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 1 presents the computational architecture of the proposed approach.

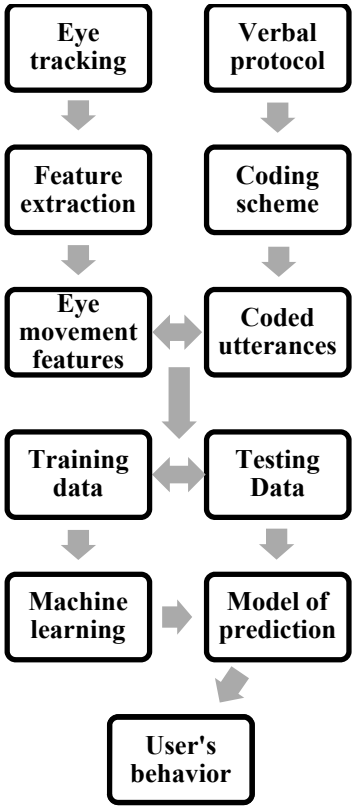


Figure 1. Procedures of the proposed mapping.

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 the automatic modeling and classification methods proposed in this paper.

In the rest of this paper we conduct a feasibility evaluation of this approach in the domain of interactive problem-solving. We propose the feature set and we evaluate the accuracy of the approach.

EXPERIMENTAL DESIGN AND PROCEDURE

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. [2]. Similar settings have been used in numerous studies investigating interactive human problem solving. 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 200×200 pixel (for each tile: width was 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 desired state of the puzzle is shown in Figure 2. Figure 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 1280×1024 and the viewing distance 60 cm [2].

Data from two participants have been removed because of low quality of eye tracking data. Preliminary eye movement data analysis has been performed with Clearview version 2.7.1 (http://tobii.se), with a default setting for fixation identification algorithms. MATLAB version R2009b and LibSVM Matlab toolbox of [10] have been used for the data analysis.

DATA ANALYSIS

To achieve the goal of predicting user characteristics and skills through the eye movement data, two main analysis techniques had to be carried out. First, outcome measures have to be defined and computed, including feature extraction and clustering of the features. Second task consisted of creation and validation of the prediction model.

Outcome measures

To address the first problem (outcome measures), verbal data were classified into six categories based on O'Hara & Payne [19] with a slight modification. The classification categories described qualitatively different utterances: *Cognitions* referred to statements describing what concrete and specific information a participant is currently attending to. *Evaluations* were conceptually similar to cognitions while, they were less accurate about the object of interest. In addition, when participants were referring to how well

Figure 2. Goal state of 8- tiles puzzle game.

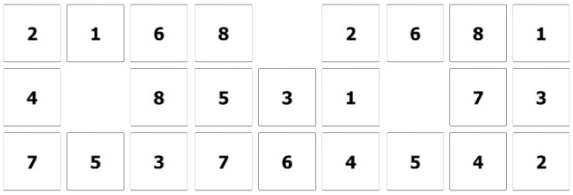


Figure 3. Three initial states for 8- tiles puzzle game.

they performed or what is the general situation in the problem-space, we coded that utterance as belonging to evaluations. *Plans and planning* were utterances containing a plan development, its specific goals and detailed actions to be taken next. *Intentions*, on the other hand, were utterances describing the general aims, without a specific descriptions how to achieve them. *Concurrent move* utterances referred to description of the changes in the problem along the manipulation with it. Finally, we applied a category of *not applicable* for other utterance; however, we do not consider those data in this analysis. More detailed description can be found in [19].

The unit of analysis was one sentence. Two independent coders conducted the coding and achieved the inter-rater agreement of 86%. Since all classified utterances included a time stamp, the action times spent on utterances were taken into an account as feature in this study.

Of all three trials and all participants, the coding yielded a total of 249 data points for Cognition states, 339 data points for Evaluation activities, 105 data points for Planning, 235 data points for Intention related utterances, and 318 utterances containing the descriptions of concurrent moves.

In this study the outcome features of eye movement data were based on fixation duration and the location of the eye movements with respect to the screen coordinates. The eye movement features that used in this experiment are listed in Table 1. Similarly as the coded utterances, eye-movement data carried a timestamp, enabling their easy mapping with the verbal protocol. In our case, each feature vector was computed from eye-movement data for the duration of the corresponding utterance. For example, during three seconds of one coded state, such as 'evaluation', we computed all eye movement features from this interval.

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 showed constantly at the left bottom of the screen.

Prediction model

To address the second task (prediction model) we employ a well-established machine learning approach. SVM is a standard and frequently applied tool that has been previously shown performing well for various classification tasks [17]. SVM has been successfully used in detection, verification, recognition, and information retrieval from a range of datasets [14]. Liang et al. [14] presented three arguments 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.

Eye movement feature	Description
Mean fixation duration	The average time of fixation duration in each state of coding scheme.
Sum of fixation duration	Sum of times of fixation durations in each state of coding scheme.
Mean path distance	The average distance of two consecutive fixations in each coding state based on eyes' coordinates
Total path distances	Summation distance of eye movements in each coding state based on eyes' coordinates
Number of fixations	Number of fixation during of each coding state
Fixation rate	Number of fixation divided by the duration of each coding state
Visited tiles (rate)	Number of visited tiles divided by number of fixation in each coding state

Table 1. List of eye movement features.

We built two prediction models. The first model learns the patterns of human cognitions (five states) and eye movement features. The second model searches for patterns between data vectors originating from different performance groups (two or three classes, high-, medium-, and low-performing participants) and eye movement features. The task is to predict, to which performance group any given data vector belongs.

For the former, the ground truth was labeled for each class (five coding states) in the sample data. For the latter, the ground truth was established by computing the task completion times. The data were split into training and testing datasets so that the data from two trials were considered as training, and the remaining trial as testing (unseen) data. Both training and testing data were normalizing between [0 1] with the same method as presented below:

$$\delta = \frac{d - d^{min}}{d^{max} - d^{min}},$$

where $\delta = normalized\ vector$, $d = original\ vector$.

Normalization of data was applied in two ways for the two prediction models. In the case of cognition recognition, we defined minimum and maximum values in training and testing dataset separately. In the case of performance recognition we defined the minimum and maximum values in the training and testing datasets for each participant individually.

We used Libsvm Matlab toolbox developed by Hsu et al. [10] to build the prediction models. In order to find best

parameter for the model, we employed a 3-fold cross validation method. Experimentally, we learned that for n>3 in n-fold cross validation, accuracy has not been changed significantly. Therefore, the training data has been divided into three subsets. Consequently, one subset was tested by using the model based on the remaining datasets (two subsets). At the end, cross validation accuracy was equal to the percentage of data that was correctly classified [10]. C-SVC support vector classification with RBF kernel has been applied.

RESULTS AND DISCUSSION

In this paper 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. We first present the results related to the classification and inference of cognitive states alone, then we introduce the classification based on performance differences, and finally we present a combination of the two.

A complete description of the mean values and standard deviations of the features computed for each of the cognitive states can be found in [5][1]. The differences in individual features related to the cognitive states were generally small and the features contained great variances.

The baseline performance was established as a classification accuracy of a majority classifier. Given the fact that most of the classes belonged to the Evaluation class, the majority classifier would perform with accuracy of 27%. Table 2 shows the recognition accuracy of the SVM for the five cognitive states (cognition, evaluation, planning, intention and concurrent moves). On unseen samples the accuracy was about 53%. These results are reported also in [5] and we report them here for completeness.

The breakdown of the results indicates that cognition is the hardest activity to automatically recognize, as seen from the confusion matrix (see Table 3). By removing cognition-data from the dataset we were able to increase the recognition accuracy up to 64%.

In addition to the cognitive states prediction, we investigate how well any given data vector informs about the originating performance group. All users were divided into three groups; we denoted these groups as high-performance, medium-performance, and low-performance groups.

[1] Before the original paper will be published (ACM Press), we provide a reference to a table describing the data: http://cs.joensuu.fi/~seivazi/koli.JPG

	Cross validation	**Unseen data**
Accuracy	75.84	53.25
Penalty parameter of the error term in RBF kernel (C)	64	64
Parameter of RBF kernel	0.25	0.25

Table 2. Cognition state activities recognition.

Prediction outcome %						
Actual class		Cognition	Evaluation	Planning	Concurrent move	Intention
	Cognition	2.6	24.67	3.9	67.53	1.3
	Evaluation	1.9	96.19	0	0.95	0.95
	Planning	7.69	25.64	48.72	12.82	5.13
	Concurrent move	5.68	28.41	1.14	64.77	0
	Intention	15.79	43.42	0	6.58	34.21

Table 3. Confusion matrix.

In this analysis, the high-performance group contains four participants who solved the puzzle with average tasks completion time less than 120 seconds.

The medium-performance group contains five participants who solved the puzzle with average tasks completion time between 120 and 240 seconds. Finally, the low-performance group contains three participants who solved the puzzle with average tasks completion time more than 240 seconds. The pair wise differences in the average completion times between the groups were significant.

The features for each of the groups are compared in Table 4. It is worth to note that the action time on utterances for the high-performance group was much shorter than that of the low-performance group and that the standard deviation of the high-performance group was low.

Other observation relates to the fact that while the high-performance group had lower number of fixations, they had longer fixation durations. In other cases, however, it is hard to visually spot eventual patterns of differences between the groups, partly due to the great variances.

Table 5 presents the recognition accuracy of predicting into which of the three groups an arbitrary vector of data belongs. The accuracy of 66% can be considered relatively low, however the baseline classifier would achieve only 55% accuracy.

To test the influence of the data from the medium-performing group, we removed the data and conducted the classification again only for the two remaining groups. We speculated the medium-group dataset can contain such

Groups (Number of participants) Type of feature	High-performance (4)		Medium-performance (5)		Low-performance (3)	
	Mean	SD	Mean	SD	Mean	SD
Mean action time on utterance (ms)	4336.03	2943.31	7495.02	9026.77	7399.38	8866.92
Mean number of fixations average	10.85	7.97	18.14	21.91	16.08	17.56
Mean fixation duration average (ms)	266.64	114.85	288.52	107.52	206.67	58.91
Mean path-distance average	168.87	68.42	169.48	84.33	191.13	91.62
Mean fixation duration summation (ms)	2704.34	1887.78	4922.55	5896.40	3213.65	3367.36
Mean path-distance summation (pixels)	1819.94	1472.00	2739.43	3135.48	3130.68	3836.88
Mean fixation rate (Hz)	2.52	0. 69	2.43	0.60	2.54	0.83
Mean visited tiles rate	0.511	0.21	0.458	0.23	0.491	0.24

Table 4. Comparison of features among High, Medium, and Low-performance groups. SD= Standard deviation.

feature-spaces that could be overlapping with data either of the two other groups.

The accuracy of 87.5% (Table 6) shows that the chance of correct prediction whether a data point belong to either to a high or low-performance group is indeed high. In fact, we achieved 66.18% accuracy of correct recognition for data vectors from the high-performance group and 96.79% accuracy for the low-performance group. In other words, if

	Cross validation	Unseen data
Accuracy	80.82	66.48
Penalty parameter of the error term in RBF kernel (C)	256	256
Parameter of RBF kernel	1	1

Table 5. Three-group recognition rate (Low, Medium, or High-performance group).

	Cross validation	Unseen data
Accuracy	96.41	87.50
Penalty parameter of the error term in RBF kernel (C)	32	32
Parameter of RBF kernel	1	1

Table 6. Two-group recognition rate.

the classifier processed a data point from a low-performing user, in about 97% the data were correctly classified. The smaller size of the high-performance group dataset (fewer participants and faster completion time leading to less data for training) can be the main reason for lower recognition rate for the high-performance group. An improvement can be achieved by adding a weight in SVM parameters (w=1.5) to the high-performance group. In that case, the individual accuracy for high-performance group users can be increased to 73.53%, however, with a trade-off as the accuracy for the low-performance group users slightly decreased to 94.87%.

Finally, we conducted the classification task separately for the five cognitive states in each performance group, see Table 7. The results show that recognizing cognitive activities of high-performers is rather difficult; again, the reason can be found in the small sample size. On the other hand, in about 75% of cases of medium- and low-performing users the classification correctly predicted what cognitive activity a user is currently undertaking. While the recognition rates are still relatively low in absolute values, they are still high when compared to a baseline recognition rate.

Descriptive statistics of Table 7 is shown in the Figure 4 and Figure 5. The Figure 4 shows the mean path distance for each performance group. The chart presents the fact that path-distance shows a U-shape behavior in at least three cognitive states: planning, concurrent move, and intension states are characterized by similar means and variances in the high- and low-performing groups, while the medium-performance group shows a decrease in the measure. The U shape behavior repeats in the Figure 5 for the measure of the rate of visited tiles, however the variance in the rates of visited tiles is high.

CONCLUSION AND FUTURE WORK

Computing and HCI researchers rooted in cognitive-science tradition frequently assume that the mind consists of mental representations and structures comparable to computer data structures, and it executes computational procedures similar to computational algorithms [24]. While we do wish to remain neutral to these views, the results presented here let us to suggest that at least some sub-part of cognition and user traits can be modeled effectively using traditional computational principles and methods.

	Baseline	Cross validation	Unseen data
High-performance (4)	53	53.38	36.36
Medium-performance (5)	53	73.30	37.10
Low-performance (3)	42	75.34	47.40

Table 7. Accuracies of cognitive state recognition for five cognitive states among High, Medium, and Low-performance groups. Size of group in parenthesis.

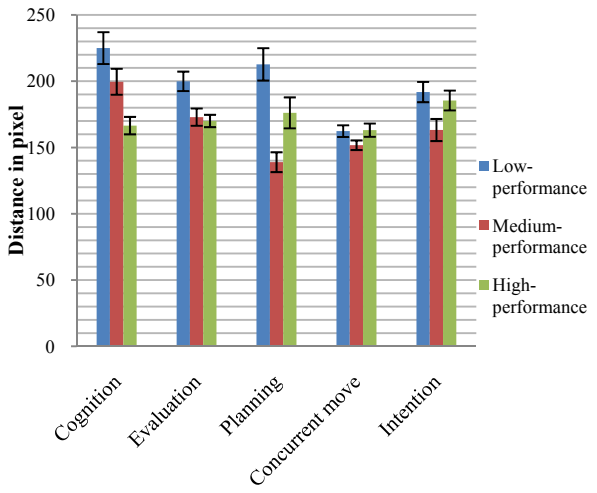


Figure 4. Mean path distance for each performance group.

We applied a SVM-based classification to predict, firstly, problem-solving cognition states and, secondly, 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 combined the approaches of machine learning methods, cognitive science, and HCI to describe a design and components of the real-time eye-tracking system for measuring user's visual attention patterns and inferring user's behavior during interaction with a problem-solving interface.

The novel result presented here shows that although the differences in problem-solving and related eye-movement data are subtle and multidimensional, they can be automatically recognized using SVM classification, with more than 87% accuracy.

This leads us to a conclusion that prediction of the user performance is possible, can be automated, and that the eye movement data carry important information about the skills of participants.

While the accuracy of classifications of cognitive activities was not extremely great, our finding shows that eye movement data carry important information about the problem solving process. We believe that increasing the sample-size to feed the training system can improve the accuracy. In addition, in the experiment we assumed that users' utterances always belong to whatever action they had just taken. The analysis of the verbal data showed that this was not always the case, particularly for high-performing experts. The expert participants often had to be prompted by the observers to talk, thus some of the thoughts were not captured in the protocol and some of the utterances that they shared were not aligned with the current eye-

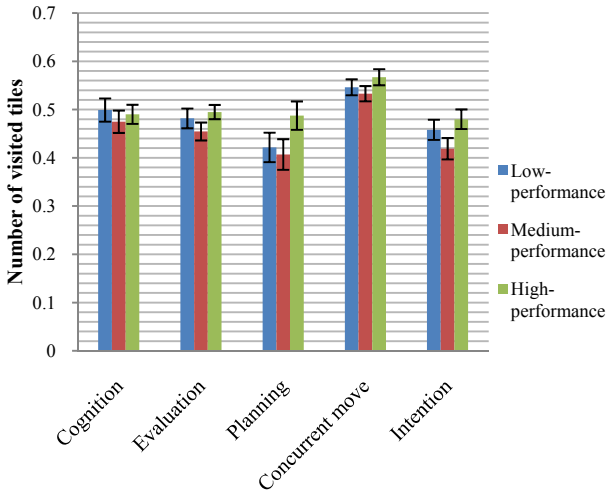


Figure 5. Mean rate of visited tiles for each performance group.

movement data simply because the style of the verbalization changed from concurrent thinking to retrospection. We plan to take this consideration into account, by both simplifying the coding protocol and extending the boundaries of the sample window to include the previous samples into mapping.

The future steps of this research include a development of a real-time system that dynamically captures and classifies user traits based on eye-movement data. In the domain of problem-solving this will enable us to build an intelligent environment that closely follows the user action and can proactively provide guidance, for example for the purposes of learning.

REFERENCES

1. ANDERSON, J. R., BOTHELL, D., AND DOUGLASS, S. 2004. Eye movements do not reflect retrieval: Limits of the eye-mind hypothesis. *Psychological Science 15*, 225–231.
2. BEDNARIK, R., GOWASES, T., AND 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. *Journal of Eye Movement Research*, 1–10.
3. BEDNARIK, R. 2005. Potentials of Eye-Movement Tracking in Adaptive Systems. In Proceedings of the

4th Workshop on the Evaluation of Adaptive Systems, held in conjunction with the 10th International Conference on User Modeling (UM'05), 1-8.

4. CONATI, C. AND MERTEN, C. 2007. Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. *Knowl. Based Syst.* 20, 557–574.
5. EIVAZI, S. AND BEDNARIK, R. 2010. Inferring problem solving strategies using eye-tracking: System description and evaluation. In *Proceedings of Koli Calling, International Conference on Computing Education Research, ACM press*.
6. ERICSSON, K. A. AND SIMON, H. A. 1993. *Protocol Analysis: Verbal Reports as Data Revised Edition*. MIT press.
7. GLÖCKNER, A. AND HERBOLD, A.-K. 2010. An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*.
8. GOLDBERG, J. H. AND KOTVAL, X. P. 1999. Computer interface evaluation using eye movements: methods and constructs. *Industrial Ergonomics 24*, 631–645.
9. SURAKKA, V., ILLI, M. AND ISOKOSKI, P. 2003. Voluntary eye movements in human-computer interaction, chapter 22, page 471. North Holland.
10. HSU, C.-W., CHANG, C.-C., AND LIN, C.-J. 2003. A practical guide to support vector classification. Tech. rep. Department of Computer Science and Information Engineering, National Taiwan University
11. JUST, M. A. AND CARPENTER, P. A. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 441–480.
12. ERICSSON, K. A. 1975. Instruction to verbalize as a means to study problem-solving processes with the 8-puzzle. Tech. rep., Department of Psychology, University of Stockholm.
13. KALLER, C. P., RAHM, B., BOLKENIUS, K., AND UNTERRAINER, J. M. 2009. Eye movements and visuospatial problem solving: Identifying separable phases of complex cognition. *Psychophysiology 46*, 818–830.
14. LIANG, Y., REYES, M. L., AND LEE, J. D. 2007. Real-time detection of driver cognitive distraction using support vector machines. *Intelligent Transportation Systems, IEEE Transactions on 8*, 340 – 350.
15. LIU, Y., HSUEH, P.-Y., LAI, J., SANGIN, M., N¨USSLI, M.-A., AND 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*, 898 –901.
16. LOBODA, T. D. AND BRUSILOVSKY, P. 2010. User-adaptive explanatory program visualization: evaluation and insights from eye movements. *User Modeling and User-Adapted Interaction 20*, 191–226.
17. MEYER, D., LEISCHA, F., AND HORNIKB, K. 2003. The support vector machine under test. *Neurocomputing 55*, 169–186.
18. MORGAN, P. L., WALDRON, S. M., KING, S. L., AND 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 4557*, 115–125.
19. O'HARA, K. P. AND PAYNE, S. J. 1998. The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology 35*, 34–70.
20. RAYNER, K. 1998. Eye movements in reading and information processing:20 years of research. *The American Psychological Association*, 372–422.
21. SALVUCCI, D. D. 2001. An integrated model of eye movements and visual encoding. *Journal of Cognitive Systems*, 201–220.
22. SIMOLA, J., SALOJÄRVI, J. AND KOJOC, I. 2008. Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research* 9, 237–251.
23. SMITH, J. D. AND GRAHAM, T. C. N. 2006. Use of eye movements for video game control. In *ACM International Conference Proceeding Series*.
24. THAGARD, P. 2007. Cognitive science, Stanford encyclopedia of philosophy. http://stanford.library.usyd.edu.au/entries/cognitive-science/
25. VAN SOMEREN, M. W., BARNARD, Y. F., AND SANDBERG, J. A. 1994. *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. Academic Press.
26. VELICHKOVSKY, B.M., ROTHERT, A., KOPF, M., DORNHOEFER, S.M. AND JOOS, M. 2002. Towards an express diagnostics for level of processing and hazard perception. *Transportation Research* 5, 145-156.
27. YARBUS, A. 1967. Eye movements and vision. *New York, Plenum Press*.

Gaze and Conversation Dominance in Multiparty Interaction

Yuki Fukuhara
Graduate School of Science and Technology
Seikei University
Musashino-shi, Tokyo 180-8633 Japan
dm106221@cc.seikei.ac.jp

Yukiko Nakano
Dept. of Computer and Information Science
Seikei University
Musashino-shi, Tokyo 180-8633 Japan
y.nakano@st.seikei.ac.jp

ABSTRACT
With the goal of designing conversational agents that can join and manage conversations with multiple participants, in this paper, we conduct an experiment to collect multiparty conversations with a virtual agent, and recognize head direction as each participant's focus of attention. Then, we analyze how gaze and mutual gaze affect floor management and conversation dominance; we assumed that the most dominant participant may control the participation framework and lead the conversation to make a decision. Based on the analysis, we found that turn-releasing success ratio is different depending on the levels of conversation dominance. We also found that the frequency of gaze and mutual gaze are different depending on the participation roles and conversation dominance. These results suggest that gaze behaviors are strongly related to conversation dominance, and may become a good predictor of dominance in multiparty conversations.

Author Keywords
Empirical study, eye-gaze behavior, conversation dominance, Wizard-of-Oz experiment.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
In public information kiosk systems which are used in museums, shopping malls, and entertainments, it usually happens that a group of people come to the system and operate it while talking with each other. Therefore, to build conversational agents that can serve as public information kiosk, such systems are required to be capable of managing multiparty conversations. However, little has been studied about interaction between a group of users and an agent. Thus, for the purpose of building conversational agents that can use bodily expressions, computational models of nonverbal behaviors in multiparty conversations are necessary.

Figure 1. Three people joining a multiparty conversation

Figure 1 shows a snapshot of a multiparty conversation by three people. Based on the face direction of each participant, we can assume that person A and person B are talking with each other, and person C is not a speaker or a hearer. If this interaction continues, it may happen that the person C cannot get enough chance to contribute to the conversation. It may also happen that one person leads and dominates the conversation, and other participants follow the leader.

These typical phenomena in multiparty conversations may be useful in managing multiparty conversation by a system or an agent. For example, if the system needs to finish the conversation, it would be effective to persuade the conversational leader to make a decision and finish this conversation as soon as possible. On the contrary, to help an inferior person get a chance to tell her/his opinion, the agent may give a turn to her/him.

With the goal of designing conversational agents that can join and manage conversations with multiple participants, in this paper, first, we conduct an experiment to collect multiparty conversations with a virtual agent, analyze head direction as each participant's focus of attention, and discuss how gaze and mutual gaze relate to floor management and conversation dominance.

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

RELATED WORK

Goffman [1] discussed the concept of participation in conversations, and based on his discussion, Clark [2] distinguished participants from overhearers. In his definition, participant includes "speaker" and "addressee (hearer)" as well as others taking part in the conversation but not currently being addressed. The non-addressed participant is "side participant". All other listeners are "overhearers", who have no rights or responsibilities in a given conversation. The role of each conversational participant is automatically assigned once a speaker starts speaking. At this moment, the addressee is chosen by the speaker, and the rest of the participants become side participants.

Previous studies showed that the participant's gaze behaviors are different depending on her/his role. People gaze more while listening (75%) than while speaking (41%) in dyad conversations [3]. Analyzing multiparty conversations, [4] reported similar results that about 1.6 times more gaze occurred while listening than while speaking. However, when a person starts speaking to all other participants, her/his gaze is distributed to all of them. In this condition, the total percentage of gaze (while speaking) rises to 59% of the time. Then, [4] applied these findings to a design of mediated communication.

The relationship between dominance and conversational behaviors was studied in [5], and they found that in dyad conversations, less dominant person is more likely to be the first to break the mutual gaze. The higher status person gaze roughly the same percentage of the time while listening and speaking, whereas the lower status person gazes relatively more while listening [6]. Moreover, dominant person who is in higher status and has more power initiates speech more and controls the whole interaction [7]. However, the notion of dominance in these studies indicates a stable social relationship, and not exactly the same as conversation dominance focused in this study where there is no power or status difference among the participants and the dominance is temporally evoked to perform the experimental task.

In studies of multiparty conversations with conversational agents, interactions between one user and multiple agents were mainly studied. [8] developed a conversational presentation system that generates conversations between two agents that answer the user's questions. [9] proposed a multiparty conversational system that can manage negotiation conversation among three parties: two virtual agents and a user.

Nagao et al [10] proposed a plan-based multiparty conversation system for two users. More recently, Bohus et al [11] proposed an open-world dialogue system where multiple people with different and varying intentions enter and leave, and communicate and coordinate with each other and with interactive systems. They also proposed a method to predict user's intention of engagement in conversation in open-world context [12]. Huang et al [13] proposed a quiz agent where a group of users join the game, and the system estimates whether the users are excited about talking with the agent. However, the previous work has not exploited gaze information to interpret multiparty participation framework, and not applied a multiparty gaze model to conversational agents.

Figure 2. Animation agent used in the experiment

EXPERIMENT

We conducted an experiment to collect multiparty conversations with a virtual agent in a Wizard-of-Oz setting, and investigate how participation framework is dynamically changed, and what types of verbal/nonverbal behaviors may be useful as a predictor of conversation dominance.

Task

A group of three people participated in the experiment as a subject group. The task of the subject group was to discuss and decide where they would go out to enjoy a weekend. To collect information about the visiting places, they could ask questions to the agent displayed in front of them. The agent answered the questions, and also recommended some places to visit. The picture of the agent is shown in Figure 2. To experimentally control the participation attitude towards the conversation, in each session, one of the subjects played a role of person who cannot go out that weekend. We expected that this person would participate in the conversation less dominantly and positively than the others, and has more possibility of becoming a side participant.

Moreover, the other two subjects were assigned different goals in selecting the places. For example, one subject was instructed to go to a shopping place, and the other was instructed to go to a scenery place. Thus, their task was to discuss and choose two places that satisfy each member's requirement as much as possible. The roles of the subjects were changed every session. The subject who could not go out was also different for each session. To motivate the subjects, we instructed them that they can get more rewards if they chose good places that satisfied all the requirements.

Procedure

We actually had another experimental condition that four people had a conversation. So, in experimental procedure, a group of subjects had 4 conversations for each condition

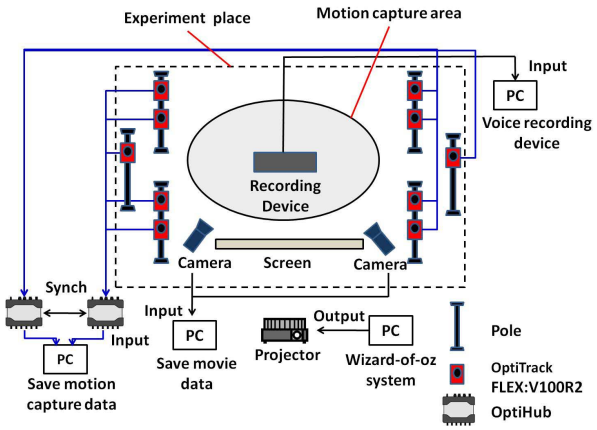


Figure 3. Experimental setting

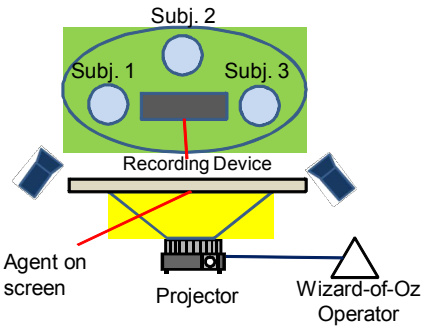


Figure 4. Proxemics of the subjects

and 8 conversations in all. The experimental contents are different in all the conversations. So, we had 8 kinds of experimental contents, and they were randomly assigned to each condition. Note that we only analyzed the corpus for three party conversations in the rest of this paper.

Experimental Setting

Figure 3 shows the setting of the equipments used in this experiment. We used 10 cameras for OptiTrack motion capture system that were mounted on 6 poles. The animated agent was displayed on a 100 inch screen in front of the subjects. When a subject asked a question to the agent, an experimenter (wizard) understood that question and chose one of the answer menus or typed in some texts to answer the question. Then, the spoken response was produced through TTS. The agent did not display any facial expressions or gestures, but her lip movement was synchronized with speech. The interaction was video-taped using two Hi-vision cameras. Each subject wore a sweater with markers for motion capturing (see Figure 1). So, the motion capture system measured upper body motions for each subject. Speech sound for each subject was recorded through individual microphones. The proxemics of subjects and coordination of audio and video equipments are shown in Figure 4. As shown in the figure, the subjects were instructed to stand around the audio device in front of the screen, so that they can see each other and the agent.

Data

From the experiment described above, we collected the following data;

- Video data shot from right front and left front of the subjects
- Speech sound for each subject
- Upper body motion data measured by a motion capture system
- Score for conversation dominance. After the experiment, we showed the video data to 5 people who did not join the experiment, and asked to rate each subject in terms of conversation dominance by asking like "Who is leading the conversation", or "who is the leader of the conversation?" Then, we calculated the average of rating scores, and used the values as the conversation dominance score for each subject.

ANNOTATION

Participation role

In multiparty conversations, to specify the participation framework, it is necessary to specify the hearer as well as the side participant. Therefore, to analyze the structure of multiparty interaction, and relationship between the subjects and their participation roles, we annotated the participation framework for each utterance.

First, we identified the time period of each utterance using Praat, and read the Praat data into Anvil annotation tool [14] to annotate utterances. Then, looking at the video, we annotated the participation roles: speaker, hearer, and side participant. When the speaker talked to the other two subjects, both of them were annotated as hearer. In addition to the participation roles, we annotated the utterance type: speak to someone, respond to someone, and soliloquy. In this paper, since we focus on human multiparty conversations, we did not analyze interaction between the user(s) and the agent.

In addition to the utterance annotation, we identified and annotated turn boundaries. A sequence of utterances whose speaker was identical and a pause between the utterances was less than 2 seconds was identified as one turn.

Automatic Annotation of Gaze

Previous studies revealed that gaze distribution is deeply related to turn taking/releasing, and conversational coordination [15, 16]. Therefore, we think that gaze data is indispensable in analyzing multiparty conversations. However, manually annotating gaze by looking at video is very time consuming. Thus, in this study, we estimated the gaze direction from the head motion data obtained from a motion capture. Decision tree algorithm implemented in Weka (J48) was applied to the head motion data: x, y, z

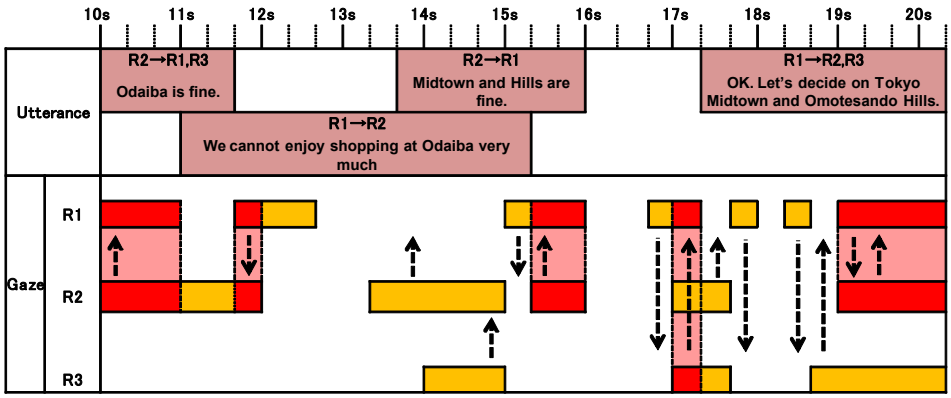


Figure 5: Example of interaction

position and rotation. We prepared training data by manually annotating gaze behaviors for each subject for 5 sessions. The decision tree judges whether the subject looks at the subject standing at the left position, the right position, or the center position towards the screen. By employing the leave-one-out method, four sessions were used for training and one for testing. We examined all the combinations of training and test sets. The averages of precision, recall, and F-measure are shown in Table 1. Since the accuracy of the decision tree was sufficient, we decided to use this model to automatically annotate subject's gaze directions.

Table 1. Accuracy of estimating gaze direction for each subject

Subject (Position)	Precision	Recall	F-Measure
Subj.1 (right)	0.872	0.914	0.893
Subj.2 (center)	0.927	0.922	0.925
Subj.3 (left)	0.954	0.95	0.952

ANALYSIS

In this section, we analyze the participant's behaviors with respect to the order of dominance in conversation. The order was determined based on the score of the conversation domination, which was described in the previous section. We call a conversational participant who got the highest score "rank 1 participant (R1)", a participant who got the second highest score "rank 2 participant (R2)", and a participant who got the lowest score "rank 3 participant (R3)". The averages of conversation domination scores are 1.16, 2.08, 2.75, for R1, R2, R3 respectively. The scores indicate clear difference between ranks. To check the validity of experimental control, we calculated how many people playing as a person who cannot go out were actually ranked as R3. As a result, in 5 out of 8 sessions, subjects who cannot go out were ranked as R3. For the rest of three sessions, they were ranked as R2. Subjects who cannot go out have never been ranked as R1. Thus, we believe that the experimental instruction successfully controlled the subjects' interaction.

Example of eye gaze exchanges in collected corpus

An example of multiparty interaction is shown in Figure 5. The upper track shows subject's utterances, and the lower track shows gaze behaviors for R1, R2, and R3 subject respectively. For instance, from 10s to 11s, R1 and R2 established mutual gaze (R1 looks at R2, and R2 looks at R1) while R2 was speaking. At 13.3s to 15s, R2 was looking at and talking to R1. At the same time (14s to 15s), R3 was looking at R2 as the speaker, but did not get R2's attention. Overall, it seems that R1 get the highest attention among three participants, and mutual gaze was more frequently established between R1 and R2, while R3 had less chance to establish mutual gaze. In the rest of this paper, we will more closely investigate such gaze behaviors.

Distribution of participation role

First, we investigated whether the distribution of participation role was different depending on the order of

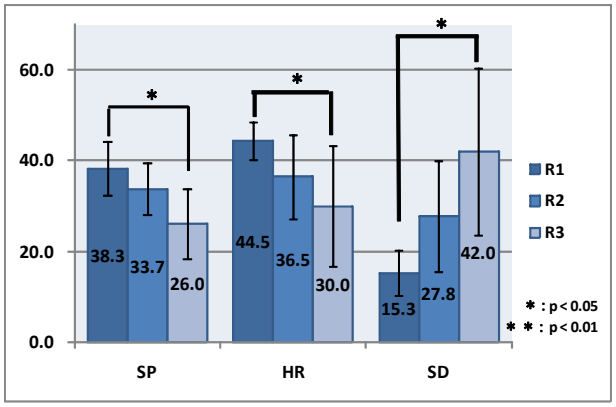


Figure 6. Distribution of participation role

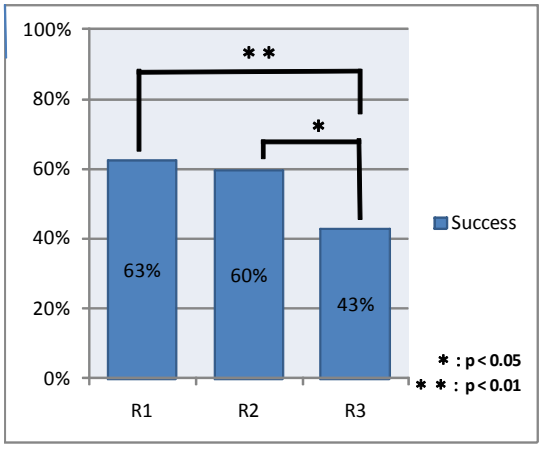


Figure 7. Turn-releasing success ratio

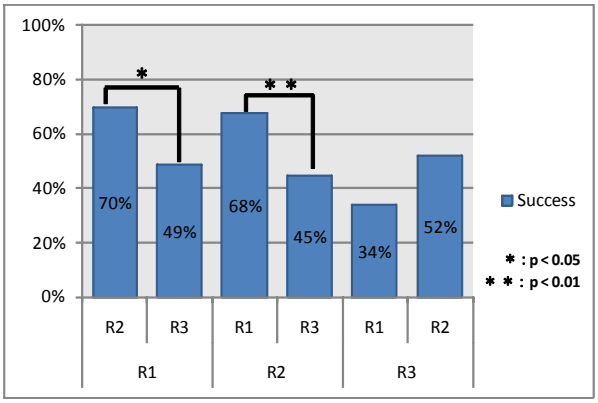
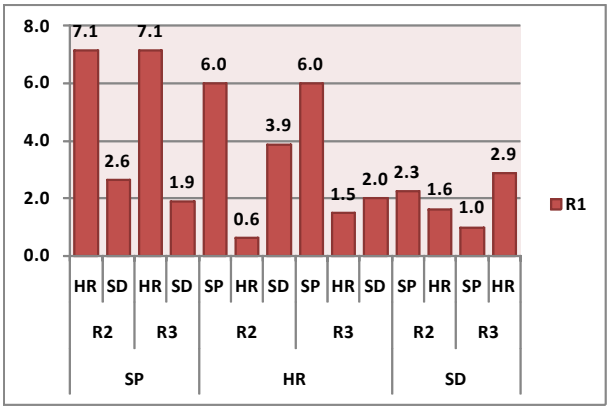


Figure 8. Turn-releasing success ratio WRT the addressee of the gaze signal

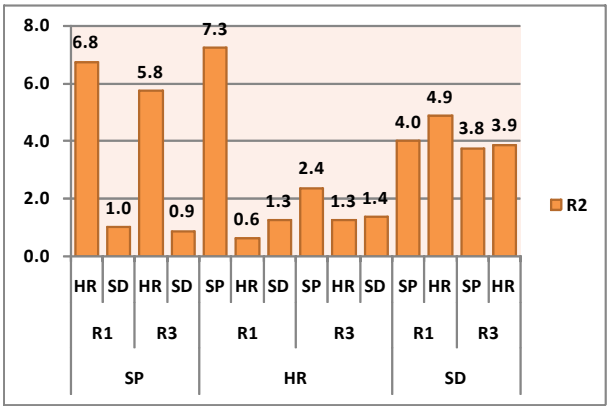
dominance in conversation. Figure 6 shows the average number of times for each participation role. R1 more frequently served as the speaker (SP) then the other two, and less frequently served as the side participant (SD). As the result of examining the difference of average by ANOVA, the average difference is statistically significant for all the participation roles (SP: $F(2,21)=6.338$ $p < .001$, Scheffe's post hoc test; R1 vs R3 $p < .05$, HR: $F(2,21)=4.481$ $p < .005$, post hoc test; R1 vs R3 $p < .05$, SD: $F(2,21)=7.302$ $p < .001$, post hoc test; R1 vs R3 $p < .05$) In the following sections, we investigate whether such difference in the distribution of participation roles is related to eye gaze.

Choosing the next speaker using gaze

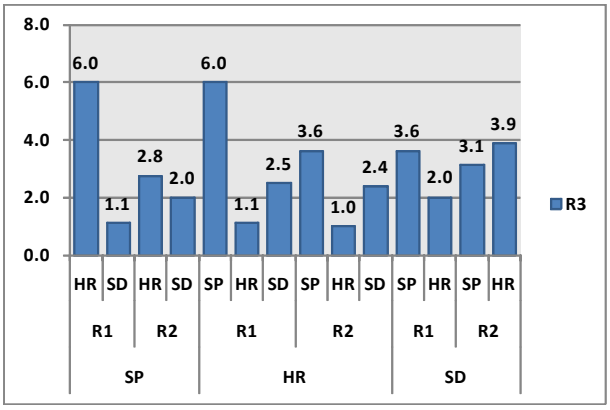
In previous studies [15, 16], it was found that the speaker chooses the next speaker using gaze. The next speaker, who receives the turn-releasing gaze signal by establishing a mutual gaze with the current speaker, looks away for a very short period of time, and then starts speaking. This is a typical gaze exchange for turn taking. We investigated whether this turn taking process is different depending on the dominance of a participant in a given conversation. We



(a)



(b)



(c)

Figure 9. Frequency of gaze

calculated the turn-releasing success ratio for each participant. For each utterance, we identified who was gazed at by the speaker for the last one second of the utterance. If the person who received the turn-releasing signal actually started speaking within 2 seconds, we counted this as the success of choosing the next speaker.

Figure 7 shows the turn-releasing success ratio with respect to the order of conversation dominance. The success ratio for rank 1 (R1) and rank 2 (R2) participants is over 60%. On the contrary, that for rank 3 (R3) participant is about 40%. As the result of statistical test for difference of ratio, the turn-releasing success ratio is different depending on the order of conversational dominance (χ^2=10.7785, p<0.0045). As the result of Post-hoc test using Ryan method, the difference between R1 and R3, and between R2 and R3 were statistically significant. These results suggest that R3 was less successful in choosing the next speaker using gaze.

Moreover, we analyzed whether the turn-releasing success ratio is different depending on who releases the turn to whom. As shown in Figure 8, the turn-releasing success ratios between R1 and R2 (R1 gives a turn to R2, and R2 gives a turn to R1) was about 70%. On the contrary, in releasing a turn to R3 or receiving a turn from R3, the success ratio became much lower than 30%. Therefore, these results suggest that the turn taking was quite successful between R1 and R2, but not for R3.

Getting attention from others

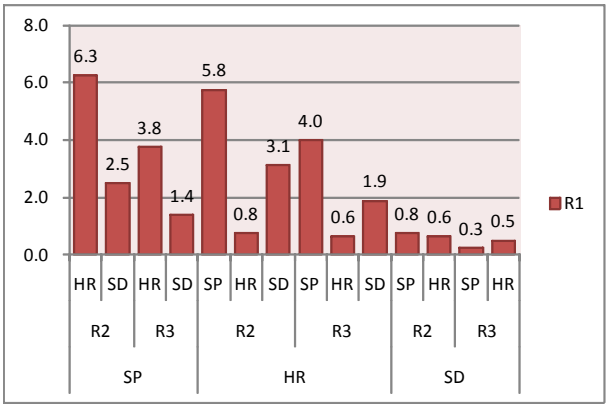
Even when the participants are not speaking, they are looking at someone. So, we analyzed participants gaze with respect to the participation role and the conversation dominance. Figure 9 (a), (b), (c) show the average number of eye gaze per conversation. When gaze shifts to other participants or elsewhere, one unit of gaze is completed. In the current analysis, we did not consider the duration of eye gaze in counting gaze behaviors. As shown in Figure 9 (a), when R1 was the speaker, the frequency of looking at R2 as the hearer and that of looking at R3 were the same (7.1). Likewise, in Figure 9 (b), when R2 was the speaker, R2 almost equally looked at R1 (6.8) and R3 (5.8). On the contrary, in Figure 9 (c), when R3 was the speaker, R3 more frequently looked at R1 (6.0) than R2 (2.8) as the hearer. This suggests that R3 more respected R1 than R2 as the hearer.

In analyzing the gaze behaviors as the hearer, when R1 was the hearer, s/he equally looked at R2 (6.0) and R3 (6.0) as the speaker. However, when R2 was the hearer, s/he more respected R1 (7.3) than R3 (2.4) as the speaker. Similarly, when R3 was the hearer, s/he more respected R1 (6.0) than R2 (3.6) as the speaker. We did not find clear trends for the distribution of gaze behaviors for side participants.

These results suggest that R1 got more attention from other participants as the speaker and as the hearer. On the other hand, while R1 equally looked at R2 and R3, gaze behaviors of R2 and R3 were not balanced (more looking at R1).

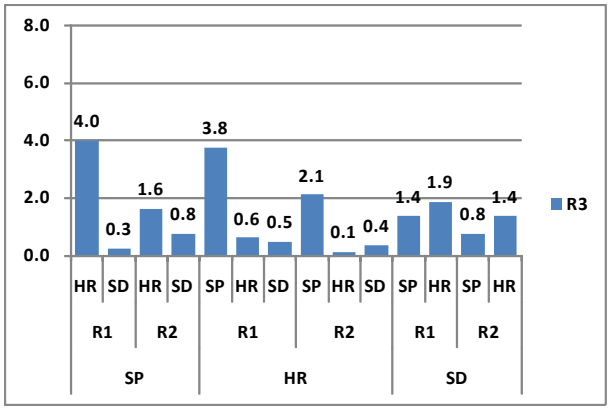
Mutual gaze

In addition to gaze, mutual gaze is also very important in investigating the interaction. Thus, we automatically identified mutual gaze as the gaze overlap between two



(a)

(b)



(c)

Figure 10. Frequency of mutual gaze

participants. Figure 10 (a), (b), (c) show the average number of mutual gaze for each participant per conversation. As shown in Figure 10 (a), when R1 was the speaker, mutual gaze was more frequently established with R2 as the hearer (6.3) than with R3 (3.8). In Figure 10 (b), when R2 was the speaker, mutual gaze was more frequently

established with R1 as the hearer (5.8) than with R3 (2.1). When R3 was the speaker, mutual gaze was more frequently established with R1 as the hearer (4.0) than with R2 (1.6). These results suggest that when R1 or R2 is the speaker, mutual gaze is more frequently established between them (not with R3). It is interesting that even though R1 and R2 equally look at other participants (shown in Figure 9), mutual gaze is more frequently exchanged between R1 and R2.

From the hearer's perspective, it is obvious that the hearer mainly looks at the speaker, but the frequency of mutual gaze is different depending on the conversation dominance. When R1 was the hearer, mutual gaze was more frequently established with R2 (5.8) than with R3 (4.0) as the speaker. When R2 was the hearer, mutual gaze was more frequently established with R1 (6.3) than with R3 (1.6) as the speaker. When R3 was the hearer, mutual gaze was more frequently established with R1 (3.8) than with R2 (2.1) as the speaker. Again, these results suggest that when R1 or R2 is the hearer, they spend less attention to R3 as the speaker even though R3 respects R1 as the speaker. More interestingly, when R1 was the hearer, the frequency of mutual gaze with R2 as the side participant (3.1) was close to that with R3 as the speaker (4.0). This suggests that R1 respects R2 even if R2 is the side participant.

In the analysis of side participant, since R1 rarely became the side participant, the frequencies were very low for all the cases. An interesting finding is that when R2 was the side participant, mutual gaze with R1 was more frequent than with R3 regardless of the participation roles.

These results suggest that both participation roles and conversation dominance affect the participants' gaze behaviors and mutual gaze between the participants. It was found that R1 got highest respect from other participants, and R2 was respected by R1. However, R3 did not get attention from other participants.

CONCLUSION AND FUTURE WORK

This paper presented our empirical study of multiparty conversations. By analyzing turn taking, gaze, and mutual gaze, we found that these nonverbal behaviors were affected by participation roles and conversation dominance. As the next step, we are aiming at estimating the participation framework based on gaze and mutual gaze behaviors, and then estimating the conversation domination.

We admit that there are many other factors that affect the dominance of conversation, such as personal relationship, and personality of each participant. Our research goal is to recognize what is happening in a given conversation, but not estimating static social relationship or personality. In our experiment, we control the relationship between the subjects by assigning a role to each subject.

Although this study focused on human-human interaction, in future work, we will analyze gaze and mutual gaze in human-agent multiparty conversation, and compare to the results found in this study.

ACKNOWLEDGMENTS

This work is partially funded by JSPS under a Grant-in-Aid for Scientific Research in Priority Areas "i-explosion" (21013042), and MEXT Grant-in-Aid for Building Strategic Research Infrastructures.

REFERENCES

1. Goffman, E., *Forms of Talk*. 1981, Philadelphia, PA: University of Pennsylvania Press.
2. Clark, H.H., *Using Language*. 1996, Cambridge: Cambridge University Press.
3. Argyle, M. and M. Cook, *Gaze and Mutual Gaze*. 1976, Cambridge: Cambridge University Press.
4. Vertegaal, R. *The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration*. in *CHI 1999*. 1999.
5. Dovidio, J.F. and S.L. Ellyson, *Patterns of visual dominance behavior in humans*, in *Power, dominance, and nonverbal behavior*, S.L. Ellyson and J.F. Dovidio, Editors. 1985, New York: Springer-Verlag. p. 129-149.
6. Knapp, M.L. and J.A. Hall, *Nonverbl Communication in Human Interaction*. 2010: Wadsworth.
7. Burgoon, J.K. and N.E. Dunbar, *Nonverbal expressions of dominance and power in human relationships*, in *The Sage handbook of nonverbal communication*, V. Manusov and M.L. Patterson, Editors. 2006, Sage publications. p. 279-297.
8. Eichner, T., et al. *Attentive Presentation Agents*. in *The 7th International Conference on Intelligent Virtual Agents (IVA)*. 2007.
9. Traum, D., et al. *Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents*. in *the 8th International Conference on Intelligent Virtual Agents (IVA08)*. 2008.
10. Nagao, K. and A. Takeuchi. *Social Interaction: Multimodal Conversation with Social Agents*. in *Twelfth National Conference on Artificial Intelligence*. 1994. Menlo Park, CA: AAAI Press.
11. Bohus, D. and E. Horvitz. *Open-World Dialog: Challenges, Directions, and Prototype*. in *IJCAI2009 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. 2009.
12. Bohus, D. and E. Horvitz. *Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings*. in *SIGdial'09*. 2009. London, UK.
13. Huang, H.-H., et al. *How multiple concurrent users react to a quiz agent attentive to the dynamics of their game participation*. in *AAMAS*. 2010.
14. Kipp, M. *Anvil - A Generic Annotation Tool for Multimodal Dialogue*. in *the 7th European Conference on Speech Communication and Technology*. 2001.

15. Duncan, S., Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology, 1972. 23(2): p. 283-292.

16. Kendon, A., Some Functions of Gaze Direction in Social Interaction. Acta Psychologica, 1967. 26: p. 22-63.

Influence of user's mental model on natural gaze behavior during human-computer interaction

Thomas Bader[1,2]
thomas.bader@kit.edu

Jürgen Beyerer[1,2]
juergen.beyerer@iosb.fraunhofer.de

[1]Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT),
Germany

[2]Fraunhofer IOSB
Intitut für Optronik, Systemtechnik und Bildauswertung
Germany

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 relations between natural gaze behavior and underlying cognitive processes still is unexplored to a large extend. In this paper we identify and characterize major factors influencing natural gaze behavior during human-computer interaction with a focus on the role of user's mental model about the interactive system in that context. In a user study 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.

Author Keywords

gaze based interaction, natural gaze behavior, multimodal interfaces

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION

In general there are two ways to incorporate eye gaze as an input modality into multimodal human-computer interfaces. The first way is to force the user to consciously look at certain locations in order to trigger actions. One example for such approaches is eye typing, which has been studied for decades [9]. 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[5].

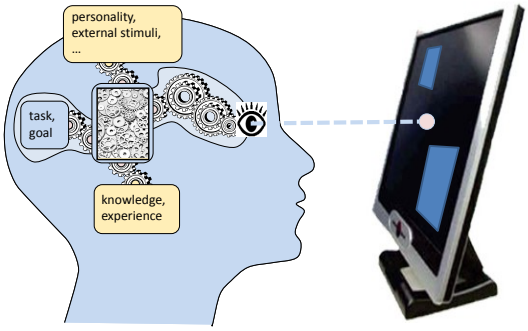
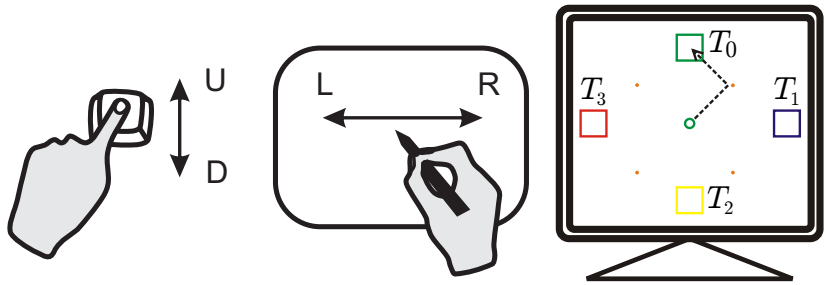


Figure 1. Dependency of natural gaze behavior

The second way to use eye gaze as input modality is to interpret natural gaze behavior during human-computer interaction, while using another modality as primary input modality. Promising examples for such interaction techniques are presented in [4] and [12]. In both approaches *natural* gaze behavior is analyzed and the user is not forced to diverge from that natural behavior for interaction purposes. *iDict* [4] 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"[12] 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. Addition-

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction. February 13, 2011, Palo Alto, California, USA.



(a) Key (b) Pen (c) Task

Figure 2. Input devices and task

ally, 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 Figure 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 [6, 8]).

Numerous studies of natural gaze behavior and hand-eye coordination during manipulative activities in natural environments like block-copying [10], basic object manipulation [6], driving [7] and playing cricket [8] revealed gaze shifts and fixations to be commonly proactive (eye-movements occured previous to movements of the manipulated object or the manipulator). In addition, a detailed study on hand-eye coordination during an object manipulation task [6] 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 in general that are critical for the completion of the task, like in [8] 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 [6] or previous experience of a person [8]).

Gaze behavior was also studied in various tasks related to HCI. In [11] 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 [1], 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. [3, 2] investigated differences in gaze behavior during action execution and observation. They distinguished three different gaze behaviors, namely proactive, reactive and tracking gaze behavior [3].

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.

In this paper 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.

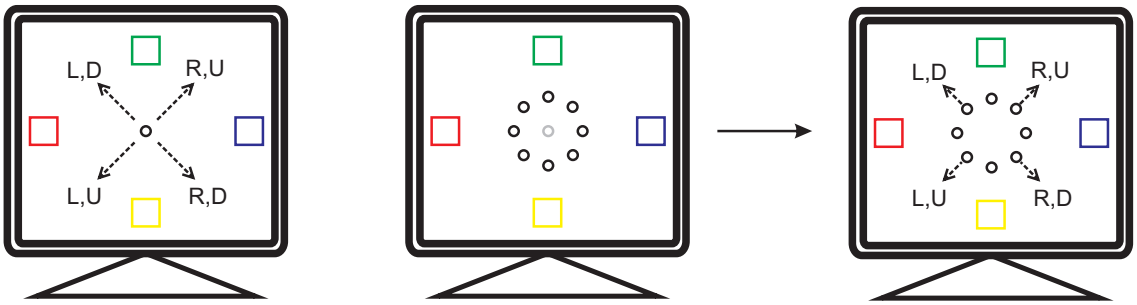
USER STUDY

Task and Apparatus

The task to be solved by participants is designed based upon a basic object manipulation task as it is common in many GUIs. The visual representation of an object has to be moved from one location to another on a display. However, in order to being 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 (Figure 2(a)) and a pen tablet, while only horizontal movements of the pen on the tablet are interpreted by the system (Figure 2(b)). The task is illustrated in Figure 2(c). A colored point which initially is displayed at the center of the display is to be moved to one of the four squares $T_0, ..., T_3$ with the same color. Note that the labels $T_0, ..., T_3$ shown in Figure 2(c) were not displayed to the user during the experiment and only serve as reference for the respective target areas within this paper.

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 Figure 3(a). For example, a horizontal movement of the pen to the right



(a) Technique1: movement direction of object given different inputs

(b) Technique2: expanded object state (left) and movement direction of individual representations given different inputs (right)

Figure 3. Mapping of input to system actions for different interaction techniques

(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 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 Figure 3(b) 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 can not 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 are removed. 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 (Figure 3(b) 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 Figure 2(c), 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 chose 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. This would, e.g., allow for robust estimation of users' intention from gaze data.

The size of the display is 33,7 x 27 cm with a resolution of 1280x1024 pixels. Eye-gaze of the users was captured during task execution by a Tobii 1750 tracking device.

Participants

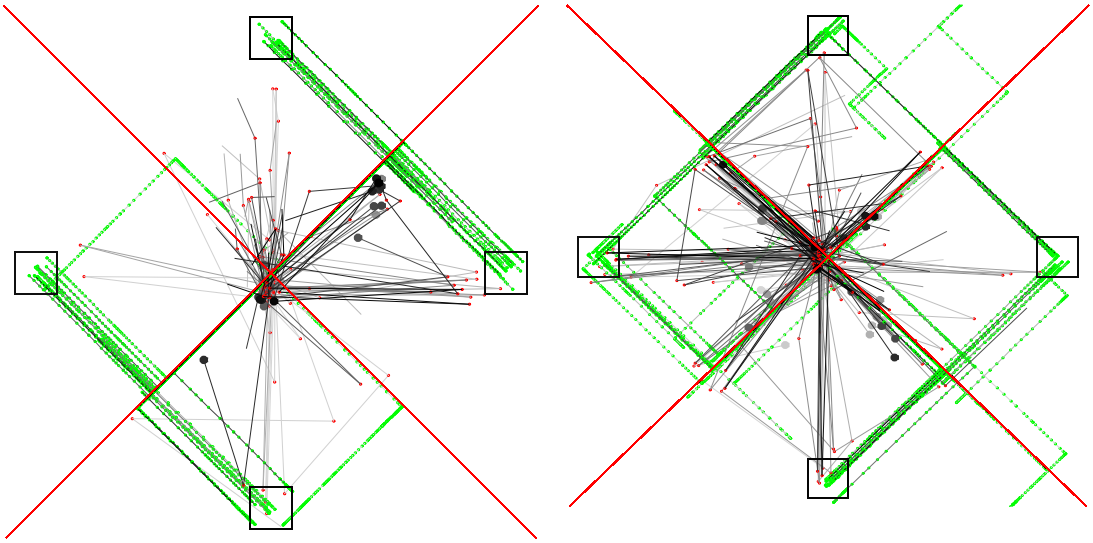
Since we want to investigate effects of mental model building on natural gaze behavior we chose a between-subjects design to avoid any prior knowledge of participants about the task or interaction techniques. We had two groups with 10 participants each. Participants were between 21 and 32 years old and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

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, while the order of tasks was chosen randomly and was the same for all participants. Except the order of tasks there was no difference between phase $A1$ and $A2$.

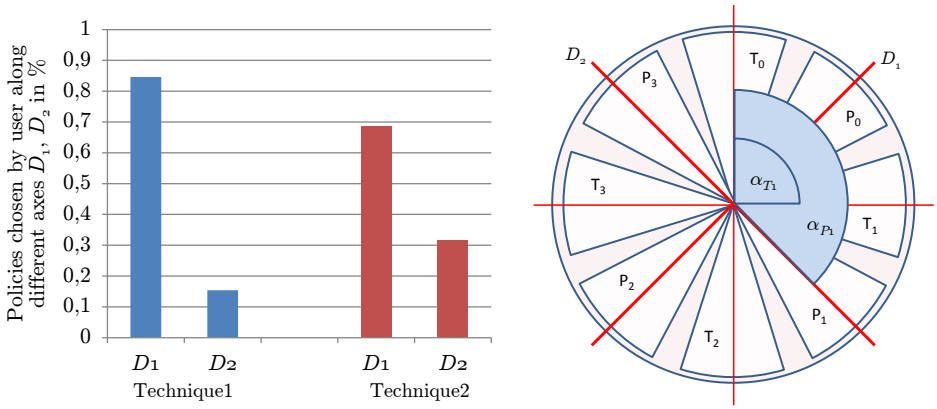
Between the two phases users were asked to fill in a questionnaire in order to capture their mental model. However, in this paper we focus on analysis of objective data only and analysis of subjective data obtained from the questionnaire will be reported in future papers.

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.



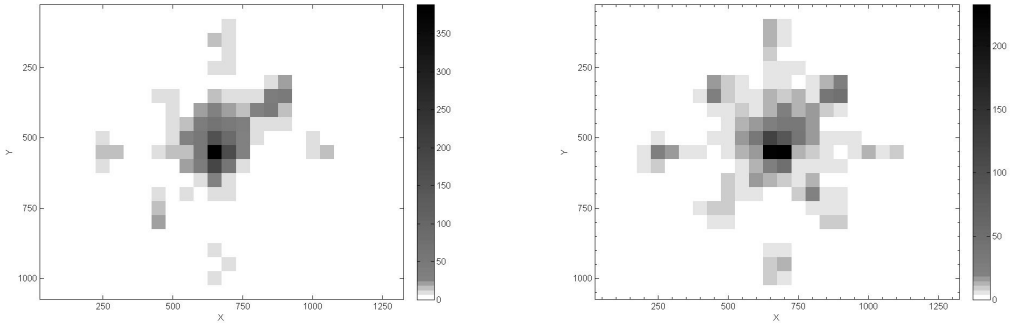
(a) Technique1 (b) Technique2

Figure 4. Data captured for different interaction techniques during phase $A1$ from one user for each technique



(a) Distribution of policies chosen by users (b) Definition of variables

Figure 5. Task solution strategy of users and definition of variables



(a) Pre-object fixations for Technique1 (b) Pre-object fixations for Technique2

Figure 6. Distribution of pre-object fixations

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 Figure 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.

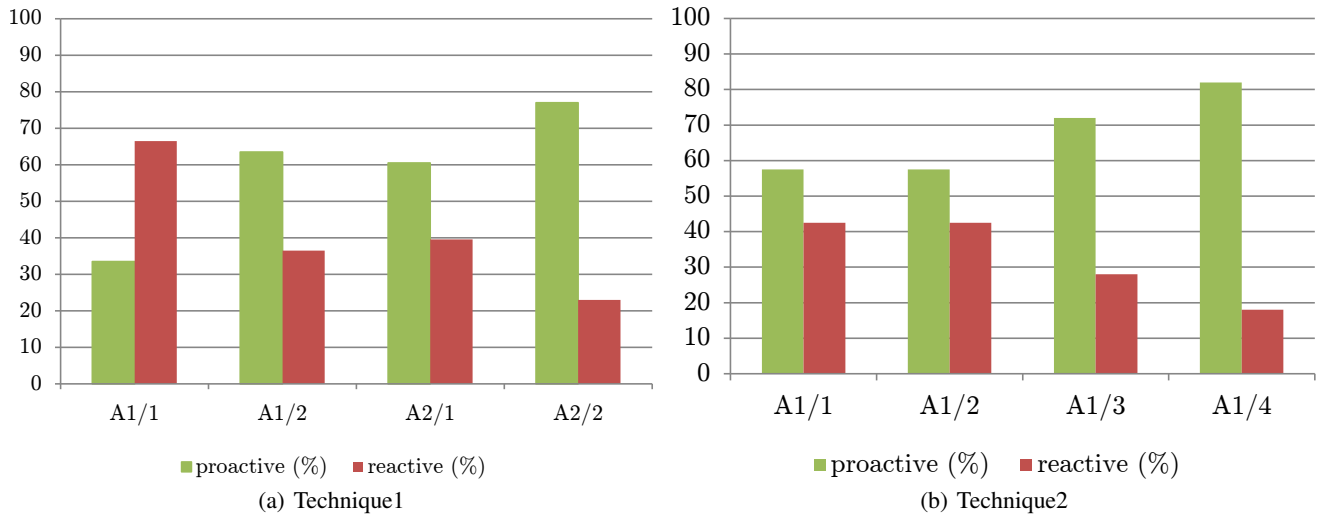
For the first interaction technique two things can be easily seen from Figure 4. First, the preferred policy for solving the task seems to be first moving the object along the diagonal line reaching from the lower left to the upper right (D_1, see Figure 5(b) for definition). This corresponds to an input 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.

Both observations in average can be confirmed for all participants. In Figure 5(a) 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 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 Figure 6(b) 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* α 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. The design of the task 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 α. We further denote the different target areas as $T_0, ..., T_3$ in clockwise direction, starting from the top. The different policies users can chose to solve a task are denoted by $P_0, ..., P_3$ in clockwise direction according to the first primary movement direction starting from the top. The definition of T_i and P_i are also illustrated in Figure 5(b). 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



(a) Technique1 (b) Technique2

Figure 7. Development of ratio between proactive and reactive fixations with increasing knowledge about the system

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 categorized as follows:

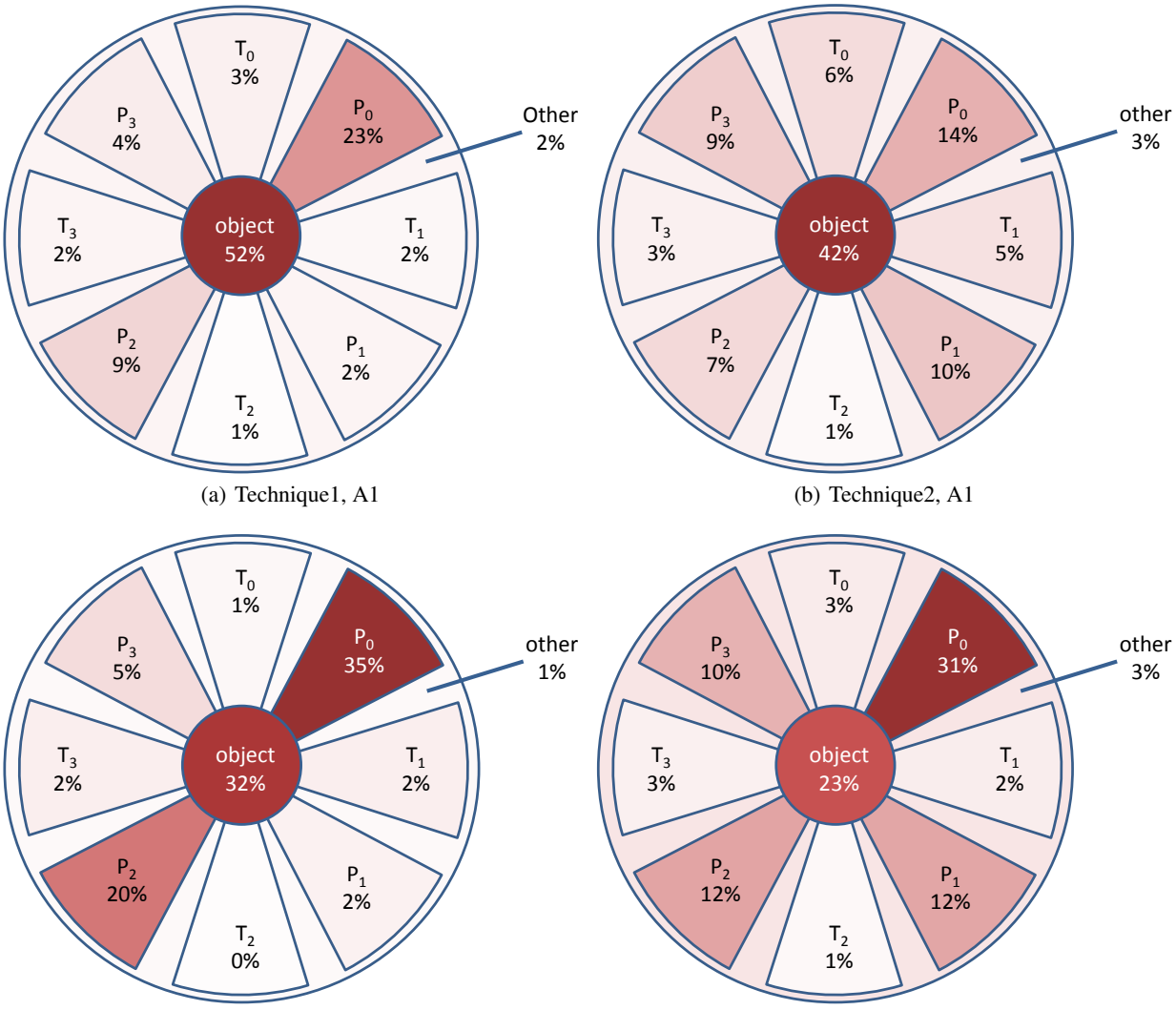
$$A = \begin{cases} T_i \text{ if } |\alpha_{T_i} - \alpha| < \alpha_{max} \\ P_i \text{ if } |\alpha_{P_i} - \alpha| < \alpha_{max} \\ other \end{cases},$$

where α_{T_i} and α_{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 Figure 5(b)).

The thresholds $r_p = 100$ and $\alpha_{max} = 20°$are 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 Figure 7. In average the ratio for Technique1 is 58.625/41.375 (proactive/reactive) and 67.25/32.75 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 8 shows the distribution of A over all possible targets $T_0, ..., T_3, P_0, ..., P_3$ for all last pre-object fixations. The different areas represent the categories as defined above by r_p and α_{max} and are colored according to the occurrence of fixations within the corresponding area on the screen.



(a) Technique1, A1

(b) Technique2, A1

(c) Technique1, A2

(d) Technique2, A2

Figure 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 $P_0, ..., P_3$ 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 Figure 5(a), where similar differences in policies chosen by the users for solving the task are depicted.

DISCUSSION

The results in the previous section show that both independent variables we used in our 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 $P_1, ..., P_3$ 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 extend 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. 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.

CONCLUSION

By the experiment described in this paper 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 in this paper show the variety of information contained in natural gaze behavior. By analyzing natural gaze behavior during human-computer interaction in-

formation like user's intention or experience can be inferred which can be used for designing proactive or adaptive intelligent user interfaces.

In future work we plan to further validate the identified dependencies with more complex tasks and to design and evaluate gaze-based multimodal interaction techniques with a focus on multimodal combination of gesture and gaze.

REFERENCES

1. T. Bader, M. Vogelgesang, and E. Klaus. Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI-MLMI)*, pages 199–206. ACM, Nov. 2009.
2. J. R. Flanagan and R. S. Johansson. Action plans used in action observation. *Nature*, 424(6950):769–771, Aug 2003.
3. B. Gesierich, A. Bruzzo, G. Ottoboni, and L. Finos. Human gaze behaviour during action execution and observation. *Acta Psychologica*, 128(2):324 – 330, 2008.
4. A. Hyrskykari, P. Majaranta, and K. Räihä. Proactive response to eye movements. In M. Rauterberg, editor, *Human Computer Interaction INTERACT 2003*, pages 129–136. IOS Press, September 2003.
5. R. Jacob and K. Karn. *Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises*, pages 573–605. Elsevier Science, 2003.
6. R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan. Eye-hand coordination in object manipulation. *The Journal of Neuroscience*, 21(17):6917–6932, 2001.
7. M. Land and D. Lee. Where we look when we steer. *Nature*, 369:742–744, 1994.
8. M. F. Land and P. McLeod. From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, 3:1340–1345, 2000.
9. P. Majaranta and K.-J. Räihä. Twenty years of eye typing: systems and design issues. In *ETRA '02: Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 15–22, New York, NY, USA, 2002. ACM.
10. J. Pelz, M. Hayhoe, and R. Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139:266–277, 2001. 10.1007/s002210100745.
11. B. A. Smith, J. Ho, W. Ark, and S. Zhai. Hand eye coordination patterns in target selection. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, ETRA '00, pages 117–122, New York, NY, USA, 2000. ACM.
12. S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (magic) pointing. In *In CHI99*, pages 246–253. ACM Press, 1999.

Awareness of Partner's Eye Gaze in Situated Referential Grounding: An Empirical Study

Changsong Liu
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
cliu@msu.edu

Dianna L. Kay
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
dnnkay855@gmail.com

Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

ABSTRACT

In situated dialogue, although artificial agents and their human partners are co-present in a shared environment, their representations of the environment are significantly different. When a shared basis is missing, 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 designed an experiment to simulate different representations of the shared environment. Our studies have shown that, when one partner pays attention to the other partner's naturally occurred eye gaze during interaction, referential grounding becomes more efficient. However this improvement is more significant under the situation where partners have matched representations of the shared environment compared to the situation with mismatched representations. This paper describes our experimental findings and discusses their potential implications.

Author Keywords

Situated Dialogue, Eye Gaze, Referential Grounding

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Miscellaneous

General Terms

Experimentation, Human Factors

INTRODUCTION

With recent advances in AI techniques, there is an increasing demand for artificial agents (e.g. robots) that can collaborate with humans in a shared environment. Examples of such applications include space exploration [7], military training [12], and autism therapy [5]. To develop this type collaborative agents, a crucial issue is to address the mismatched abilities between artificial agents and their human partners.

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

Although artificial agents and their human partners are co-present in a shared environment, their perception, representation and knowledge of the shared environment could be significantly different. When a shared basis of the environment is missing, communication between partners becomes more challenging [4]. 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 – human eye gaze during speech communication.

Eye gaze serves many functions in mediating interaction [1, 4] and managing turn taking [17] and grounding [16]. Previous psycholinguistic findings have shown that eye gaze is tightly linked with language production and comprehension [10, 21, 15, 8]. Eye gaze has also been shown efficient for providing early disambiguating cues in referential communication [9], for intention recognition during object manipulation [2], and for attention prediction [6]. Specifically, related to the referential process, recent work has incorporated eye gaze in resolving exophoric referring expressions [18, 19].

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

1. *How difficult is it for partners with mismatched representations of the shared environment to collaborate?* When a shared basis of the environment is missing, partners may not be able to communicate as they normally do. We are interested in how the mismatched representations may impact collaboration, conversation, and automated language processing.

2. *To what extent does the collaboration benefit from the awareness of partner's eye gaze? Is such awareness more helpful for partners with mismatched representations?* Our hypothesis is that partners with mismatched representations could benefit more from gaze information. This is because on one hand verbal communication could be more difficult, and on the other hand gaze may allow many joint actions to be done non-verbally [3].

To validate this hypothesis and address the above questions, we designed an experiment where a director and a matcher

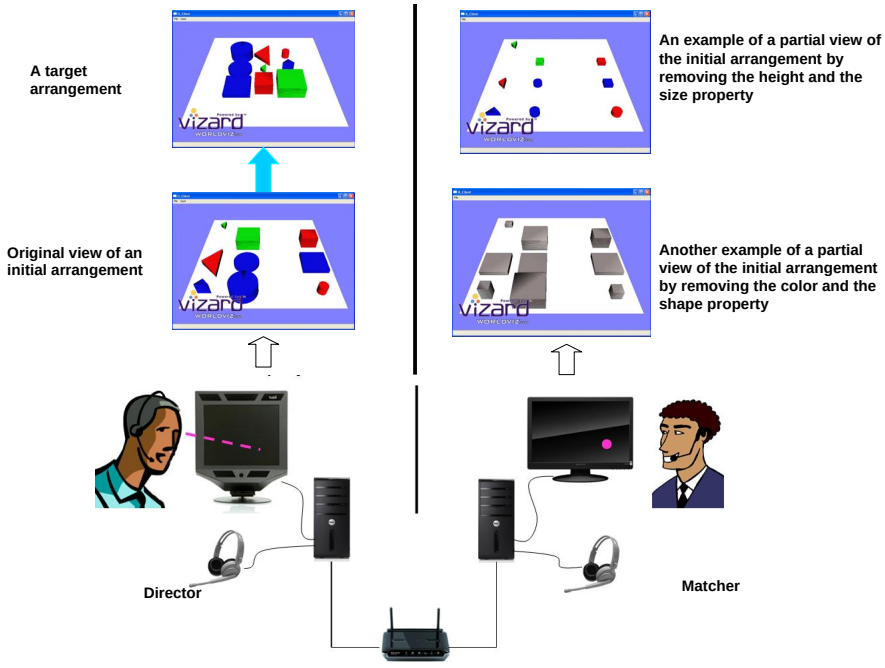


Figure 1. Architecture of our collaborative system. Two partners in the same room are separated by a divider. The director is seated in front of a display-mounted Tobii 1750 eye tracker (Tobii Technology) and the matcher in front of a regular computer. Two computers are connected and synchronized via an Ethernet hub. The director's eye gaze positions are captured and can be displayed as gaze cursors (a 32 by 32 pixels pink dot) superimposed over the matcher's display based on experimental conditions. A bi-directional microphone-speaker system was used as the speech channel for two partners to verbally communicate.

collaboratively solve a block 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 different representations of the shared environment between the director and the matcher (i.e., matched representations or mismatched representations). In addition, during the interaction, we keep track of the director's eye gaze and alter the conditions based on whether the matcher is aware of the director's eye gaze by controlling whether the director's eye gaze will be made available on the matcher's screen. These settings allow us to investigate the role of eye gaze in different experimental conditions.

Our results indicate that human partners can overcome the mismatched representations by switching their communicative strategies. The matcher's awareness of the director's eye gaze during interaction improves referential grounding. However, such an improvement is more significant under the matched representations compared to mismatched representations. This finding is somewhat surprising given our original hypothesis. This is possibly due to the more complex gaze pattern that interacts with the spatial information occurred in utterances, which opens up new interesting questions for future research.

METHOD

Apparatus

An architecture of our collaborative system is shown in Figure 1. Two partners (a director and a matcher) collaborate on a block arranging task. The director instructs the matcher to arrange a set of blocks according to a given configuration (i.e., a target configuration) which is only available to the director. They both face the same virtual environment, however, it can be displayed differently on their respective screens to simulate different internal representations of the environment. The director's eye gaze positions are captured by a TOBII display-mounted eye tracker. Depending on experimental conditions, the director's naturally occurred gaze will be made available to the matcher (shown as gaze cursors on the matcher's screen) real time during interaction.

Experiment Design

A pair of participants collaborate to arrange a set of blocks in a given order. In each task, there are seven blocks with randomly assigned color (red, green or blue), shape (triangle, circle or pentagon), size (small, medium or large) and height (short, medium or tall). At the beginning of the task, seven blocks are randomly placed on a plane (Figure 2a). The director is given a target arrangement of the blocks (Figure 2b). The director can view the final arrangement at any time by pressing the 'space' key. The director can not move the blocks by himself. He has to instruct the matcher to pick up one block and move it to the desired location, and then continue with the next one till all the blocks are in the right positions. They also have to follow a given order when they are moving the blocks one by one. The block that should be moved at the current step is indicated by an arrow pointing to it, which can only be seen by the director.

The director and the matcher both face the same environ-

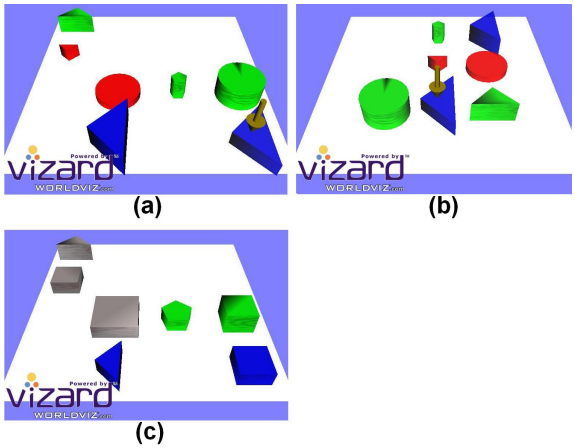


Figure 2. Example of different views in our experiment. (a) shows a random initial arrangement of the blocks on the director's screen at the beginning of the task. The block that should be moved at the current step is indicated by an arrow pointing to it. (b) shows the target arrangement of the blocks. (c) shows the corresponding partial view displayed on the matcher's screen to simulate mismatched representations between the director and the matcher.

ment, but may have different representations of this environment. The blocks can be displayed identically or differently on their respective screens, depending on the experimental conditions based on the following dimension:

- Whether the director and the matcher has matched (*view+*) or mismatched (*view–*) representations of the shared environment. Under the mismatched condition, the display on the matcher's screen is different from the director's in the following way: each attribute (color, shape, size or height) of a block has a 50% chance to be set to an 'unknown' default status (i.e., color is set to gray, shape is set to square, size is set to medium and height is set to medium). Figure 2(c) shows a mismatched-view display of the same environment as in Figure 2(a).
- Whether the matcher is aware of the director's gaze direction during interaction. This is controlled by whether or not displaying the director's naturally occurred eye gaze as cursors on the matcher's display. This results in two levels: the matcher is aware of the director's eye gaze (*gaze+*) and the matcher is not aware of the eye gaze (*gaze-*)

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

- (*view+,gaze+*): matched-view with gaze awareness
- (*view–,gaze+*): mismatched-view with gaze awareness
- (*view+,gaze–*): matched-view without gaze awareness
- (*view–,gaze–*): mismatched-view without gaze awareness

Participants and procedure

Sixteen (eight pairs) undergraduate/graduate students from Michigan State University were recruited to participate in our studies. 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 first had two practice trials. The first practice trial was the (*view+,gaze+*) condition and the second was (*view–,gaze–*). After the first two practice trials, all our participants had no problem understanding the task and familiarizing themselves with gaze cursors. They then proceeded with four actual trials, each of which was based on one of the four conditions, in a completely random order. The experiment lasted approximately 40 minutes.

RESULTS AND DISCUSSION

During the experiment, each trial was logged by keeping track of both the director's and the matcher's displays and speech communication. Two kinds of information were extracted from the logged data to facilitate our investigation: the total time spent to finish each trial and the total number of utterances issued by both partners in each trial. Table 1 shows the time (seconds) spent to finish each trial. Table 2 shows the number of utterances in each trial.

The Role of Mismatched Representations

We compare the means of time and number of utterances between different conditions. Surprisingly, we did not find significant difference between (*view–,gaze–*) and (*view+,gaze–*) (for time, $t = -0.25, p < 0.594$; for number of utterances, $t = -0.46, p < 0.671$). The result implies that collaboration based on mismatched representations of the environment may not be more difficult. A further investigation of our data indicate that partners employed different communicative strategies in the (*view–,gaze–*) condition compared to the (*view+,gaze–*) condition. There were basically two kinds of strategies for collaboratively referring to the intended block: one is to describe its object-based properties (color, shape, size and height), and the other is to describe its spatial information locally (with respect to another block(s) close by) or globally (with respect to the environment). Sometimes the two strategies can also be used together. Here are examples of the two strategies from our data:

> *Object-based properties*: the red pentagon / the big blue one / the tallest one
> *Spatial information*: the object underneath the one we just moved / the middle one of the three objects on the top / the object in the right bottom corner
> *Combined*: the green one to the left of that one / the triangular-looking one on the right side of this environment

It is natural to use object-based properties with matched views, whereas using spatial information is a better strategy for mismatched views. In our experiment, participants had no problem switching to and relying on the spatial information when they detected they might have mismatched representations. Sometimes they even tried to explicitly query or request their partners about the condition or strategy at the beginning of a trial, e.g., "*Can you see color this time?*", "*I only have gray squares, so you would better always describe the spa-*

Condition	Pair of participants								$\bar{y}_{i.}$
	1	2	3	4	5	6	7	8	
view+,gaze–	126.2	158.6	129.8	208.2	156.0	70.1	188.7	160.4	149.8
view–,gaze–	71.4	107.8	217.0	150.2	169.2	103.2	257.2	77.0	144.1
view+,gaze+	72.8	134.8	155.3	155.8	97.1	94.0	120.3	53.6	111.5
view–,gaze+	86.7	129.9	135.7	210.2	84.9	83.6	146.8	78.9	119.6
$\bar{y}_{.j}$	89.3	132.8	159.5	181.1	126.8	87.7	178.3	92.5	$\bar{y}_{..} = 131.0$

Table 1. Time (seconds) spent to finish each trial. The last row of the table is the average time of each pair of participants ($\bar{y}_{.j}$). The last column of the table is the average time of each experimental condition ($\bar{y}_{i.}$).

Condition	Pair of participants								$\bar{y}_{i.}$
	1	2	3	4	5	6	7	8	
view+,gaze–	49	45	45	99	121	23	52	65	62
view–,gaze–	24	44	90	53	103	30	84	29	57
view+,gaze+	17	41	51	43	72	33	27	19	38
view–,gaze+	40	33	48	98	66	24	38	30	47
$\bar{y}_{.j}$	33	41	59	73	91	28	50	36	$\bar{y}_{..} = 51$

Table 2. Number of utterances in each trial. The last row of the table is the average number of utterances of each pair of participants ($\bar{y}_{.j}$). The last column of the table is the average number of utterances of each experimental condition ($\bar{y}_{i.}$).

tial location". The spatial information based strategy, although shown to be less preferable when object-based properties were available [13], appears equally efficient compared to the object-based strategy in our experiment. Actually, it provides an easier and more reliable way for collaborative referring when the shared basis was missing as in the mismatched-view case. Therefore, all our participants spontaneously relied on spatial information and accomplished the task with little trouble under the mismatched-view condition.

The Role of Gaze Awareness

The other aspect we are interested in is the role of gaze awareness in this collaborative task. As we expected, when the two partners' views were matched, allowing the matcher to see the director's gaze positions significantly reduced the time and the number of utterances needed to accomplish the task. This is demonstrated by the comparison between (*view+,gaze+*) and (*view+,gaze–*). Under the (*view+,gaze+*) condition, participants spent 39.3 seconds shorter time ($t = 2.43, p < .05$) and issued 24.5 fewer utterances ($t = 2.69, p < .05$) compared to the (*view+,gaze–*) condition. This result is in accordance with previous findings (e.g. [11, 1]) that shared gaze facilitates the communication of task-relevant spatial information.

Is gaze more helpful when two partners' representations of the environment are mismatched? Our results have shown some interesting observations. Although the (*view–,gaze+*) condition on average took 24.5 seconds shorter time and 10 fewer utterances than the (*view–,gaze–*), the differences are only marginally significant (for interaction time, $t = 1.14, p < 0.146$; for number of utterances, $t = 0.9, p < 0.199$). Why does gaze appear to be less helpful under the mismatched-view condition compared to the matched-view condition? To answer this question, more in-depth analysis has to be done to investigate how the gaze patterns are different under two conditions and how the difference may affect the partner's acceptance of gaze information. This is left for our future research. Here we only present one possible explanation based on our intuition.

We hypothesize that gaze patterns under the mismatched condition can be different from the patterns under matched condition. In the matched-view case, since the speaker's strategy is mainly based on describing the referent's own properties, his/her gaze should mostly fixed on the object that is being described at the moment. In other words, the object that draws most of the speaker's attention is very likely the current referent, and thus allows the listener to directly use this cue to infer the intended referent. However, such a pattern may be weakened in the mismatched-view case as the speaker switches to the strategy based on spatial information. Using spatial language often involves complex mental computations [14], which can possibly interact with gaze patterns. For example, instead of steadily fixating on the object that is being referred to, the speaker may need to look back-and-forth between two objects while he is selecting a proper spatial term to describe the relation. Also, the gaze could be circling among a group of objects if the speaker intends to use a group-based description (e.g., "*the middle object*", "*the third one from left to right*"). The gaze may not fixate on any object but rather scan through the environment if the speaker intends to describe a global direction (e.g., "*the northwest one*", "*the one that is to the right bottom corner*"). All these possible gaze behaviors may be closely related to the internal computations of spatial language, and thus may not precisely indicate the intended referent.

CONCLUSION

This paper investigate the role of eye gaze in the referential process in situated dialogue. Our findings indicate that, when the partners' representations of the shared environment are matched, eye gaze and awareness of eye gaze facilitate the communication and significantly reduces the time and verbal communication needed for grounding references. But when the partners have mismatched representations of

the shared environment, the effect of awareness of eye gaze appears less significant. This is possibly due to more complex patterns of gaze behaviors in production of utterances involving complex spatial information. However, more in-depth investigation has to be done to have a clear understanding.

Our experiments further indicate that, participants spontaneously employed communication strategies based on spatial information to describe objects of interest when the shared representation was missing. By using spatial information, participants grounded references as efficiently as applying strategies based on object properties. This finding again (see [20] for another example in human robot interaction) has revealed the importance of spatial language, especially when the artificial agents and their human partners have mismatched perceptions and representations of the shared environment. Therefore, spatial language understanding and spatial information based dialogue management serve as key components to enabling collaborative and situatd interaction between humans and artificial agents.

This paper only describes some initial results of our investigation. To have a more clear understanding of the role of eye gaze in this collaborative referential process, we also need to capture the gaze behaviors from the matcher and examine how that may affect the director's behaviors and thus the overall discourse of interaction. In addition, the interactions between the gaze from both partners will be even more interesting. Our future work will explore these directions.

ACKNOWLEDGMENT

This work was supported by CNS-0957039 and IIS-1050004 from the National Science Foundation.

REFERENCES

1. M. Argyle, M. Cook, and M. Argyle. *Gaze and mutual gaze*. Cambridge University Press Cambridge, 1976.
2. T. Bader, M. Vogelgesang, and E. Klaus. Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 199–206. ACM, 2009.
3. S. Brennan, X. Chen, C. Dickinson, M. Neider, and G. Zelinsky. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477, 2008.
4. H. H. Clark. *Using language*. Cambridge University Press, 1996.
5. K. Dautenhahn and I. Werry. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, 12(1):1–35, 2004.
6. R. Fang, J. Chai, and F. Ferreira. Between linguistic attention and gaze fixations in multimodal conversational interfaces. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 143–150. ACM, 2009.
7. T. Fong and I. Nourbakhsh. Interaction challenges in human-robot space exploration. *interactions*, 12(2):42–45, 2005.
8. Z. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11(4):274, 2000.
9. J. Hanna and S. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615, 2007.
10. M. Just and P. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480, 1975.
11. A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1):22, 1967.
12. P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, and D. Piepol. Building interactive virtual humans for training environments. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, volume 2007. NTSA, 2007.
13. S. Kriz, J. Trafton, and J. McCurry. The Role of Spatial Information in Referential Communication: Speaker and Addressee Preferences for Disambiguating Objects. In *D. S. McNamara & J. G. Trafton (Eds.), Proceedings of the 29th Annual Cognitive Science Society Austin, TX: Cognitive Science Society*, 2007.
14. C. Liu, J. Walker, and J. Chai. Ambiguities in Spatial Language Understanding in Situated Human Robot Dialogue. In *2010 Fall Symposium on Dialogue with Robots*, 2010.
15. A. Meyer, A. Sleiderink, and W. Levelt. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2):B25–B33, 1998.
16. Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 553–561. Association for Computational Linguistics, 2003.
17. D. Novick, B. Hansen, and K. Ward. Coordinating turn-taking with gaze. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1888–1891. IEEE, 2002.
18. Z. Prasov and J. Chai. 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*, pages 20–29. ACM, 2008.
19. Z. Prasov and J. Y. Chai. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 471–481, October 2010.

20. M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock. Spatial language for human-robot dialogs. *IEEE transactions on systems, man and cybernetics*, 34:154–167, 2004.

21. M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632, 1995.

Combining Multiple Types of Eye-gaze Information to Predict User's Conversational Engagement

Ryo Ishii*, †, Yuta Shinohara††, Yukiko, I. Nakano††, Toyoaki Nishida*

* Dept. of Intelligence Science and Technology
Kyoto University
Kyoto-Shi 606-8501, Kyoto, Japan
{ishii, nishida}@i.kyoto-u.ac.jp

† NTT Cyber Space Laboratories
NTT Corporation
Yokosuka-Shi 239-0847, Kanagawa, Japan
ishii.ryo@lab.ntt.co.jp

†† Dept. of Computer and Information Science
Seikei University
Musashino-shi 180-8633, Tokyo, Japan
y.nakano@st.seikei.ac.jp

ABSTRACT

In face-to-face conversations, speakers are continuously checking whether the listener is engaged in the conversation by monitoring the partner's eye-gaze behaviors. The goal of this study is to build an intelligent conversational agent that can recognize user's engagement from multiple types of eye-gaze information. In our previous work, we developed a method of estimating the user's conversational engagement using user's gaze transition patterns. However, it was not accurate enough. In this study, we added new variables in our gaze analysis: occurrence of mutual gaze, gaze duration, and distance of eye movement. Then, based on the results of the analysis, we used the variables as estimation parameters, and set up prediction models which consist of different combinations of parameters. To test which parameters are effective, the performance of these models were compared. As the result, it was revealed that a model using the gaze transition patterns, occurrence of mutual gaze, gaze duration, distance of eye movement, and pupil size as prediction parameters performed the best and was able to predict the user's conversational engagement quite well.

Author Keywords

Empirical study, eye-gaze behavior, conversational engagement, Wizard-of-Oz experiment

ACM Classification Keywords

H5.2. Information System: User Interfaces

General Terms

Algorithms, Design, Human Factors

INTRODUCTION

In face-to-face conversations, not only the speaker presents communicative signals accompanying speech, but also the listener display nonverbal signals, such as eye-gaze and head nods as feedback to the speaker. Argyle & Cook [1] showed that a listener's eye-contact expresses his/her attention toward the conversation.

We expect that exploiting eye-gaze information in designing user interfaces improves the naturalness of human-computer interaction specifically in conversational agent systems. If the system can judge the user's engagement in the conversation, it becomes possible to change the agent's actions and responses according to the user's attitudes. However, there has not been enough research in mechanisms of judging the user's engagement and implementation of such mechanisms into fully automatic conversational agents.

In our previous work [2-4], we demonstrated the close connection between user's conversational engagement and gaze transition patterns: patterns of shifting her/his attention. Then, we developed a method of estimating user's conversational engagement using the gaze transition patterns. Then, we implemented the method in a dialogue system. As the results of evaluation experiment, we found that in interacting with the implemented system, users felt that the agent's conversational functions were improved. However, we still have not been satisfied with the estimation accuracy.

In this study, to improve the engagement estimation accuracy, we re-analyze the user's gaze behaviors by adding new variables: occurrence of mutual gaze, gaze duration, distance of eye movement, and pupil size. For this purpose, this paper addresses the following issues:

- Analyze human-agent conversation corpus to investigate whether user's engagement is correlated with gaze transition patterns, mutual gaze, gaze duration, distance of eye movement, and pupil size.
- Evaluate the estimation methods and compare which combination of the parameters performs the best.

In the following sections, first we discuss related work, and then explain our collected data in an experimental setting. Based on the analysis of the data, our new engagement estimation methods will be proposed. Finally, the methods will be evaluated and compared, and then we conclude our empirical study.

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

RELATED WORK

In communication science and psychology, many studies investigated functions of eye-gaze in face-to-face communication. Kendon [5] very precisely observed eye-gaze behaviors by employing the ethnomethodological method and discussed various kinds 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 [1, 6]. This kind of mutual gaze also contributes to smooth turn-taking [7]. On the contrary, 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 [8-10].

These findings were later used as the basis of conversational humanoids. Nakano et al [11] proposed a gaze model for nonverbal grounding in ECAs that judges whether the user understands what the agent said or not. They used this model to control human-agent conversations. In human-robot communication research, Sidner et al [12] proposed a gaze model for conversational engagement, and implemented a model using a head tracker to judge whether the user is engaged in the conversation or not from the user's head movements.

In addition to the gaze direction, some other information obtained from "eye(s)" is also useful in HCI. Qvarfordt et al [13] developed an interactive system, iTourist, for city trip planning, which encapsulated knowledge of eye-gaze patterns and duration gained from studies of human-human collaboration systems. User study results showed that it was possible to sense users' interest based on the combination of eye-gaze patterns and gaze duration. Iqbal et al. [14, 15] investigated the use of task-evoked pupillary response to provide a measure of mental workload for interactive tasks. Their results showed that a more difficult task demands longer processing time, induces higher subjective ratings of mental workload, and reliably evokes greater pupillary response at salient subtasks. Eichner et al [16] used eye-trackers in detecting an object which the user was interested in, and integrated an eye-tracker as a component of interactive interfaces.

Based on the discussion above, we believe that exploiting multiple types of information in eyes improves human-agent conversational engagement.

CORPUS DATA

Wizard-of-Oz experiment in human-agent conversation

We conducted a Wizard-of-Oz experiment where the agents' speech and behavior were controlled by an operator [3]. In the experiment, two subjects participated in each session. One of the subjects participated as a user (called the "user"), and the other subject as an observer (called the "observer"). The user listened to the agent's explanation lasted for about 3 to 5 minutes for each 6 cell phone (see Figure 1). The user can communicate with the agent using speech. A push-button device was given to both the user and the observer. The user was instructed to press the button if the agent's explanation was boring and the user would like to change the topic. The observer was instructed to press the button when the user looked bored and distracted from the conversation. In this study, these button pressing behaviors were used as the human judgment of user's conversational engagement.

Collected corpus

We collected 10 conversations whose average length was 16 minutes, and built a multimodal corpus consisting of verbal and nonverbal behaviors shown below;

- Verbal data: The user's speech was transcribed from the recorded speech audio, and the agent's utterances were extracted from the log of the Wizard-of-Oz system. The total number of the agent's utterances was 951 and that of the user's was 61.
- Nonverbal data: Since the agent's behaviors were pre-set in the Wizard-of-Oz system, the agent's gestures and gaze behaviors were able to be extracted from the system log. As for the user's nonverbal behaviors, we collected user's gaze data using Tobii X-50 eye-tracker in 50Hz.
- Human judgment of engagement: When the user and/or the observer pressed her/his button, lights went on in another room, and these lights were recorded as video data.

All these data were integrated in xml format, and can be visually displayed using the Anvil annotation tool [17].

ANALYSIS

As a preliminary analysis, we examine four factors in eye gaze: occurrence of mutual gaze, gaze duration, distance of eye movement, pupil size. Then, we will investigate the correlation between these factors and human judgments of engagement.

We regarded the user as disengaged from the conversation when either the user or the observer pressed the button. We

Figure 1. Conversational agent projected on a screen

found some cases that the observer pressed the button slightly after the user displayed their disengagement gaze behaviors. So, we annotated the user's disengagement time period from 10 second before the observer's pressing the button to the time that s/he released it.

In following sections, we analyze a correlation between gaze transition patterns and human judgment of user's engagement. Then, we investigate other types of eye-gaze information (i.e. mutual gaze, gaze duration, distance of eye movement, and pupil size).

Analysis of Gaze 3-gram

Defining 3-gram patterns

By following our previous study [2-4], gaze transition patterns were defined as gaze 3-gram: three consecutive gaze behaviors. The constituents of gaze 3-gram are the following three types of eye gaze.

- *T*: look at the target object of the agent's explanation, which is the discourse focus.
- *A*: look at the agent's head
- *F*: look at non-target objects, such as other cell phones and an advertisement poster ($F1 \neq F2 \neq F3$).

Since the agent was designed to look at the current target object most of the time, it is presumed that joint attention is established between the user and the agent when the user's gaze label is *T*.

Since the eye-tracker cannot measure the pupils movement during blinks, small blanks often occurred in gaze data. So, we counted two consecutive gaze data as one continuous eye-gaze if the same object was continuously looked at in both data and the data blank was less than 200 msec. On the contrary, if the focus of the attention changed after the data blank or the blank was longer than 200msec, these two gaze data were not merged. Suppose that the user's gaze direction shifts in the following order: *T*, (100msec blank), *A*, (50msec blank), *A*, (150msec blank), and *F1*. There is a 50msec data blank between two consecutive *A*s. Thus, these two data are merged into one block. As the result, a 3-gram constructed from this sequence is *T-A-F1*. If the data loss is longer than 1sec, such data was discarded as an incomplete 3-gram.

Correlation between 3-gram patterns and engagement

The results of the 3-gram analysis in 8 users are shown in Figure 2. In this graph, the probability of 60% means that this pattern co-occurs with the human disengagement judgment (i.e. overlap with the time period of pressing the buttons plus ten seconds before this) 60% of the time. For each 3-gram pattern, we calculated the probability of co-occurrence with the disengaging judgment.

As shown in Figure 2, the probability of disengagement judgment is different depending on the types of 3-gram. For example, *A-F1-F2* has the highest probability over 60%. This means that the user was judged as disengaged over 60% of the time when this pattern occurred. On the other hand, the probability for the *A-F1-T* 3-gram is only 13.3 %. These results suggest that the 3-grams with higher probability violate proper engagement gaze rules, and those with lower probability contribute to conversational engagement.

Analysis of mutual gaze

In face-to-face conversations, it is obvious that mutual gaze serves as feedback to the speaker. Mutual gaze gives a good opportunity for the listener to give feedback to the speaker, and for the speaker to receive the feedback and determine or perform conversational actions. Therefore, focusing on mutual gaze, we analyze correlation between mutual gaze and user's engagement.

To distinguish mutual gaze from use's gaze at the agent, the gaze category *A* (looking at the agent) was subcategorized into the following two labels: *M* (mutual gaze; the user establishes eye contact with the agent) and *A* (non-mutual gaze). Then, the probability of co-occurrence with disengagement judgment was calculated.

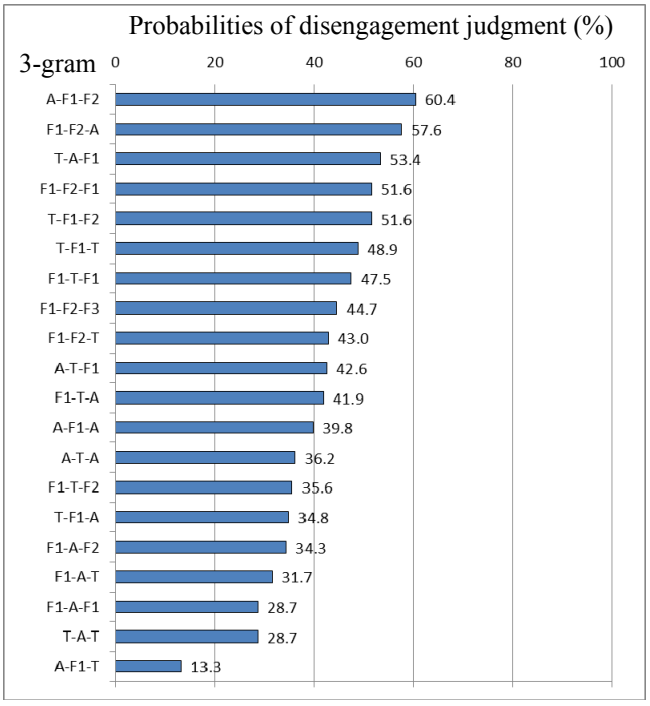


Figure 2. Eye-gaze 3-grams and probabilities of disengagement judgment

As shown in Figure 3, the new 3-grams with *M* label are indicated with red diagonal line's box. Some of these 3-grams have very high or low probability of disengagement judgment such as 100% or 0%. For example, *A-F1-A* 3-gram has the probability of 39.8% in Figure 2. By distinguishing mutual gaze from gazing at the agent, *A-F1-A* 3-gram is divided into four types of 3-grams, such as *A-F1-A*, *M-F1-M*, *M-F1-A*, and *A-F1-M* whose probabilities of disengagement judgment are 37.1%, 59.7%, 9.8%, and 100% respectively in Figure 3. Thus, by considering mutual gaze, the correlation between 3-grams and dis/engagement judgment becomes clearer. This result suggests that considering mutual gaze in categorizing gaze behaviors may improve the accuracy of estimating the conversational engagement.

Analysis of gaze duration

Our previous work [3] analyzed correlation between gaze duration and engagement, and suggested that looking at the target object and looking at the agent would be the positive signs of user's engagement. On the other hand, looking at non-target objects for a long time signals the user's disengagement from the conversation.

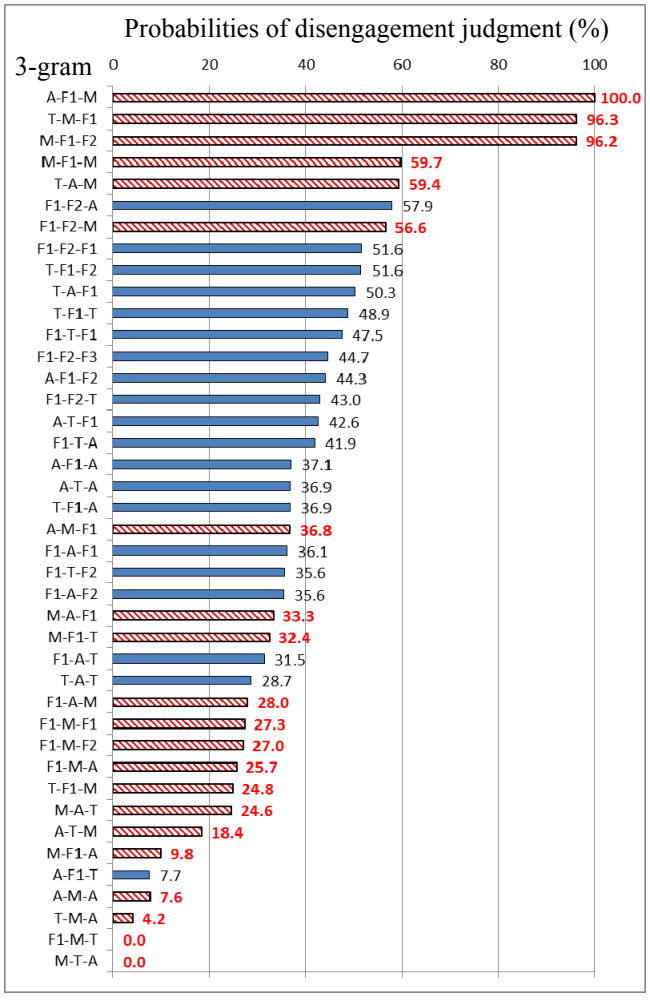


Figure 3. Probabilities of user's button pressed in mutual gaze or not situation

Based on these results, in this study, we considered the gaze duration in categorizing 3-grams. In the current analysis, we only considered the duration of the third constituent of a given 3-gram. We will extend this analysis by considering the duration for all the constituents in a 3-gram. According to the length of the third constituent, we classified 3-grams into long 3-grams and short ones. When the duration of the third constituent of a given 3-gram is shorter than the threshold, we labeled it with an extension "*_S*" at the end of the label. When the duration of the third constituent is longer than the threshold, we labeled it with an extension "*_L*". The thresholds were set at the values that clearly distinguish engagement from disengagement. For example, in *T-A-F1*, the probability of disengagement judgment was 0% when the duration of *F1* was 0 to 34 msec. The probability was 100% when the duration *F1* was 34 to 78 mesc. In this case, if the threshold is set at 34 msec, in long 3-grams (longer than 34 msec), the button pressing probability is 100%. By applying this threshold, we calculated the average probability of disengagement judgment for *T-F1-F_S* and *T-F1-F1_L*. Then, we found that the button pressing probability for *T-F1-F_S* (18.18%) was much lower than that for *T-F1-F_L* (100%).

Figure 4 shows the probability of co-occurrence between 3-grams with duration information and disengagement judgment. For example, *A-F1-F2* has the probability of 44.3% in Figure 3. By considering the duration information, a 3-gram of *A-F1-F2* is divided into *A-F1-F2_S* and *A-F1-F2_L* whose probabilities are 18.2% and 100% respectively. The new 3-grams containing *_S* or *_L* label at the end have very high or low probabilities of disengagement judgment such as 100% and 0% compared to the probabilities shown in Figure 3. Therefore, considering the duration in categorizing 3-grams is useful in distinguishing engagement patterns from disengagement ones. In particular, 3-grams whose probabilities are close to 100% or 0% almost always have "*_L*" label (indicated with box with orange diagonal lines). In other words, when the user looks at an object for a long time in the third constituent of a 3-gram, the user's engagement can be judged by looking at the preceding gaze behaviors (the first two gaze behaviors in a given 3-gram). Thus, these results suggest that considering gaze duration in categorizing gaze 3-grams may improve the accuracy of estimating conversational engagement.

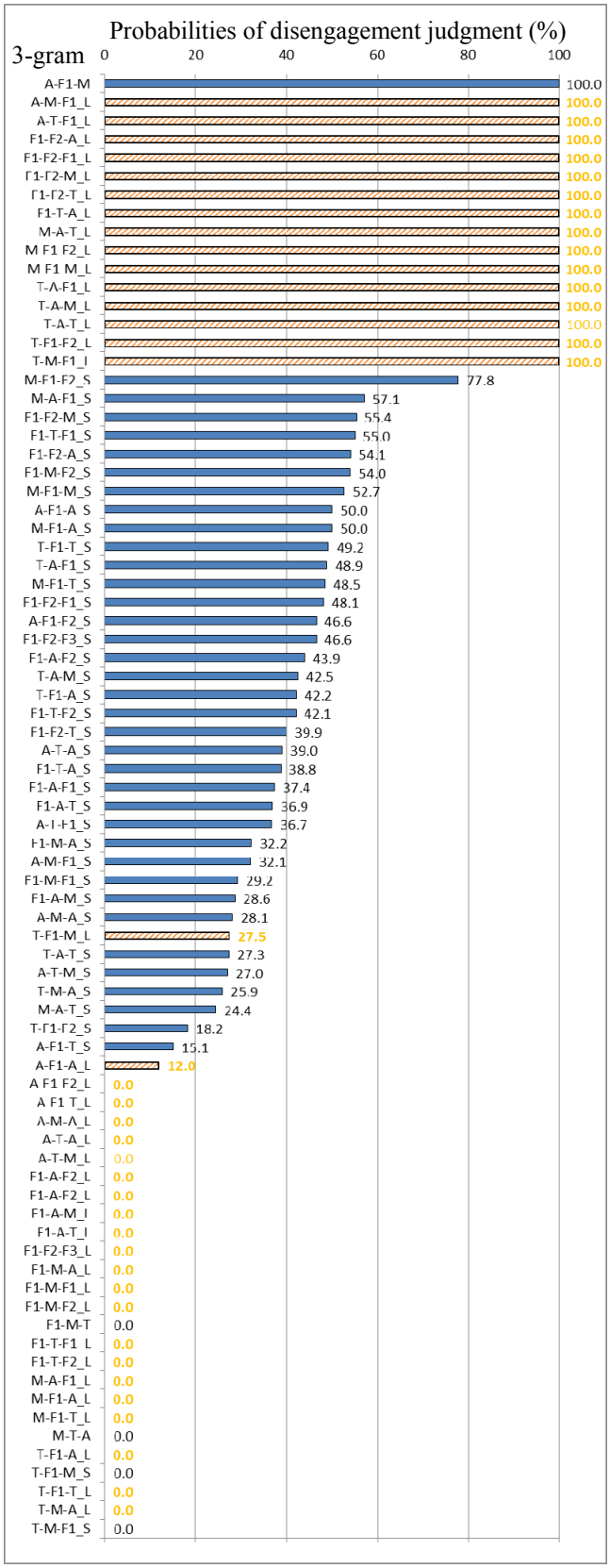


Figure 4. Eye-gaze 3-grams with duration and probabilities of button pressing

Analysis of distance of eye movement

We assume that user's conversational engagement is somehow related to the distance of eye movement during the conversation. For example, if the user is positively engaged in the conversation, distance of eye movement may be short because the user carefully looks at the object explained by the agent. On the other hand, if the user is not engaged, the distance of eye movement may become longer because the user looks at the objects which are not focused in the discourse, such as advertisement and cell phones currently not-explained. Therefore, we analyzed correlation between distance of eye movement and user's conversational engagement.

Figure 5 shows the moving average of 400 msec window for the distance of eye movement. The graph shows the proportion of occurrence of each distance value (this actually indicates the frequency of each distance value), the difference between engagement (dashed blue line) and disengagement (solid red line) is not very big, and average for each graph is 12.15 and 12.61 pixels, respectively. Although the peak of these graphs are not very different (Engagement: 5.25~5.50 pixels, Disengagement: 5.00~5.25 pixels), the distribution in disengagement situation is skewed to the right (the right tail is longer) compared to the distribution for engagement situation (the blue graph has larger values when the distance is 0 to 8 pixels). In other words, when the user is actively engaged in the conversation, the distance of eye movement is shorter than that in disengagement situation.

Thus, the results suggest that the distance of eye movement might be a weak predictor in estimating the conversational engagement.

Analysis of pupil size

It has already been known that pupil size becomes larger when people look at something interesting and exiting [18]. Therefore, it may be a reasonable assumption that the user's conversational engagement is somehow related to pupil size. For example, if the user is engaged in the conversation, the pupil size may become larger because the user carefully looks at the object explained by the agent. On the contrary, if the user is not engaged, the pupil size becomes smaller because the user doesn't seriously look at object. To examine this hypothesis, we analyzed the correlation between pupil size and user's conversational engagement.

Figure 6 shows the distribution of left eye pupil size data in engagement and disengagement situations. The x-axis indicates the pupil size, and the y-axis indicates the proportion of occurrence for each pupil size value. The average of pupil size for engagement is 5.20 cm and that is 5.16 cm for disengagement. Similar to the result of eye movement, although the averages of pupil size are not very different between engaged and disengaged situations, the distribution for engagement slightly shifts to the right compared to that for disengagement. As reported in previous studies, this result suggests that pupil size

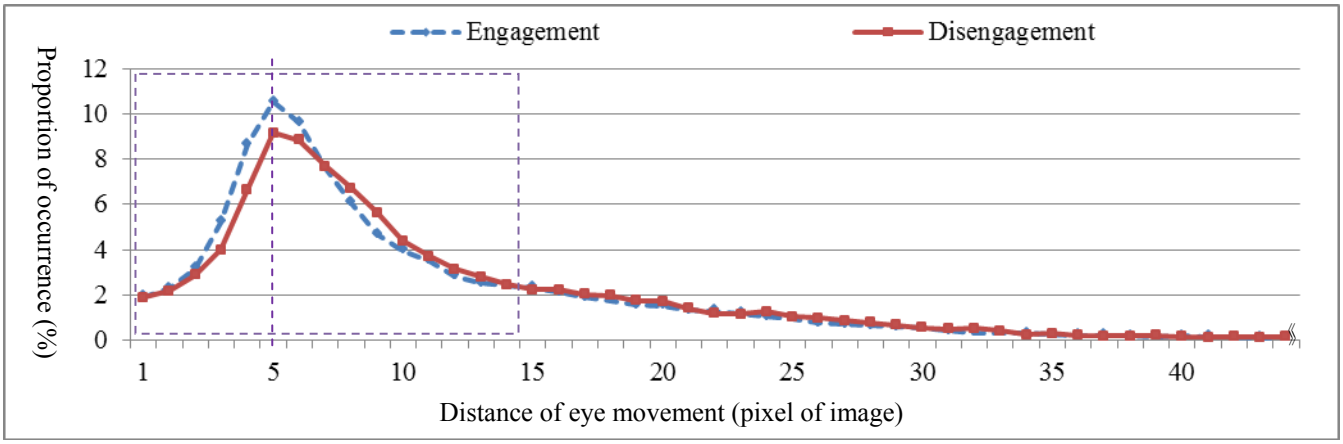


Figure 5. Distribution of distance of eye movement

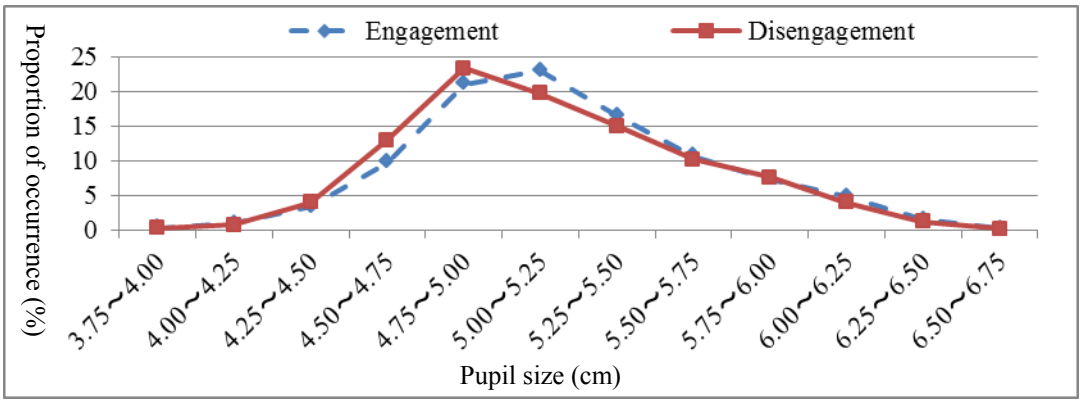


Figure 6. Distribution of pupil size

becomes larger when the user is engaged in the conversation and becomes smaller when the user is disengaged. Thus, pupil size may be a useful predictor of conversational engagement.

ESTIMATING USER'S ENGAGEMENT

Based on the analyses in previous sections, we found that 3-grams, mutual gaze, duration of eye-gaze, distance of eye movement, and pupil size may be useful as predictors of user's engagement in conversations. In this section, we estimate the user's engagement by employing SVM (Support Vector Machine) [19], and evaluate the accuracy and the effectiveness of each parameter. We used e1071 package[1] with R for SVM. The settings of SVM are RBF (radial basis function) kernel and default parameter of C (C = 1.0). The data used in SVM contains the user's conversational engagement state as a class, and the five eye-gaze parameters as features. Since we measured the user's behaviors 50 times per second (50Hz), the time resolution of the training data is also 50 Hz. The detailed explanation about each parameter is shown below;

- User's conversational engagement state: We have two classes in this model; engagement and disengagement. A disengagement state was a time period from ten seconds before the observer pressed button to when the button is released.
- 3-gram: The original types of gaze transition 3-gram
- 3-gram with mutual gaze: Distinguishing mutual gaze (*M* label) and gaze at the agent (*A* label) in constructing 3-grams.
- Duration of eye-gaze: The time duration from the start to the end of the third constituent of a given 3-gram.
- Distance of eye movement: The sum of the distance of eye movement for the last 400 msec.
- Pupil size: The average of both-eyes' pupil size data measured by an eye-tracker.

We tested the following estimation models that use these parameters above:

- 3-gram-based model (3-gram): Using original types of 3-gram without considering mutual gaze.

[1] http://cran.r-project.org/web/packages/e1071/

Model	State	Precision	Recall	F-mesure
3-gram	Engagement	**0.634**	**0.955**	**0.762**
	Disengagement	**0.532**	**0.085**	**0.147**
3-gram + M	Engagement	**0.645**	**0.950**	**0.768**
	Disengagement	**0.579**	**0.115**	**0.192**
3-gram + M + Dr	Engagement	**0.721**	**0.873**	**0.790**
	Disengagement	**0.658**	**0.419**	**0.512**
3-gram + M + Ds	Engagement	**0.674**	**0.877**	**0.762**
	Disengagement	**0.560**	**0.269**	**0.363**
3-gram + M + PS	Engagement	**0.664**	**0.937**	**0.777**
	Disengagement	**0.627**	**0.183**	**0.283**
ALL	Engagement	**0.769**	**0.918**	**0.837**
	Disengagement	**0.793**	**0.534**	**0.638**

Table 1. Results of evaluation

- 3-gram+M: 3-gram with considering mutual gaze.
- 3-gram+M+Dr: 3-gram with mutual gaze and the duration of eye-gaze.
- 3-gram+M+Ds: 3-gram with mutual gaze and the distance of eye movement.
- 3-gram+M+PS: 3-gram with mutual gaze and pupil size
- All parameters model (ALL): 3-gram with mutual gaze, the duration of eye-gaze, the distance of eye movement, and the pupil size.

The results of SVM are shown in Table 1. As the overall evaluation, F-measure of All parameters model (ALL) is 0.837, which is the best score among the other models. This result suggests that all the parameters contribute to estimate the user's conversational engagement. In comparing 3-gram and 3-gram+M, we did not find a big difference between them. In disengagement judgment, F-measure for 3-gram model is 0.147, and that for 3-gram+M is 0.192. This is because the agent rarely looks at the user in this corpus, and we did not collect enough number of 3-gram data accompanied by mutual gaze (indicated by *M* label). The performance of 3-gram+M+Dr is much better than 3-gram+M, specifically the performance of estimating the disengagement is improved (F-measure goes up from 0.192 to 0.512). This result suggests that duration of gaze is a strong predictor of user's engagement. In comparing a 3-gram+M and a 3-gram+M+Ds, the F-measure was 0.192 in 3-gram+M, and that was 0.363 in 3-gram+M+Ds. Moreover, in comparing 3-gram+M and 3-gram+M+PS, F-measure of 3-gram+M+PS (0.283) are better than that of 3-gram+M (0.192). This suggests that distance of eye movement and pupil size are useful in estimating the conversational engagement.

DISUCUSSION

The evaluation results showed that all the parameters proposed in this study are useful in estimating user's attitude towards the conversation. However, the parameters may need to be adjusted according to the user, situation, and conversational content or theme. For example, pupil size may change depending on the user's emotion and the brightness of visual stimuli. Individual difference is also not very small. Moreover, the distance of eye movement may different depending on the place and layout of visual stimuli. Therefore, to estimate the user's attitude more precisely, it is necessary to establish a situation adaptive model. On the contrary, 3-gram itself is not seriously affected by these environmental factors. It is suggested that because of such robustness, the performance of 3-gram+Dr model is better than 3-gram+Ds and 3-gram+PS model.

CONCLUSION

Aiming at estimating user's conversational engagement from eye-gaze information, this paper analyzed different kinds of gaze information: mutual gaze, gaze duration, distance of eye movement, and pupil size.

Based on the results of the analysis, we integrated all kinds of eye-gaze information investigated in our empirical study, and applied them to SVM to test whether each parameter contribute to the model or not. As the results of testing various combinations of these parameters, it was revealed that a model with all the parameters performs the best, and can predict the user's conversational engagement quite well.

We have already implemented a fully autonomous conversational agent that incorporates the user adaptive engagement estimation method. As future work, we will improve the estimation method by adding more information by other nonverbal behaviors, such as head movement, facial expression, and posture.

ACKNOWLEDGMENTS

We would like to express great thanks to Yoshihiko Suhara in NTT Cyber Solution Laboratories for this professional advice about SVM technique.

This work is partially funded by the Japan Society for the Promotion of Science (JSPS) under a Grant-in-Aid for Scientific Research in Priority Areas "i-explosion" (21013042), and a Grant-in-Aid for Scientific Research (S) (19100001).

REFERENCES

1. Argyle, M. and Cook, M., *Gaze and Mutual Gaze*. (1976), Cambridge: Cambridge University Press.
2. Ishii, R and Nakano, Y.I., *Estimating User's Conversational Engagement Based on Gaze Behaviors.* in *the 8th international conference on Intelligent Virtual Agents (IVA'08).* (2008): Springer. pp. 200-207.
3. Ishii, R. and Nakano, Y.I., *An Empirical Study of Eye-gaze Behaviors: Towards the Estimation of Conversational Engagement in Human-Agent Communication*. in 2010 International Conference on Intelligent User Interfaces (IUI2010), Workshop on Eye Gaze in Intelligent Human Machine Interaction. (2010).
4. Nakano, Y.I. and Ishii, R., *Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations*. in 2010 International Conference on Intelligent User Interfaces (IUI2010). (2010).
5. Kendon, A., *Some Functions of Gaze Direction in Social Interaction.* Acta Psychologica, (1967). 26: pp. 22-63.
6. Clark, H.H., *Using Language*. (1996), Cambridge: Cambridge University Press.
7. Duncan, S., *Some signals and rules for taking speaking turns in conversations.* Journal of Personality and Social Psychology, (1972). 23(2): pp. 283-292.
8. Argyle, M. and Graham, J., *The Central Europe Experiment - looking at persons and looking at things.* Journal of Environmental Psychology and Nonverbal Behaviour, (1977). 1: pp. 6-16.
9. Anderson, A.H., et al., *The effects of face-to-face communication on the intelligibility of speech.* Perception and Psychophysics, (1997). 59: pp. 580-592.
10. Whittaker, S., *Theories and Methods in Mediated Communication*, in *The Handbook of Discourse Processes*, Graesser, A., Gernsbacher, M., and Goldman, S., Editors. (2003), Erlbaum, NJ. pp. 243-286.
11. Nakano, Y.I., et al. *Towards a Model of Face-to-Face Grounding*. in *the 41st Annual Meeting of the Association for Computational Linguistics (ACL03)*. (2003). Sapporo, Japan. pp. 553-561.
12. Sidner, C.L., et al., *Explorations in engagement for humans and robots.* Artificial Intelligence, (2005). 166(1-2): pp. 140-164.
13. Qvarfordt, P. and Zhai, S., *Conversing with the user based on eye-gaze patterns. In Proceedings of the SIGCHI Conference on Human Factors in Computing System.* CHI '05. ACM Press, New York, NY, 221-230, (2005).
14. Iqbal, S.T. Adamczyk , P.D. Zheng, X.S. and Bailey B.P., *Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution*. *Proc. CHI'05.* Portland, OR: ACM, 311-320, (2005).
15. Iqbal, S.T. Zheng, X.S. and Bailey, B.P., *Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction*. *Proc. CHI'04.* Vienna: ACM, 1477-1480, (2004).
16. Eichner, T., et al. *Attentive Presentation Agents*. in *The 7th International Conference on Intelligent Virtual Agents (IVA)*. (2007). pp. 283-295.
17. Kipp, M. *Anvil - A Generic Annotation Tool for Multimodal Dialogue*. in *the 7th European Conference on Speech Communication and Technology*. (2001). pp.1367-1370.
18. Hess, E.H., *Attitude and Pupil Size.* Scientific American, (1965). 212: pp. 46-54.
19. Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, (1995).

Model-Based Eye-Tracking Method for Low-Resolution Eye-Images

Takashi FUKUDA[*1] **Kosuke MORIMOTO**[*1] **Hayato YAMANA**[*1*2]

[*1.] Dep. of Computer Science and Enginerring, Waseda Univ.
3-4-1 Okubo, Shinjuku, Tokyo 169-8555 JAPAN
{t_fukuda, morimoto, yamana}@yama.info.waseda.ac.jp

[*2.] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430 JAPAN

ABSTRACT

Techniques for estimating gaze without restricting user movements are highly desired for their potential applications. Although commercial gaze estimation systems achieve high accuracy using infrared light, gaze estimation systems with webcams have become indispensable because of their low price. The problem using webcams is that their resolution is too low to estimate the direction of eye-gaze accurately without restricting user movements. Low-resolution eye-images have strong sources of noise that result in inaccurate estimation. In this paper, we propose a novel gaze-estimation method that uses both image processing and geometrical processing to reduce various kinds of noise in low-resolution eye-images and thereby achieve high accuracy. In our experiments, the mean horizontal error of 5 participants is 3.0° and the mean vertical error is 1.5° with calibration.

Author Keywords

Gaze, Eye tracking, Low resolution, webcam.

ACM Classification Keywords

H1.2 User/Machine Systems: Human factors

INTRODUCTION

Gaze information has been widely used for studies in neuroscience, psychology, human engineering, marketing strategy, advertising, and so on [4], because the direction of eye-gaze gives clues about a person's interest and attention. In such studies, gaze information extracted from a variety of people in natural situations is highly desired to avoid any effects of the measurement system on the testees. In other words, the system should not restrict the movement of testees and should not require testees to wear any equipment.

Moreover, in order to gather gaze information from a variety of people, low-cost systems are desirable. However, conventional commercial eye-tracking systems [2,11,13] are too expensive, costing as much as several tens of thousands of US dollars.

One way to solve the abovementioned problems is to adopt low-cost webcams that cost just a few tens of US dollars but with a resolution of only 640 × 480 pixels. Although webcams are low in cost, eye-images extracted from webcams have low resolution because it is impossible to adopt a zoom-in feature to capture eye-images without restricting the movement of users. Agustin et al. [1] used low-cost web cams. However, in their method, users have to wear some equipments to set webcams near by their eyes and to capture high resolution eye-images. This is uncomfortable for users. Therefore, eye tracking using low-resolution eye-images becomes indispensable to solve the problems.

Previous studies based on low-resolution eye-images have some drawbacks. Ono et al. [9] used machine learning correlating the appearance of low resolution eye-images to the points of gaze. Their method shows good accuracy in that the mean error was 2.4°; however, their method has one drawback—a learning process that is too complicated. A user must gaze twenty times at twenty points from various positions for each point. On the other hand, Yamazoe et al. [16] used a model matching method with low-resolution eye-images. The mean errors were 5.3° horizontally and 7.7° vertically. This low accuracy is resulted from their rough eye model.

In this paper, we propose an accurate model-based gaze-estimation method with low-resolution eye-images. In our method, eye orientation is estimated accurately without quantization. We observe the pupil as a circle and then estimate the normal direction of the pupil as the gaze direction. We approximate the observed pupil contour with an ellipse and estimate the direction normal to this ellipse.

Low-resolution eye-images have strong sources of noise distorting the pupil contours. The approximation of the pupil contour with an ellipse using hough transformation is robust against noises. However, hough transformation has two drawbacks—the huge computation and the nonunique result. Hough transformation is a kind of vote. Its voting procedure is carried out in a parameter space exhaustively. Therefore, it has huge computation in ellipse approximation that has five-dimensional parameter. Furthermore, hough

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

(a) (b) (c) (d) (e) (f)

Figure 1 Examples of noise in low-resolution eye-images

transformation could have two or more results which receive the most votes. Takegami et al. [12] and Tsuji et al. [14] resolved these problems by constraining some parameters in hough transformation. However, they constrained parameters by restricting the movements of users.

We use the least square method (LSM) to allow users to move. The LSM is fast and results uniquely with no constraint. However, the LSM results wrong with the distorted ellipse contour. We adopt image processing and geometrical processing to remove the distortions from the pupil contour in low-resolution images. In our method, we expand low-resolution eye-images with the Bicubic convolution and estimate the sub-pixel pupil contour to reduce the approximation error. Furthermore, we fit the convex hull to the pupil contour to remove the distorted parts of the contour. The rest of this paper is organized as follows: In Section 2, we review related work. In Section 3, we propose our method. We then describe the experimental evaluation of the method in Section 4. Finally, we conclude the paper in Section 5.

RELATED WORK

Related work based on low-resolution eye-images is classified into two categories; appearance-based and model-based methods.

Appearance-based method

In appearance-based methods, the gaze direction is estimated using Machine Learning (ML). The classifier learns the relation between the appearance of the eye-images and gaze direction. Ono et al. [9] used N-mode Singular Value Decomposition (SVD). The resolution of the eye-images is 24 × 12 pixels, and the mean error is 2.6°. This shows a high accuracy for low-resolution eye-images. However, in this method, the calibration procedure is extensive in that each user must gaze at each point at least 20 times from various positions. This is too demanding for practical use.

Model-based method

In model-based methods, pre-generated 3D eye-models are used to estimate the gaze direction. The major model-parameters are the position, tilt, and rotation of the eye. The system calculates the parameters without any learning process.

However, most researchers assume that high-resolution eye-images are available with model-based methods [3,5,7,8,10,12,14,15]. They capture high-resolution eye-images by zooming in on movement-restricted users. To the best of our knowledge, there exist a few model-based gaze tracking methods using low-resolution eye-images.

Yamazoe et al. [16] used 30 × 15 pixel eye-images to extract the direction of eye-gaze with a model-based method. The mean errors were 5.3° horizontally and 7.7° vertically. They consider a line from the center of the eyeball to the center of the pupil as the line of gaze. They estimate the position of the center of the eyeball by matching the eye-images with pre-generated 2D eye-images using maximum-likelihood estimation. They generate 2D images by projecting variously oriented 3D eye-models that contain eyeballs, pupils, and eyelids. The 3D eye-models are rotated 5° each. However, this quantization (i.e., 5° each) is too coarse to achieve high accuracy. When the eye-image does not completely match with that of a pre-generated 3D model, the estimated center of the eye deviates considerably from the real position. This results in a low accuracy of gaze estimation.PROPOSED METHOD

Problem Formulation

Previous model-based eye-tracking methods using low-resolution eye-images face the difficulty of estimating detailed eye-model parameters.

The low-resolution eye-images are too coarse to determine the position of the pupil contour in the image. Figure 1 (a) shows an example of part of a low-resolution eye-image, especially around the pupil contour. Figure 1 (a) is generated by reducing the resolution of Figure 1 (b) to one tenth. The black pixels represent the pupil and the white pixels represent all other areas of the eye. The contour of the pupil must be placed on the gray pixels; however, we cannot determine where the true contour is in the gray pixels because of the low image resolution. The error in pupil contour estimation caused by these gray pixels results in a significant error in eye tracking. Consider that the true contour of the pupil is a circle whose diameter is 15 pixels, and that the estimated pupil contour is an ellipse whose major diameter is 15 pixels and minor diameter is 14 pixels. This 1-pixel error causes an error of 21° in the estimated gaze direction.

At first glance, it might appear appropriate to expand the image using Bicubic convolution in order to estimate the pupil contour more precisely. However, this expansion generates other noise. Figure 1 (c) shows Figure 1 (a)

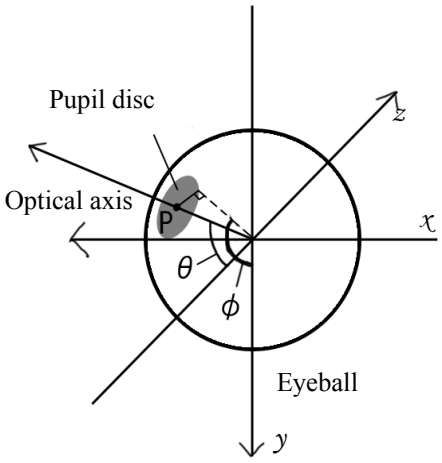


Figure 2 3D eye-model

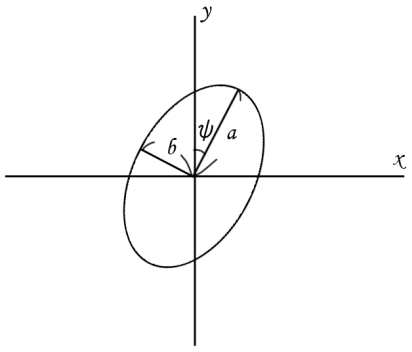


Figure 3 Example of ellipse approximation

expanded ten-fold. The wave of the contour is generated using Bicubic convolution.

In addition, the expansion emphasizes noise in the original images. Figure 1 (d) shows a low-resolution version of Figure 1 (e). The gray ellipse is included in Figure 1 (e) because of the light reflection on the pupil. In this case, we obtain Figure 1 (f) by expanding Figure 1 (d). The gray circle in the pupil in Figure 1 (e) becomes a dent in the pupil contour as shown in Figure 1 (f).

Such noise causes distortion of the pupil contour. As a result, the approximation of the pupil contour with an ellipse fails and the accuracy of eye tracking decreases. Reducing these distortions is the main problem we address in this paper.

Overview of our proposed method

We propose a method to remove the distorted portions of the pupil contour to approximate the pupil contour accurately with an ellipse. This increases the accuracy of model-based eye tracking using low-resolution eye-images.

We adopt both geometrical processing and image processing to remove distortions of the pupil contour. We expand low-resolution eye-images with Bicubic convolution and fit the convex hull to the pupil contour. Furthermore, we estimate the sub-pixel pupil contour to reduce the approximation error. After noise reduction, we approximate the pupil contour with an ellipse.

Our eye-model itself is the same as that Wang et al. [15]. Figure 2 shows the eye-model. Wang et al. modeled an eyeball as an ideal sphere, and the pupil as a disc fitted into a hole on the sclera sphere. The line normal to the pupil disc from the center of the disc represents the optical axis of the eye. In Figure 2, $\theta\left(0 \le \theta < 2\pi\right)$ shows the angle between the optical axis of the eye and the z axis. On the other hand, $\phi\left(-\pi < \phi < \pi\right)$ shows the angle between the optical axis of the eye projected onto the x-y plane and the y axis. We refer to θ as "eye tilt angle" and call ϕ the "eye rotate angle." We define $\phi > 0$ when the x-coordinate of the center of the pupil disc is positive.

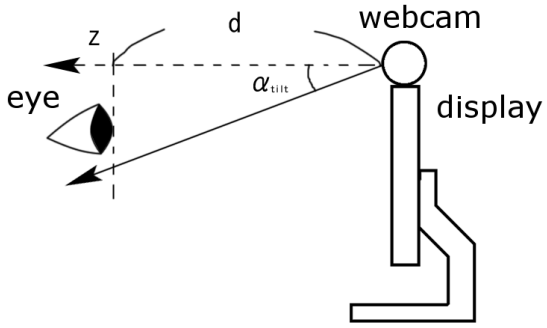


Figure 4 Tilting cameras against the display

By considering this model, we are able to approximate the pupil contour observed by cameras with an ellipse.

Figure 3 shows an example of approximating of the pupil with an ellipse. In Figure 3, ψ is the angle of rotation, *a* is the major radius, and *b* is the minor radius. We calculate θ and ϕ using these parameters as shown below.

$$\theta = a\cos\left(\frac{b}{a}\right) \tag{1}$$

$$\phi = \psi \text{ or } \pi - \psi \tag{2}$$

We must determine the value of ϕ uniquely. Thus, we set webcams tilted downward relative to the display as shown in Figure 4. The user's eyes are then above the optical axis

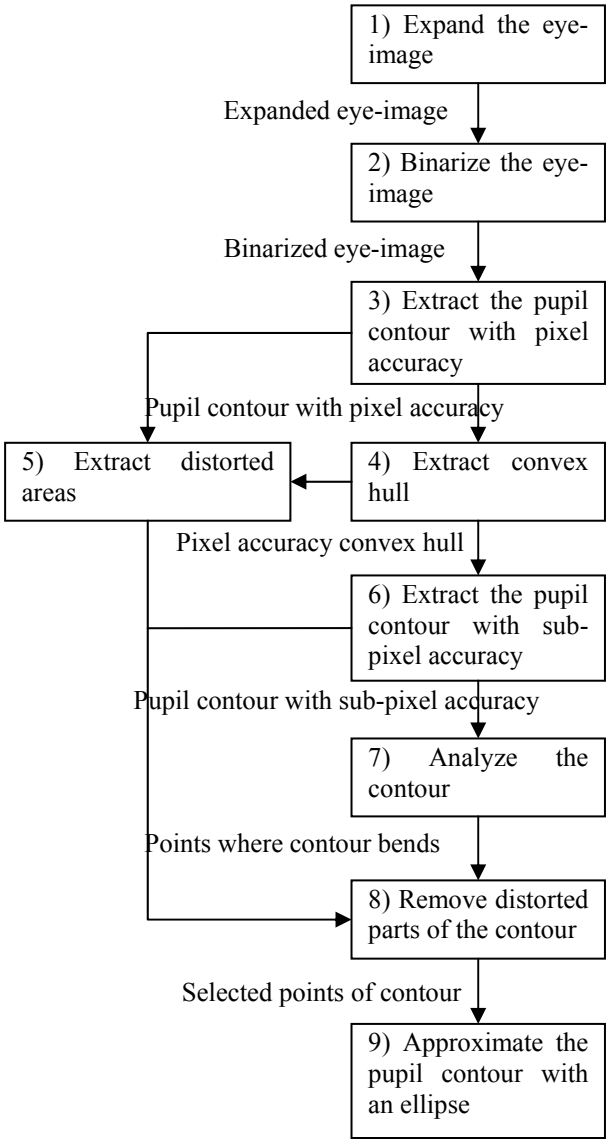


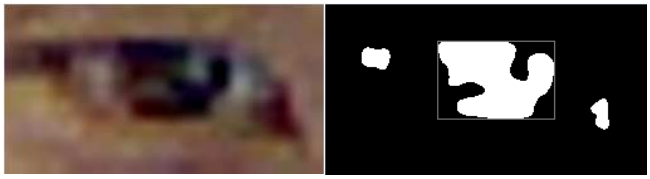
Figure 5 Image processing and geometrical processing in the proposed method

of the webcam. We can bind the range of the value, $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$.

Ellipse approximation

Low-resolution images have strong sources of noise that distort the contour of the pupil. The contour distortion causes errors of ellipse approximation.

To reduce the approximation error, we expand low-resolution eye-images using Bicubic convolution and estimate the sub-pixel pupil contour to reduce the approximation error. Subsequently, we fit the convex hull to the pupil contour to remove the distorted parts of the pupil contour.



(a) (b)

Figure 6 Result of binarization

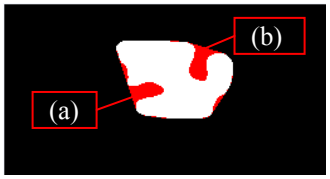


Figure 7 Examples of dents in a pupil and convex hull

In this paper, we categorize the causes of ellipse approximation error into two types: dents by light reflection, and distortions of the pupil contour by noise.

As shown in Figure 5, we apply image processing and geometrical processing to the eye-image.

Expanding the eye-image

We expand the eye-image to reduce the influence of 1-pixel error in ellipse approximation to gaze estimation. We adopt Bicubic convolution to expand the eye-image. Figure 6(a) shows an example of expanded eye image.

Binarizing the eye-image

Binarization cuts out the pupil from the eye-image using only the red channel of the eye-image. We call this R-channel. This is because the R-channel is useful for distinguishing the pupil from the other parts of the eye for Japanese people. Here, Japanese skin color has an R-channel value higher than both G and B, and the R, G, and B values are all high for the sclera because of its white color. On the other hand, the R, G, and B values are all low for the pupil because it is black. Figure 6 (b) shows the result of binarizing Figure 6 (a). In Figure 6 (b), the white areas are darker than a specified threshold. The system determines the largest white area to be the pupil. In Figure 6 (b), we treat the white area in the gray square as the pupil. In our method, we set the threshold manually.

Extracting the pupil contour with pixel accuracy

To obtain the precise location of the pupil, we need edge points. Using the information of the boundary pixels where the value of binary image changes, the pupil contour can be extracted with pixel accuracy. We assume that contiguous boundary pixels form a contour.

Fitting the convex hull to the pupil contour

In Figure 7, the white area represents the pupil area with pixel accuracy. In Figure 7, two dents, (a) and (b), represent the bright areas caused by light reflection before binarization. These distortions on the pupil contour reduce the accuracy of ellipse approximation.

Convex hulls fill dents such as (a) and (b). Figure 7 shows an example of dents filled by a convex hull. In order to ignore such dents, the system fits the convex hull to the pupil contour with pixel accuracy. However, the convex hull fills dents with line segments that cannot match the true contour of the pupil perfectly. Given that the convex hull represents the rough shape of the pupil, the system uses the convex hull of the pupil contour with pixel accuracy when analyzing the contour process.

Extracting the dent index

The dent index shows the start points and end points of the dents filled by the convex hull. This is calculated by matching the pupil contour with pixel accuracy with its convex hull. The dented portions of the contour do not match with its hull convex. The both ends of these portions of the convex hull are the dent index.

The system uses the dent index to remove the areas filled by the convex hull in the ellipse approximation phase. As the convex hull fills dents in the pupil contour with line segments, the filled areas do not match the true pupil contour completely. In the ellipse approximation process, these areas act as sources of noise. Therefore, the system should remember the dent index and remove filled areas in the ellipse approximation phase.

Estimating pupil contour with sub-pixel accuracy

In this step, pupil contour is estimated with sub-pixel accuracy. As pixel values are discrete, we apply linear convolution over the x-axis to obtain the pupil contour with sub-pixel accuracy. Figure 8 shows an example. The equations to calculate a point on the pupil contour with sub-pixel accuracy $p_{sub} = (x_{sub}, y_{sub})$ are as follows.

$$x_{sub} = \frac{x_0|th - v_1| + x_1|th - v_0|}{|th - v_0| + |th - v_1|} \quad (3)$$

$$y_{sub} = y_0 = y_1 \quad (4)$$

$$(p_0 = (x_0, y_0), p_1 = (x_1, y_1))$$

In Equations (3) and (4), v_0 is the intensity value of R-channel of p_0, v_1 is the value of p_1, and *th* is the threshold value.

The system fills the area of the pupil contour dented by the convex hull with sub-pixel accuracy.

Analyzing the contour

In order to remove the distortion by the eyelids, the system analyzes the pupil contour with sub-pixel accuracy whose dents are filled by the convex hull. The eyelid covers the pupil and distorts portions of the pupil contour. These covered areas are almost horizontal, and located between two bending points where the pupil contour bends sharply.

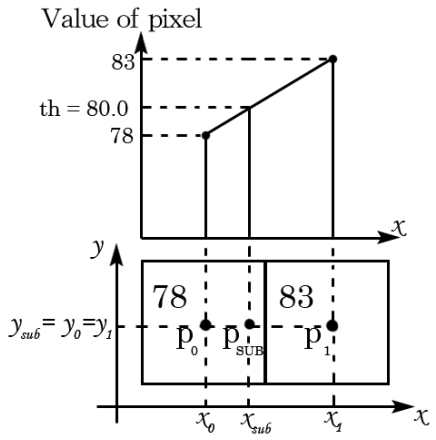


Figure 8 Pupil contour with sub-pixel accuracy

Figure 9 The bending points

To select bending points, the system calculates κ_i, the angle at point *i*. In equation (5), p_i is the *i* th point of the pupil contour with sub-pixel accuracy.

$$\kappa_i = \sum_k^n \frac{a\cos\left(\frac{p_i p_{i+k} \cdot p_i p_{i-k}}{|p_i p_{i+k}||p_i p_{i-k}|}\right)}{2n} \quad (5)$$

The point at which κ shows a local minimum becomes a candidate bending point. However, many points around the sharply bent point tend to show a local minimum. Therefore, we place two constraints. First, the bending points must have an interval longer than the threshold. Second, bending points must have a small κ. In particular, the bending point with a κ smaller than all other bending points must have the smallest κ among all contour points. Figure 9 shows an example of the result of contour analysis. Circled points represent the bending points. In this example, the length of the pupil contour with sub-pixel accuracy is 244 points where the threshold is 30 points.

Removing distorted portions of the contour

The system removes the distorted areas within the pupil contour with sub-pixel accuracy using the dent index and the bending points. First, the system simply removes areas where the dent index is evident. On the other hand, the bending points do not show distorted areas but indicate candidates of the start or end point of the area distorted by the eyelid. Therefore, the system removes areas that are nearly horizontal between two bending points. Specifically,

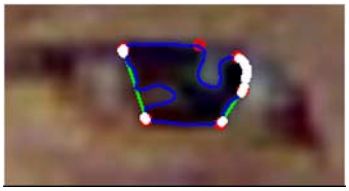


Figure 10 Selected points (White points)

User	*x*	*y*
A	0.46	0.62
B	0.65	0.79
C	0.57	0.35
D	0.57	0.61
E	0.70	0.54
Avg.	0.59	0.58

Table 1 Correlation between the estimation result and error

the system selects the areas to remove as follows. First, it selects two adjacent bending points. Second, it calculates the gradient of a line that includes the selected bending points. Finally, if the gradient is less than the threshold value, the system removes the area between the selected bending points. Figure 10 shows an example of the result of this phase. The remaining white points should comprise the portions of the true contour of the pupil.

Approximation of the pupil contour with an ellipse

In our method, the system approximates the pupil contour with an ellipse using the least square method (LSM). The sample points for the LSM are the points selected in the phase of removing distorted parts of the contour. As a result, the system obtains the major radius a, the minor radius b, and the rotation angle ψ of the approximated ellipse.

Gaze point estimation

The direction of eye-gaze is estimated in this step. Here, the eye-gaze point refers to location toward which the user is looking. In our method, the eye-gaze point for each eye is calculated using equations (1) and (2). Subsequently, the system determines the gaze point as the midpoint of the two eye-gaze points.

Calibration

The calibration process compares measurements and true values. The accuracy of measurements increases as calibration reduces sources of noise with systematic trends. In terms of eye tracking, calibration refers to the adjustment of parameters in gaze estimation. The eye-tracking system measures the difference between raw measurements and true values by measuring the gaze of the user gazing at known points. For examples, Nagamatsu et al. [6] and Ohono[8] et al. adjust the angle between the measured optical axis of an eye and the gaze direction. Chen et al. [3] adjust the distance between the center of eyeball and the center of the pupil.

User	*x*	*y*
A	9.3×10^{-10}	5.7×10^{-11}
B	1.6×10^{-11}	8.5×10^{-20}
C	1.3×10^{-7}	2.6×10^{-6}
D	2.7×10^{-9}	2.0×10^{-9}
E	4.2×10^{-9}	4.3×10^{-7}

Table 2 P-value of the t-test

	Max [cm]	Min [cm]	Range [cm]
X	11.7	-8.5	20.2
Y	5.1	-16.0	21.2
Z	89.6	36.2	53.4

Table 3 Width of users' head movements

In our method, the system adjusts gaze direction directly using a linearly approximated curve for the scatter plot of estimated direction and error in degrees.

Table 1 shows the correlation between the estimated result and error in our experiment. The estimated result is the gaze direction relative to the optical axis of the camera. The error is the angle between the estimated gaze line and the line from the midpoint of the two pupil centers to the correct point. In our experiment, we established 9 target points. We obtained 225 estimated results for the calibration. These results consist of 5 results per 9 target points for each of 5 users. Table 2 shows p-values of the t-test. Here, the null hypothesis is that no correlation exists between the estimated results and errors. As shown in Table 2, we can reject the null hypothesis at a significance level of 1%. Therefore, we expect calibration with linear approximation to raise the accuracy of gaze estimation. In fact, this is precisely what we observe in our next experiment.

EXPERIMENTAL EVALUATION

Procedure

We evaluated our method with and without calibration. The experimental procedures of both cases were almost the same. There were 5 subjects. All were males in their twenties without glasses. Only one subject wears contact lens. They sat in a chair as they would when normally using a PC. Table 3 shows the range of the users' head movements.

The subjects gazed at 9 points on the display, and the system estimated their gaze points. The coordinates of the target points were (50,50), (800,50), (1550,50), (50,600), (800,600), (1550,600), (50,1150), (800,1150), and (1550,1150) on a display with a resolution of 1600 × 1200 pixels.

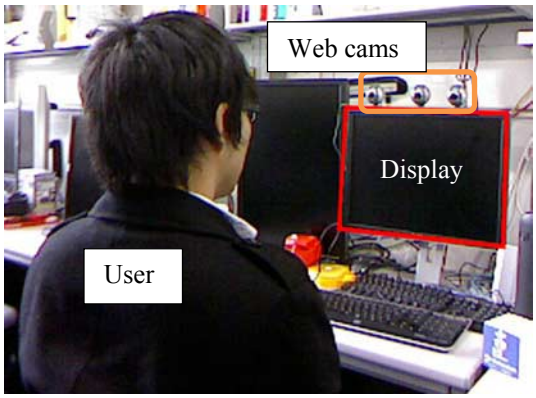


Figure 11 Environment of evaluation

User	$\overline{\lvert\theta_x\rvert}$ [deg]	σ_x [deg]	$\overline{\lvert\theta_y\rvert}$ [deg]	σ_y [deg]
A	5.4	7.3	2.0	2.5
B	3.5	4.2	2.2	2.6
C	3.6	4.0	2.0	1.9
D	3.3	4.1	1.5	2.3
E	3.3	3.9	2.0	2.2
Avg.	3.8	4.7	1.9	2.3

Table 4 Mean and standard deviation of the error degree in the case without calibration

In the case without calibration, each user observed 9 target points for 10 times. In total, we obtained 450 eye-gaze results from five users.

In the case with calibration, we divided the results into two sets: the learning and test sets. Each set included 225 results. These sets consisted of 5 results per target point. The system adjusts the result, i.e., applies calibration, using the learning set. Subsequently, another set, i.e., the test set, is used for testing after calibration.

Evaluation environment

Figure 11 shows the evaluation environment. We implemented a gaze-tracking system with three webcams (Logicool C500), one 20-inch display, and one PC (Windows 7, Core i7 950 3.07 GHz, Memory 12 GB). These cameras operated with a resolution of 640 × 480 pixels. The mean resolution of eye-images was 21.5 × 14.4 pixels. The software ran with a single thread.

Two webcams measured the 3D position of the pupil center. The other webcam captured eye-images for gaze-direction estimation.

Results

In the case without calibration

Table 4 shows the results without calibration. Subject E wears contact lens. We divide the error into two axes: horizontal and vertical. The horizontal axis is shown as x and the vertical axis is shown as y. $\overline{\lvert\theta_x\rvert}$ and $\overline{\lvert\theta_y\rvert}$ show means of the absolute error value. σ_x and σ_y show the standard deviation of the error value.

	Without calibration		With calibration	
User	$\overline{\lvert\theta_{Tx}\rvert}$ [deg]	$\overline{\lvert\theta_{Ty}\rvert}$ [deg]	$\overline{\lvert\theta_{Tx}'\rvert}$ [deg]	$\overline{\lvert\theta_{Ty}'\rvert}$ [deg]
A	6.0	2.0	4.5	1.7
B	3.3	2.1	2.4	1.3
C	3.3	1.6	2.9	1.3
D	3.1	1.7	2.3	1.6
E	3.1	1.8	3.0	1.5
Avg.	3.8	1.8	3.0	1.5

Table 5 Mean error of the test set in cases without and with calibration

	Without calibration		With calibration	
User	σ_{Tx} [deg]	σ_{Ty} [deg]	σ_{Tx}' [deg]	σ_{Ty}' [deg]
A	8.0	2.6	6.1	2.0
B	4.1	2.5	3.2	1.6
C	4.0	1.9	3.5	1.6
D	3.5	2.0	2.8	1.7
E	3.6	1.8	3.5	1.7
Avg.	4.6	2.2	3.8	1.7

Table 6 Standard deviation of the test set in cases without and with calibration

In the case with calibration

Table 5 and Table 6 show the results with calibration. According to these tables, the means and standard deviations decrease after calibration.

DISCUSSION

We compare our method to the method proposed by Yamazoe et al. [16], who used low-resolution eye-images in eye tracking. The mean resolution they used was 30 × 15 pixels. In addition, their method is a model-based method. These constraints are similar to those of our method.

In their experiment, Yamazoe et al. evaluated horizontal and vertical errors extracted from five users. The mean horizontal error was 5.3°, and the mean vertical error was 7.7°.

On the other hand, in our method, the mean horizontal error is 3.0° and the mean vertical error is 1.5°. In conclusion, our method reduced horizontal and vertical errors by 43.4 and 80.6%, respectively, compared to the method proposed by Yamazoe et al.

CONCLUSION

Gaze estimation with low-resolution eye-images is indispensable for achieving low-cost eye tracking. However, low-resolution eye-images have strong sources of noise distorting the pupil contour. This distortion results in inaccurate eye tracking. In this paper, we propose a detailed model-based eye-tracking method for low-resolution eye-images. We adopt image processing and geometrical processing to remove distortions from the pupil contour in low-resolution images. The resolution of eye-images we used was 21.4 × 14.4pixels and thus sufficiently low. The mean errors were 3.0° horizontally and 1.5° vertically. Our method reduced horizontal and vertical errors by 43.4 and 80.6%, respectively compared to a previous method.

REFERENCES

1. Agustin, J. S., Skovsgaard, H., Hansen, J. P. and Hansen, D. W. Low-Cost Gaze Interaction: Ready to Deliver the Promises. *Proc. of ACM CHI EA ' 09: extended abstracts on Human Factors in Computing Systems.,* (2009) 4453-4458.
2. CRS: http://www.crsltd.com/
3. Chen, J., Tong, Y., Gray, W., and Ji, Q. A Robust 3D Eye Gaze Tracking System using Noise Reduction. *Proc. of Eye Tracking Research & Applications Symp*., (2008) 189-196.
4. Duchowski, A.T. A Breadth-First Survey of Eye Tracking Applications. *Behavior Research Methods, Instruments, and Computers*, 34, 4 (2002), 455-470.
5. Hennessey, C., Noureddin, B., and Lawrence, P. A single Camera Eye-Gaze Tracking System with Free Head Motion. *Proc. of Eye Tracking Research & Applications Symp.*, (2006), 87-94.
6. Nagamatsu, T., Kamahara, J., Iko, T., and Tanaka, N. One-point Calibration Gaze Tracking based on Eyeball Kinematics using Stereo Cameras. *Proc. of Eye Tracking Research & Applications Symp.*, (2008) 95-98.
7. Nagamatsu, T., Kamahara, J., and Tanaka, N. Calibration-free Gaze Tracking using a Binocular 3D Eye-model," *Proc. of ACM Conf. on Human Factors in Computing Systems*, (2009) 3613-3618.
8. Ohono, T. and Mukawa, N. A Free-head, Simple Calibration, Gaze Tracking System that Enables Gaze-Based Interaction. *Proc. of ACM Eye Tracking Research & Applications Symp*., (2004) 115-122.
9. Ono, T., Okabe, T., and Sato, Y. Gaze Estimation from Low Resolution Images. *LNCS*, 4319 (2006) 178-188.
10. Shih, S.W. and Liu, J. A Novel Approach to 3-D Gaze Tracking Using Stereo Cameras. *Proc. of IEEE Trans. on System, Man, and Cybernetics*, 34, 1 (2004) 234-245.
11. SMI: http://www.smivision.com/
12. Takegami,T., Goto, T. and Ooyama. G., An Algorithm for Model-Based Stable Pupil Detection for Eye Tracking System, *IEICE Trans. Information and Systems, vol.J86-D-II(2)* (2003) pp.252-261.
13. Tobii: http://www.tobii.co.jp/japan/home.aspx
14. Tsuji, T., et al., Iris Detection Using LMedS for Eye Tracking System, *Proc. MIRU2004* (2004) pp.I-684-689.
15. Wang, J.G., Sung, E., and Venkateswarluye, R. Eye Gaze Estimation from a Single Image of One Eye. *Proc. of IEEE Int'l. Conf. on Computer Vision*, 1 (2003) 136-143.
16. Yamazoe, H., Utsumi, A., Yonezawa, T., and Abe, S. Remote Gaze Estimation with a Single Camera Based on Facial-Feature Tracking without Special Calibration Actions. *Proc. of Symp. on Eye Tracking Research & Applications*, (2008) 245-250.

Simulated Crowd: Towards a Synthetic Culture for Engaging a Learner in Culture-dependent Nonverbal Interaction

Sutasinee Thovuttikul, Divesh Lala, Hiroki Ohashi,
Shogo Okada, Yoshimasa Ohmoto, Toyoaki Nishida
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{thovutti@ii.ist, lala@ii.ist., ohashi@ii.ist., okada_s@, ohmoto@, nishida@}i.kyoto-u.ac.jp

ABSTRACT
Difficulties in living in a different culture are caused by different patterns of thinking, feeling and potential actions. A good way to experience cultural immersion is to walk in a crowd. This paper proposes a simulated crowd as a novel tool for allowing people to practice culture-specific nonverbal communication behaviors. We present a conceptual framework of a simulated crowd using an immersive interactive environment. We discuss technical challenges concerning a simulated crowd, including real-time eye gaze recognition in a dynamic moving situation, sensing of nonverbal behaviors using multiple range sensors, and behavior generation based on novel temporal data mining algorithms.

Author Keywords
Synthetic Culture, Embodied Conversational agents, Multi-modal interaction, learning by mimicking

ACM Classification Keywords
I.2.11 [Artificial intelligence] Distributed Artificial Intelligence –Artificial agents, multi-agent system.

INTRODUCTION
In daily life, humans use nonverbal communication to expand the meaning of verbal feelings. Eye movement plays a critical role. It helps the actor to both effectively and efficiently communicate her/his intention to the partner without speaking.

As pointed out by Knapp and Hall [1], nonverbal behaviors depend on numerous factors, including age, status, gender, role, context, emotion, mood, personality, and cultural background. When one enters another culture, s/he might feel uneasy from time to time until s/he has learned the patterns of thinking, feeling and potential actions shared in that culture [2].

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

Practice is often better than precept. A good way to practice culture is to walk in a crowd in which one need to pay attention to culture-specific nonverbal signals by which people cooperate or negotiate with each other to avoid collision and achieve the respective goals.

This paper proposes a simulated crowd as a novel tool for allowing people to practice culture-specific nonverbal communication behaviors. A simulated crowd can be characterized as one possible instantiation of synthetic culture [3], which specifies an artificial environment habited by synthetic agents behaving according to a parameterized norm. To realize a simulated crowd, we introduce an immersive interactive environment. We discuss real-time eye gaze recognition in a dynamic moving situation, sensing of nonverbal behaviors using multiple range sensors, and behavior generation based on novel temporal data mining algorithms as key technical contributions.

In what follows, we first discuss how nonverbal communication depends on culture. Then we introduce the idea of synthetic culture and simulated crowd as its instantiation. We then present the realization of simulated crowd using an immersive interaction environment. Finally, we discuss technical challenges together with preliminary results.

NONVERBAL COMMUNICATION
Nonverbal communication "involves those nonverbal stimuli in a communication setting that are generated by both the source (speaker) and his or her use of the environment and that have potential message value for the source or receiver (listener)" [4].

In principle, sending and receiving of messages occur in a variety of ways without the use of verbal codes, i.e., touch eye contact (eye gaze), volume, vocal nuance, proximity, gestures, facial expression, pause (silence), intonation, posture and smell. There are two basic categories of non-verbal language, nonverbal messages produced by the body and nonverbal messages produced by the broad setting (time, space, silence).

Eyes not only function to see things but also serve as a stimulus to be seen by others. Eye gaze cues are used to make inferences about others' cognitive activity, including their focus of attention, intention, desire, and knowledge about the current state of affairs. Kleinke [5] summarized how gaze functions (a) provide information, (b) regulate interaction, (c) express intimacy, (d) exercise social control, and (e) facilitate service and task goals. One of the important roles of eye gaze is as a social cue guiding attention. People share information about intentions and future actions using eye gaze.

Communicative functions implying future actions are particularly important in situations where a person is moving with others, such as playing sports, collaborating in physical tasks or walking in a crowd. The eye gaze manages the directions and the timing that people move and provides their intentions as to whether each person accepts the movement or not. The most significant point is that we can communicate that information in a short time by using eye gaze and other information. The function of eye gaze plays an important role in establishing quick and smooth interaction among the people.

Since nonverbal communication is rather polysemic, integrating multiple modalities is mandatory for robust interpretation of nonverbal behaviors. Morency et al [6] constructed a probabilistic model which predicts when to give listener backchannel using not only eye-gaze of the speaker but also the prosody and spoken words. They achieved better prediction of visual backchannel cues than previous studies by enabling the system to automatically select the relevant features. Huang et al [7] also used multimodal cues to realize a quiz agent that is attentive to its users' situation. They used audio and visual information to estimate the atmosphere of the interaction and who among the users is leading the conversation. They changed the agent's behavior according to those barometers. We believe the multi-modal information is very useful for our setting as well because we obviously use not only eye gaze but also some other cues to walk through the real crowd without conflicts. To infer the situation of the leaner, we propose using multi-modal information such as eye gaze, hand gesture, head direction, and torso direction.

Maia et al [8] devised an experiment to study the impact of eye gaze of humanoid avatars in conversation. They set four conditions: video, audio-only, random-gaze avatar and informed-gaze avatar. The four experiments compared the impact between avatar and non-avatar communication conditions. The informed-gaze avatar can relate the agent's eye gaze and conversation better than the random-gaze avatar. They concluded that the best result of the conversation is on agent who can relate the gaze to the conversation.

CULTURAL DIFFERENCES IN NONVERBAL COMMUNICATION

Different cultures have different type of communicative expressions, where verbal and nonverbal behavior is different. For instance, in some cultures standing close shows a familiar relationship between people. On the other hand, other cultures consider standing close as being rude. There are also differences in gestures, touching behaviors and eye gaze patterns [1].

Fehr and Exline [9] suggested that gaze is associated with dominance, power, or aggression. For example, when people are in the crowd or an elevator, they can adjust their personal space if they agree or limit eye contact [10]. Eye gaze has different meaning in each culture. Argyle et al [11] reported that English and Italians use direct eye gaze during conversation, whereas for Japanese and Chinese from Hong Kong seldom use direct eye contact in conversation. Watson [12] classified two categories, "contact culture" and "noncontact culture". Contact cultures such as Arabs, Latin Americans and Southern Europeans engage in more gaze, touch and close interpersonal distance during conversation than noncontact cultures.

The behavior of a human crowd is diverse. For example, the differences of Thai and Japanese culture in a crowd are walking speed, posture and eye contact. There is evidence that walking speed differs across cultures[13] Preliminary through observation, the posture of walking of Japanese people seems more gentle and graceful than Thai people, and eye contact of Japanese people is focused more toward their destination more than Thai people.

In the physical age, the learner has had to go to a different country for observation, recognition, and imitation of cultural behavior, if they would like to practice nonverbal culture. In the information age, however, we can use a simulated system to represent the virtual world where human communication takes place in a different place, culture and language. The virtual world allows us to design various types of synthetic culture which are introduced in the next section.

Christopher [14] presented the theory of mind which is used as a computing process of agent behavior. This theory controls agent interaction behavior, their eye, head and body directions, locomotion and greeting gestures. The direction detector detects the agent's eye gaze, i.e., direct or averted. The intentionality detector decides if the goal object is the desired one. Theory is the module store of the mental state of the agent. The result of this module is based on the interaction between the agents. In our approach, we use cultural behavior as agent behavior. The selection of suitable agent reaction behavior is for future work. The proper selection of agent behavior is based on cultural.

SYNTHETIC CULTURE

The idea of synthetic culture has been put forward by Hofstede [3] and is described as "role profiles for enacting dimensions of national culture". Hofstede [2] defines five dimension of national culture based on then aspect of each culture when measure relative with other cultures.

Dimension 1 Power distance. High power distance cultures believe the power in institution is distribute unequally.

Dimension 2 Individualism versus collectivism. This represents the difference between people. The collectivism culture feels about in-group or out-group.

Dimension 3 Masculinity versus femininity or achievement oriented versus cooperation oriented. This dimension describes how gender influences roles. In high femininity cultures both genders are assume to be cooperation oriented.

Dimension 4 Uncertainty avoidance. Weak uncertainty avoidance cultures believe "what is different is curious", but strong uncertainty avoidance cultures tend to think "what is different thing is curious is dangerous".

Dimension 5 Long-term versus short-term orientation. Long-term orientation cultures are driven by future-orientation (perseverance and thrift), while short-term orientation refers to cultures that are driven by past and present-orientation (respect for tradition).

Synthetic culture "does not exist in reality", but only exists in a gaming or training context. Hofstede himself gave ten such profiles of synthetic cultures, which could be of use in simulation and games.
The ten Synthetic-Culture Profiles [3] extend from the five dimension of national culture.
Dimension 1 Power Distance: *Hipow* is high power distance culture. *Lopow* is low power distance.
Dimension 2: *Indiv* is highly individualistic. *Collec* is extremely collectivity.
Dimension 3: *Achievor* is highly achievement orientation. *Caror* is highly cooperation orientation.
Dimension 4: *Uncavo* is strong uncertainty avoidance. *Unctol* is weak uncertainty avoidance.
Dimension 5: *Lotor* is absolutely long-term orientation and *Shotor* is very short-term orientation.

The variety characters are considered in the ten Synthetic-Culture Profiles and eye gaze is an important expression of many intentions. For instance, the expression of interest in *Hipow* is positive and animated, with no eye contact but the expression of interesting *Lopow* is animated, with eye contact and interjections. It is clear that the way to express the same intention is different in a contrary synthetic culture.

The concept of synthetic cultures is useful for the analysis of different cultures and intercultural awareness. In particular, our goal is to create agents which exhibit cultural attributes. Although infusing agents with these characteristics is difficult, synthetic cultures give us a method to achieve this by not having to recreate complex social norms, experiences or behavior that has been built throughout many generations. Using eye gaze is just one of the initial steps which we can take to achieve agents that possess cultural attributes. For example, we may find that avoidance behavior can be expressed through the lack of eye contact. In this case, a synthetic culture can be created which uses eye gaze in a way that can be recognized by the learner. By the same token, agents with a synthetic culture may recognize eye gaze from the learner and react to it accordingly.

SIMULATED CROWD AS AN INSTANTIATION OF SYNTHIC CULTURE

The simulated crowd is an instantiation of synthetic culture in which the learner can gain virtual experiences with a culture through interaction with synthetic characters that behave according to the parametric model of the culture. It allows the learner to interact with characters that embody various kinds of the behavior in the given culture. Figure 1 shows a person who is interested in nonverbal behavior in a different culture in the synthetic crowd. From the information above, we can create scenarios and role plays to drive learners to understand the differences of people from other cultures. In figure 1, the learner walks through the crowd among many people with varying cultures. Due to learner observation of nonverbal behavior, s/he can interact with people in the synthetic crowd.

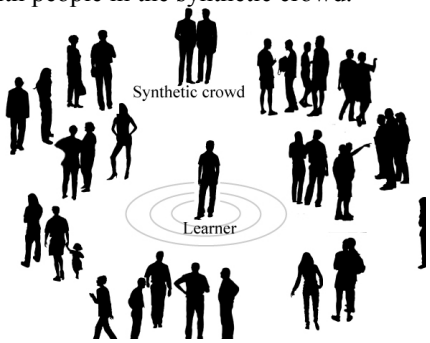


Figure 1. Synthetic Crowd and learner.

Klaus Dorfmueller-Ulhaas [15] developed an immersive 3D game consist of a 3D projection screen with shutter-glasses, a 3D sound system and an optical tracking system. Their system created a crowd simulation for a 3D game. The player controls his upper body and avoids the characters in crowd. The 3D technique is an ideal interaction for a player. s/he feel not only width and height but also depth on the screen. From this research we know this type of environment is important for a learning system.

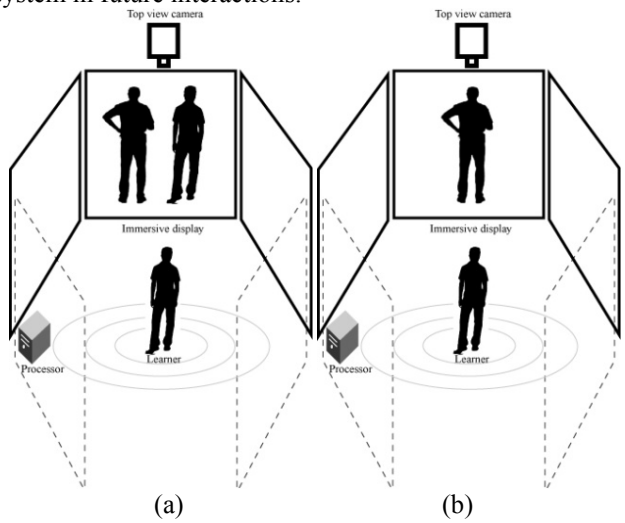
We plan to set the environment to enclose the learner like s/he is walking among people in the crowd. We use the learner's head direction as the future walking direction of learner. When the learner walks, s/he feels like they are walking through real crowd. We set various actions for agent behavior. The learner can observe the agent behavior for learning and recognizing the cultural behavior in a crowd or interacting with the agent such as walking close to, looking at, waving their hand to the agent.

Figure 2 shows our implementation of a simulated crowd using an immersive display system. The learner is among the virtual crowd. The player walks through the crowd for play the game. This research is a good example of walking

in crowd. For our approach we discuss a including nonverbal behavior from a variety of cultures into an agent. The learner can then practice interaction with various cultural agents.

Figure 2. Implementation of a simulated crowd using an immersive display

One may or may not display an avatar of the learner on the screen. In the avatar mode, both the agent and avatar represented learner are displayed on the screen as figure 3(a). In contrast, there is only agent on the screen in non-avatar mode as figure 3(b). From the learner's viewpoint, non-avatar mode represents a directly mutual the gaze between learner and agent (on the screen). Moreover, user distance is directly detected from the length between user and screen. On the other hand, in avatar mode the user sees both the avatar and the agent. The learner interprets the distance between the avatar and the agent on the screen. The learner's gaze need to be mapped as the avatar behavior and the learner is expected to interpret the gaze interaction of the avatar and the agent. In this approach we selected the non-avatar be proper in the initial step. However each approach has its own advantages and disadvantages, we will decide the best solution for the system in future interactions.



(a) (b)

Figure 3. show avatar and the agent on the screen (a) with-avatar (b) without-avatar Mode

We focus on several scenarios that are possible for investigating eye-gaze and the simulation of crowds. The following is one such scenario as the simple act of walking past a person coming from the opposite direction. For illustrate the scenario in figure 2-4, the avatar is a man on the left (in figure 2) and the agent is a woman in the right (in the figure 2). We show both the avatar and the agent on the screen because the interaction between them is mapped in the immersive environment.

Step 1: In figure 4, the agent recognizes the learner walking towards them and can see that they are going to collide if they keep going along their current paths. It reasons that somebody will have to change direction if the goals of both parties are to be achieved.

Figure 4. Step 1: recognizing the situation

Step 2: The agent observes the learner and recognizes via eye gaze that the learner cannot see the agent. In figure 5, the learner looks down and cannot see the agent, causing the agent to change his walking direction.

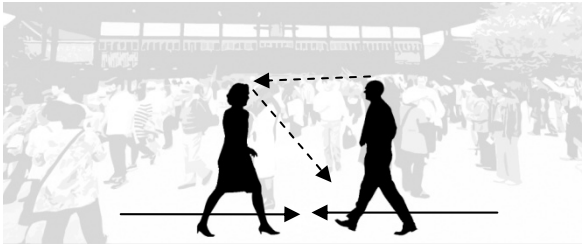


Figure 5. Step 2: Checking each other

Step 3: The agent subtly motions with their body that they are changing direction, while checking the gaze of the learner to see if they recognize their action. On the other hand, the agent may also recognize the subtle movement of the learner and use eye gaze to acknowledge their intention as the figure 6.

Figure 6. Step 3: Collision avoidance by changing motion

Step 4: Figure 7 is when the agent is sure that collision can be avoided, both parties' goals can be achieved, and the agent continues, reverting to its higher level goal.

Figure 7. Step 4: Confirmation of the safety

The above scenario involves activities that we take for granted in everyday life. Humans instinctively manage to achieve this many times in a day, through subtle non-verbal interactions and eye gaze. These gestures must be able to be identified quickly and an appropriate action performed, in order for agents to react accordingly in those situations.

Eye gaze and behavior differ across cultures using this scenario. We can use the virtual environment to create a number of similar scenarios in which eye-gaze coordinates human behavior.

Simulated crowd in immersive interaction environment

We are building a system that implements a simulated crowd using an immersive interaction environment. It permits the learner to achieve a goal by using eye gaze and multiple actions of nonverbal behavior. Our main focus is to create agents that can exhibit cultural behavior by characterizing synthetic cultural attributes.

The immersive interaction environment consists of 7 immersive displays, 4 range sensors and 3 top-down cameras. The immersive environment is shown in Figure 7. The immersive display screens the simulated crowd around the learner. There are 7 immersive displays for represent in 315 degrees. We use a regular octagon less one display for use as the immersive environment entrance. When we use the scenario for training, the learner has a 315 degree view of the scenario of the immersive crowd (Figure 8).

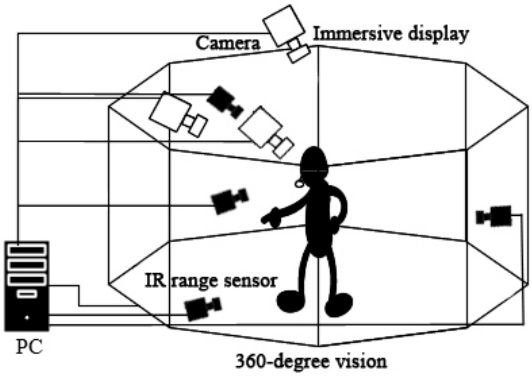


Figure 8. The immersive environment

The range sensors and top-down cameras are used for capturing learner behavior e.g. the direction of the head (as the eye gaze direction), walking speed and position of the learner in the immersive environment. We apply this behavior as an input of the nonverbal communication behavior learning system. The learner can interact with the system in 315-degree as well. Range sensors are the optimal sensing devices for our environment because they give us distance information in real time without any complicated calculation. This enables us to easily distinguish the leaner who stands in the cylindrical display from other people who are possibly shown on the display. Contact-type sensors are powerful but inadequate for our environment since they possibly prevent the leaner from acting freely. Among various contactless sensors, optical motion sensors will allow us to easily capture the motion of people. However, they cannot properly get data when the markers are hidden by something. In order to deal with this problem, we must place many cameras and try to observe the person from various points. We, however, cannot use optical motion sensors in this study because we try to capture the motion of a leaner, who is in a cylindrical display and the places for putting cameras are limited in such an environment.

Figure 9 represents the system architecture that we plan to implement for simulated crowd. The system is designed to recognize head gesture, torso, eye movement, and hand gesture and generate the behaviors of multiple synthetic characters based on the game script. A machine learning system (temporal data mining) is used to extract patterns from the log of the user-agent interactions to generate an action model of the synthetic characters. The simulated crowd has not been implemented but we have started to develop some techniques for supporting the simulated crowd learning system. Ohmoto et al implemented head gesture recognition and torso recognition [16]. This approach focused on eye movement recognition for interact ion with an agent and part of a learning system.

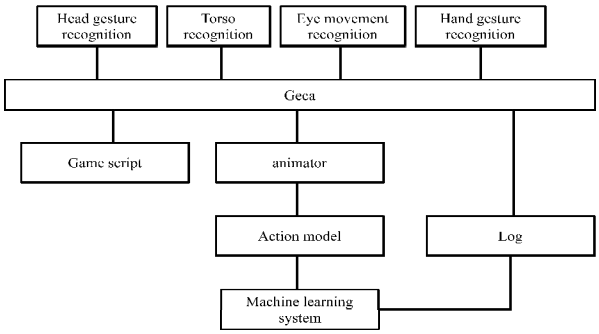


Figure 9. The system diagram

We now discuss three technical problems that we have to address in order to realize the idea mentioned above: real-time eye gaze recognition in a dynamic moving situation, sensing of nonverbal behaviors using multiple range sensors, and behavior generation based on novel temporal data mining algorithms.

REAL-TIME GAZE ESTIMATION IN MOVING SITUATION SUCH AS WALKING IN A CROWD.

When humans walk, they usually look towards the front. However, humans are not always gazing at the direction because peripheral vision is often used to recognize their surroundings, especially when they are moving in a crowd. While humans are walking in a crowd, they have to communicate with each other in following manner; they recognize whether they have eye contact, they imply their future walking directions (future action) using their eye gaze, and they give their acceptance if necessary. Mutual eye gaze occurs when human are looking at each other. Normally, mutual eye gaze helps human organize interaction [17]. In a crowd, sometimes when mutual eye gaze occurs, humans cannot predict the future walking direction of each other because they are seeing each other and cannot recognize the other eye direction which cues the future walking direction. We often avoid collisions by quickly looking in at the future direction. This point is one challenge of our research for developing the simulated crowd.

When learner walk in the simulated crowd, the system has to capture learner behaviors such as eye gaze, head gesture, hand gesture and torso movement. While the learners are walking in the simulated crowd, the main challenge is to develop a method recognizing the object which the learner is actually watching in real-time.

There are many eye gaze recognition techniques [18-21]. Some of this research uses the front of the face as the input to gaze recognition. There are many process to recognize gaze e.g. face detection, eye detection, and eye gaze analysis.

The users of many gaze direction measuring systems have to wear some devices or fix their head in a small region for measurement. These conditions prevent natural communication. On the other hand, the precise measuring of gaze direction by using image processing is a time-consuming process.

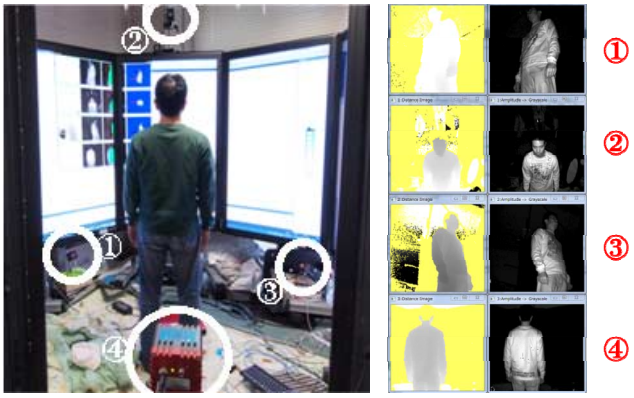
The recognition of precise gaze direction is less important than recognizing whether the agent is watching the learner or whether the learner is watching the agent in a simulated crowd. Therefore, we roughly approximate gaze direction by head direction by taking the time to integrate information such as the learner's eye movements, the agent's gaze direction and the timing of their movements.

Our prototype [16] measuring system can detect the learner's head direction and eye movements in 20 frames per second at least. The prototype system spends most of the time detecting head direction. The body motion measuring system can already detect head direction. The implementation of eye movement detection is left for future research

REAL-TIME ESTIMATION OF THE LEARNER'S 3D MOVEMENT

The estimation of leaner behavior will be calculated from the learner's 3D movement. Learner movements are captured from multi range sensors. Movement data is interpreted for the input of the learning system.

There are two main difficulties in achieving this. One is that since the learner can exhibit various behaviors in the situation, some parts of the learner's body can often be hidden by other parts, and this prevents us from accurate tracking. The other is that the tracking process usually requires a good deal of computational time, and it is difficult to develop a tracking algorithm which works in a reasonable computational time.



(a) **(b)**

Figure 10. (a) The configuration of the 4 range sensors. (b) Distance image (left) and amplitude image(right) measured by each range sensor in Figure 9. (a)

We need to realize accurate and robust sensing of the learner's motion by using multiple range sensors. We can deal with the problem of occlusion by setting multiple range sensors at mutually complementary position and integrating them. Our configuration of the 4 range sensors is shown in Figure 10 (a), and distance and amplitude images measured by each range sensor are shown in Figure 10 (b). In order to track the learner's motion, we use the 3D human body model which consists of 7 body parts; torso, head, upper arms, lower arms, and lower abdomens (see Figure 11). These parts are approximated by hemisphere cylinders or elliptic cylinders. We update the rotation matrices and translation vectors of each body parts to fit the learner's posture at each frame.

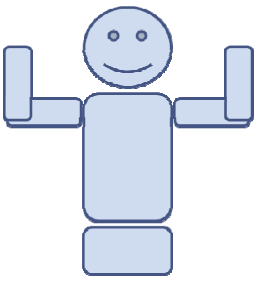


Figure 11. The 3D human body model we use.

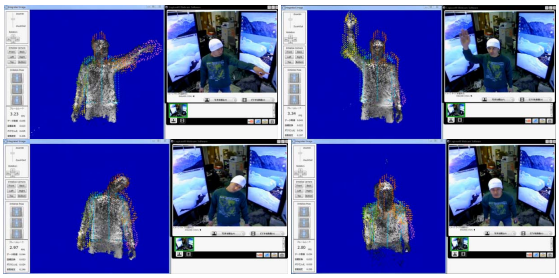


Figure 12. Some examples of recognized body parts

BEHAVIOR GENERATION BASED ON MACHINE LEARNING APPROACH

Agents have to generate appropriate actions (behavior) corresponding to the user's behavior. To generate appropriate actions, we have to model the correlation between the user's multimodal nonverbal pattern and the behavior of the agent. We call this the correlation interaction rule. The numbers of nonverbal patterns which each user uses are unknown because the behavior of users changes depending on the situation and task. Thus it is difficult to define the model prior to execution.

Mohammad and Nishida proposed a novel unsupervised learning technique [22, 23] for discovering the interaction rule from interaction data which is obtained from sensors. It allows for modeling of the user's gesture command, with the robot's action corresponding to the gesture and their association [22]. A constrained motif discovery algorithm elicits gestures and action patterns from continuous time-series data.

Okada et al proposed an incremental clustering approach: Hidden Markov Model based Self-Organizing Incremental Neural Network (HB-SOINN) [23]. HB-SOINN is a hybrid approach which integrates a self-organizing incremental neural network (SOINN), a tool for incremental clustering, and HMM which is used for modeling of time-series data. HB-SOINN has markedly improved the clustering performance over that of traditional clustering methods. However, HB-SOINN cannot be applied for continuous time-series data.

We plan to realize an incremental motif discovery algorithm by integrating CMD and HB-SOINN for discovering the interaction rule from multimodal interaction data which is obtained from sensors. Figure 13 shows the procedure of the machine learning approach.

The main idea of our approach for behavior generation is to learn the interaction model in three main steps which in we plan to realize (observing, learning and acting) the same manner as the approach which is proposed in [22]. The inputs to our system are the learner's behavior (nonverbal patterns) and the agent's action sequence; both are continuous multi-modal time series data. Each modal time-series data is multidimensional. The learning is accomplished in three main phases. First, during the discovery phase the input streams are segmented into meaningful primitive patterns and converted into a discrete integer sequence. Second during the association phase, the discrete integer sequence is analyzed to find associations and correlations between learner and agent behaviors and this information is used to build the behavior generation model. The discovery phase is done using a novel constrained motif discovery algorithm based on HB-SOINN [23].

The association phase is done using simple Bayesian network induction. The controller generation phase is done by modeling actions using Hidden Markov Models. We take into account the correlation between learner and agent behaviors. To implement the incremental multimodal motif discovery, we need a method for matching between multimodal patterns and a clustering approach for multimodal patterns. We plan to implement a new kernel for calculating the similarity between multimodal patterns.

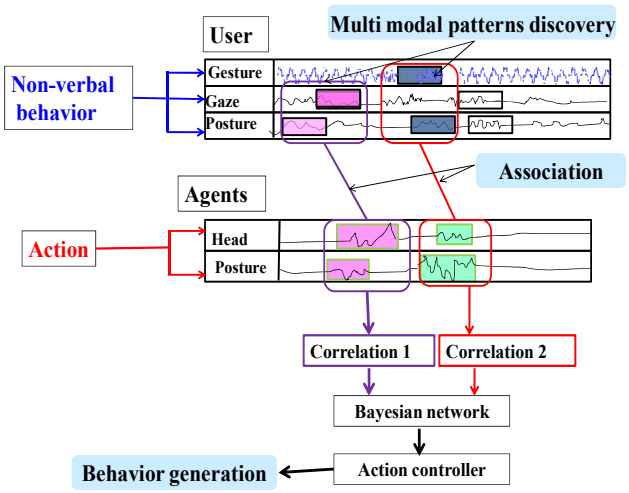


Figure 13. The procedure of proposed machine learning approach

CONCLUSION

This paper presents our research concept for developing nonverbal communication in a cultural learning system. When the learner walks through the simulated crowd, s/he will observe the agents in different culture behavior and interact with agent in real-time. The scenarios are created for helping the learner understand the culture. Eye gaze is important nonverbal behavior because humans usually use eye gaze for explanation or emotion. For eye gaze estimation we plan to analyze head direction that we have implemented as our prototype but for this approach real-time eye estimation is required. This issue is our challenge. We will use real-time estimation of the learner's movement to recognize the torso movement. Currently we have developed human body detection. The challenge of this process each identifying part of the learner body. The last technique described in this paper is behavior generation based on a machine learning approach; this technique generates an appropriate agent behavior corresponding to the learner's behavior. The learning system requires more

technique required but for the initial step we propose only the technique that we have been implementing.

REFERENCES

1. Knapp, M.L., and J.A. Hall. Nonverbal Communication in Human Interaction. Fort Worth, TX: Harcourt Brace Jovanovich College Publisher, 1992
2. Greert, H., Gert, J. H., and Michael, M., Cultures and Organizations 3rd edition, Mc Graw Hill, 2010.
3. Hofstede, G., and Pedersen, P. Synthetic Cultures: Intercultural Learning Through Simulation Games, 1999. *Simulation Gaming,* 30:415. Sage Publishing
4. LARRY, A. S., Communication Between Cultures 7th edition, Wadsworth, 2010.
5. Kleinke, C.L., Gaze and Eye Contact: A Research Review, 1986.
6. Morency, L.P., de Kok, I., and Gratch, J. Predicting listener backchannels: A probabilistic multimodal approach. Intelligent virtual agents: 8th international conference, IVA 2008 5208 (September 2008), 176–190
7. Huang, H.H., Furukawa, T., Ohashi, H., Cerekovic, A., Pandzic, I.S., Nakano, Y., and Nishida, T., How Multiple Concurrent Users React to a Quiz Agent Attentive to the Dynamics of Their Game Participation. In van der Hoek, Kaminka, Lesperance, Luck, Sen, eds.: Proceedings of 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), Toronto, Canada (May10-14, 2010), 1281–1288
8. Maia, G., Mel, S., Simon, B., and Martina, A. S., The Impact of Eye Gaze on Communication using Humanoid Avatarsz. In Proceeding CHI '01 Proceedings of the SIGCHI conference on Human factors in computing systems. (2001), 309-316.
9. Fehr, B. J., and Exline, R. V., Social visual interactions: A conceptual and literature review. In A. W. Siegman & S. Feldstein (Eds.), *Nonverbal behavior and communication* (Vol. 2nd). Hillsdale, NJ: Lawrence Erlbaum. (1987), 225-326
10. Scheflen, A.E., Body language and the social order. Englewood Cliffs, NJ: Prentice-Hall, 1972.
11. Argyle, M., Henderson, M., Bond, M., Iizuka, Y., and Contarello, A. Cross-Cultural Variations in Relationship Rules. International Journal of Psychology 21, 1 (1986), 287-315.
12. Watson, O.M., Proxemic behavior: A cross-cultural study. Mouton De Gruyter, 1970.
13. Geert, H., Culture's consequences: comparing values, behaviors, institutions, and organizations across nations 2nd revised edition, SAGE, 2001
14. Christopher, P., A Perceptually-Based Theory of Mind for Agent Interaction Initiation. International Journal of Humanoid Robotics © World Scientific Publishing Company, (June 8, 2006).
15. Klaus, D.-U., Dennis, E., Oliver, G., Nicolas, S., Volker, W., and Elisabeth, A., An Immersive Game - Augsburg Cityrun, In Perception and Interactive Technologies (PIT), ser. LNAI 4021, E. Andrè et. al., Eds., Springer-Verlag Berlin Heidelberg (2006), 201-204.
16. Ohmoto Y., Takahashi A., Ohashi H., and Nishida T., Capture and Express Behavior Environment (CEBE) for realizing enculturating human-agent interaction. Culture and Computing, Lecture Notes in Computer Science, Volume 6259/2010,(2010), 41-54
17. Argyle, M. Bodily Communication 2nd edition. London: Methuen, 1988.
18. Helton, M.P., Anna, M.G.G., and Adriao, D.D.N., Image Processing for Eye Detection and Classification of the Gaze Direction. Neural Networks, 2009. IJCNN 2009, page 2475 – 2480, 14-19 June 2009.
19. Ba, L. N., Eye Gaze Tracking. Computing and Communication Technologies, 2009. RIVF '09, page 1 – 4, 13-17 July 2009.
20. Cadavid, S., Mahoor, M.H., Messinger, D.S., and Cohn, J.F., Automated classification of gaze direction using spectral regression and support vector machine, Affective Computing and Intelligent Interaction and Workshops, 2009., page 1-6, 10-12 Sept. 2009.
21. Kristiina, J., Nishida, M., and Seiichi, Y., On Eye-gaze and Turn-taking. IUI 2010 Workshop: Eye Gaze in Intelligent Human Machine Interaction, February 7, 2010.
22. Mohammad, Y, Nishida T., and Okada, S., Unsupervised Simultaneous Learning of Gestures, Actions and their Associations for Human-Robot Interaction, IEEE International Conference on Intelligent RObots and Systems (IROS2009) 2009
23. Okada, S., and Nishida, T. Incremental clustering of gesture patterns based on Self-Organizing Incremental Neural Network. IEEE International Joint Conference on Neural Networks (IJCNN2009), (2009)
24. Lala, D., and Nishida, T., A Layered Construct-Based Model to Analyze Trust in Culture Contexts. (submitted)

Evaluation of Simulated Visual Impairment

Margarita Vinnikov
Department of Computer Science and Engineering
York University
mvinni@cse.yorku.ca

Robert S. Allison
Department of Computer Science and Engineering
York University
allison@cse.yorku.ca

ABSTRACT
We have developed two novel evaluation techniques for gaze-contingent systems that simulate visual defects. These two techniques can be used to quantify simulated visual defects in visual distortion and visual blur. Experiments demonstrated that such techniques could be useful for quantification of visual field defects to set simulation parameters. They are also useful for quantitative evaluation of simulation fidelity based on measurement of the functional relation between the intended simulated defect and psychophysical results.

Author Keywords
Gaze-contingent displays, foveation, visual simulations, evaluation of visual simulations

ACM Classification Keywords
H.5.2.e Information Interfaces and Representation (HCI): User Interfaces Evaluation/Methodology[Evaluation of gaze-based UI]

INTRODUCTION
Humans as a visual species rely on their sight in numerous everyday situations. One approach to study and modify visually-guided behaviour is through the simulation and control of the visual field. For instance, several studies have looked at visual defects associated with spatial resolution [1, 4, 3]. A few researchers have experimented with other possible visual defects such as visual distortions (metamorphopsia) or glare [11, 9, 2]. Nevertheless, what patients with visual deficits really see remains unknown. In order to close the gap between the simulation and the real experience a quantifying technique to evaluate the amount of simulated visual defect is required.

Visual simulations are most effective when combined with gaze-contingent display (GCD), in which the content and the quality of the rendered displays primarily depends on the user's eye and head position. GCD systems are usually assessed in terms of bandwidth performance, GCD latency or successful task completion. However, there are no systematic approaches that evaluate the GCD system through its primary components - the effectiveness of the system's contingency and how well the visual simulation matches a desired result. Several studies looked at different methods to evaluate the GCD systems and to qualify the property of the displayed image. For example, Loschky at al. [7] have looked at explicit measures such as subjective image quality and implicit measures such as eye movement durations. However, their approach does not quantify the actual acuity represented by their algorithm and does not separate contingent the component from the rendering component. Hence, it is impossible to attribute the experimental results to a particular component. On the other hand, if the visual field is modeled correctly then any noticeable distraction can be attributed to problems with contingency of the system. In our earlier work [10], we described a technique to evaluate the effectiveness of a system's contingency. This technique was based on localizing natural visual field features. In contrast, the present experiments look at the second aspect.

In this paper, we present new quantitative methods to verify or to model the simulated visual field for a GCD system. These techniques psychophysically quantify simulated visual disorders such as visual distortions and visual blur with the goal of validating the simulated behaviour against the expected degradation. The presented evaluation techniques are based on visual testing procedures that are usually used with real patients to examine their visual abilities.

SYSTEM DESCRIPTION
To evaluate our techniques, we have developed a real-time GCD system that simultaneously tracks the user's eye movements and head pose. The system is separated into three primary components - tracking, virtual environment rendering, and image processing. The tracking module is responsible for determining the gaze location from the head and eye coordinates that are received from the tracking devices. More details can be found in [11]. The virtual environment rendering module can render any content from 2D images to 3D scenes. For both experiments the stimuli were rear projected with a resolution of 1024 x 768 pixels at a frame rate of 120 Hz. The image processing module transforms the rendered content to reflect users's visual field and gaze position. The system can also combine several visual defects together (see [11] for more details). All image processing operations were done with Nvidia GeForce 8800 support. A chin-rest was used to stabilize the head and to maintain a

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

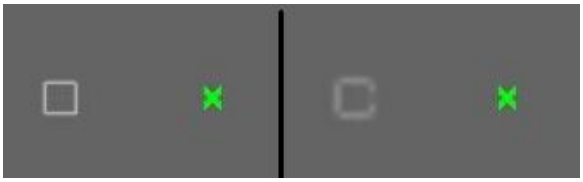


Figure 1. Stimuli examples: highest blur & gap on the right (left); no gap and no blur (right).

viewing distance of 100 cm.

PARTICIPANTS

Three female university students (average age 25) participated in the study. All participants had normal or corrected to normal visual acuity (20/20 or better) and passed a series of visual tests. Participants viewed the display monocularly. One participant tan twice viewing first with right eye and then with the left eye.

EVALUATION OF VISUAL BLUR TECHNIQUE

Visual blur is the most common visual defect or limitation that is simulated in the context of GCDs. Our evaluation technique was developed to quantify the amount of simulated blur and to compare it to visual acuity measurements. This technique capitalizes on a standard procedure that is used to examine visual acuity. We hypothesized that the participants' visual acuity would change as a function of the amount of simulated blur. Errors and imprecisions in GCD could lead to perceptual disruptions such as increase in blur in regions that result in mismatch with user's visual field. On the other hand, if the simulated blur falls below the visual threshold then it will not be perceived. Hence, a more conservative simulation can be still considered to be realistic simulation.

Visual Blur Simulation & Stimuli

The approach for simulating visual blur is a scene-based approach that used graphical hardware support, to achieve blur simulation during a single rendering pass. During the pass the system aligns visual field resolution map and the image of the scene based on the user's gaze position. The degradation level of each pixel is then determined based on an appropriate mipmap level that is encoded in visual field resolution map.

The stimuli used in this experiment was similar to the Landolt C optotype [8]. However, to simplify the stimulus, a square shape was used instead of a ring. Each side of the stimulus subtended 1°. The gap was vertical and was shown either in the middle of the left side or the right side of the square (Figures 1). In addition, three different levels of blur (highest blur (bias 5.4), intermediate blur (bias 2.4), no blur (bias 0)) were applied.

Procedure

During each trial the participant fixated a cross at the center of the screen. Then, a square with a gap on one side was shown for 250 ms at an eccentricity of 5° from the center of

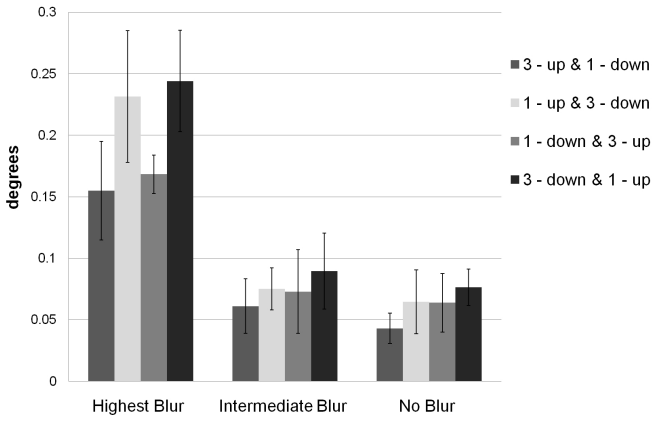


Figure 2. Averaged threshold estimates for all participants. The variability is reflected with the standard deviation bars

fixation. At the end of each trial, the participant had to indicate whether the gap was displayed on the left side of the square or on the right side. The participant's acuity was measured with a transformed Up-Down staircase procedure [6]. Whether the gap changed depended on the staircase rule for sequence (see below). To prevent guessing of the sequence and to further enhance precision, four sequential staircases ran concurrently and randomly intermixed.

Two of the staircase strategies were ascending and started at the lowest gap size (0°), while the other two were descending strategies and started at the maximum gap size (1°). For each case, two strategies were used. For three-up and one-down strategies, the gap was increased after three incorrect responses in succession, and immediately decreased after correct response. The one-up and three-down increased with a single incorrect response, while a decrease required three correct responses in a row. The three-up, one-down and one-up, three-down target the 25% and 75% correct points on the psychometric function respectively and they estimate the upper and lower bounds of the participant's blur sensitivity.

There were two sessions. In the first crude sensitivity thresholds were determined and in the second session more refined thresholds were determined. In order to determine the refined threshold, the threshold was estimated by averaging last ten peaks and valleys of each staircases [6].

Results

Figure 2 shows averaged visual acuity thresholds across participants. From the figure it is apparent that there is a significant decrease in threshold gap size from the highest blur condition to the intermediate blur condition. There was no significant difference between the intermediate and no blur conditions. The simulated ratios of the blur kernel width between different blur levels were 3.0:1.5:1 for high: intermediate: no blur cases. However, the ratio of thresholds

between no blur and intermediate blur conditions was about 1, which implies that there was no difference between the results for these two conditions.

Discussion

The experimental hypothesis was that the threshold gap size would grow as a function of amount of simulated blur. Indeed, this hypothesis was supported by the obtained results. Nevertheless, data for the intermediate blur and no blur trials were very similar. However, when directly viewed, the difference in the blur is easily perceived. The explanation for this result is the fact that, since the stimulus was shown 5° away from the center of fixation, the simulated blur was enhanced with the subject's natural acuity decline across the visual field. At 5° eccentricity the subjects natural acuity was the limiting factor. Furthermore, this can also explain why participants could detect smaller gaps on the right rather then on the left. Overall, our technique offers a way to measure perceived blur across the entire visual field and have results comparable with user's visual acuity.

EVALUATION OF SIMULATED VISUAL DISTORTION

Distortion (metamorphopsia) of the visual field is a common symptom of ocular disease such as Age-related Macular Degeneration (AMD). It is typically assessed clinically by sketches or descriptions, such as Amsler grid. However, such assessments do not quantify the degree of distortion. Therefore, our the technique is based on the ability of the participant to discriminate differences in spatial positions of two different segments as a function of imposed distortion. The procedure is similar to a vernier visual acuity test, where one has to line up a point or a segment with a second fixed point or segment as precisely as possible. Shifts in the apparent point of alignment of two objects reflects spatial distortion between them. Our hypothesis was that this technique can also be used to measure perceived visual distortion and to equate a desired distortion with a simulated value.

Visual Distortion Simulation & Stimuli

Visual distortion was achieved by using a modified version of a bump mapping shading techniques. The bump mapping shading technique perturbates each pixel according to the surface normal. The current algorithm deforms the input image at the pixel level by using a normal map as an input.

Two line segments were shown on the screen. Each line segment was 1° long and the two lines were separated with a gap of 3°. The grey background intensity was 3.5 cd/m^2 and the lines were black (0.5 cd/m^2). The lines were either drawn horizontally or vertically. The stationary line was always rendered in the same position on the screen, while the other line segment was randomly placed to the right or to the left of the first segment in the vertical case and above or below the first segment in the horizontal case (Figure 3). For each trial, different levels and direction of distortion were applied to the target area in the visual field. In total, there were 4 levels of horizontal distortions (0.573°, 0.43°, 0.286°, 0.143°), 4 levels of vertical distortions (0.573°, 0.43°, 0.286°, 0.143°) as well as one case where there distortions where applied together. There were 10 conditions x 4 initial offsets x 2 viewing options x 2 repeats. Thus, in total there were 160 trials par subject.

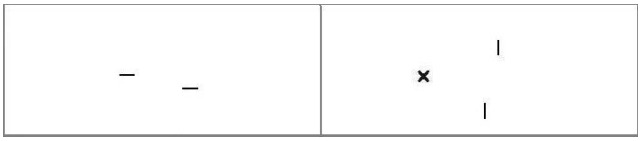


Figure 3. Possible fixation positions and stimuli locations

Procedure

During each trial two line segments were displayed. The participant had to use the arrow keys on the keyboard to align one of the line segments with the other. Once the participant was satisfied with the alignment results, they pressed a key to indicate the end of the trial. There were two types of viewing condition (Figure 3). One set of trials allowed for free viewing, while during the second set of trials, the participant fixated a location 5° away from the two segments. The latter is applicable for GCD evaluation where the segments would be presented in fixed location in the visual field to evaluate distortion of the visual field. We did not present in a GCD manner in order to evaluate the technique without dependence on particulars of a given GCD implementation. However, we conducted an additional experiment, where participant's fixation was tracked. The results were consistent with the main experiment and thus are not presented here.

Results

The alignment settings for both free viewing trials and fixation trials averaged across participants are presented in Figure 4. For free viewing, the alignment adjustment shifted with imposed distortion as predicted when the distortion was in the direction of the alignment task (e.g. horizontal distortion with vertical lines, which need to be horizontally shifted for alignment). On another hand, if the alignment task was perpendicular to the direction of distortion then alignment errors were small, on average 0.04°. This was expected since the distortion should have no effect on the alignment of the line segments (the gap between the lines changes but not their alignment). When the psychophysical results are compared with the predicted results, one can see that the observed values are closer to the predicted results for horizontal alignment and slightly underestimated in the vertical alignment trials. In the fixation tasks, it can be observed that participants had more variability and tended to overshoot or undershoot in their matches. However, these errors were not systematic and on average the data were similar to those in the free viewing tasks.

Discussion

The hypothesis that the bias exhibited in the alignment task would be directly related to the amount of imposed visual distortion was supported by the experimental results from the free viewing trials. A similar pattern was observed in trials with fixations. The task was very precise under free viewing, but it was much more variable with fixation. This

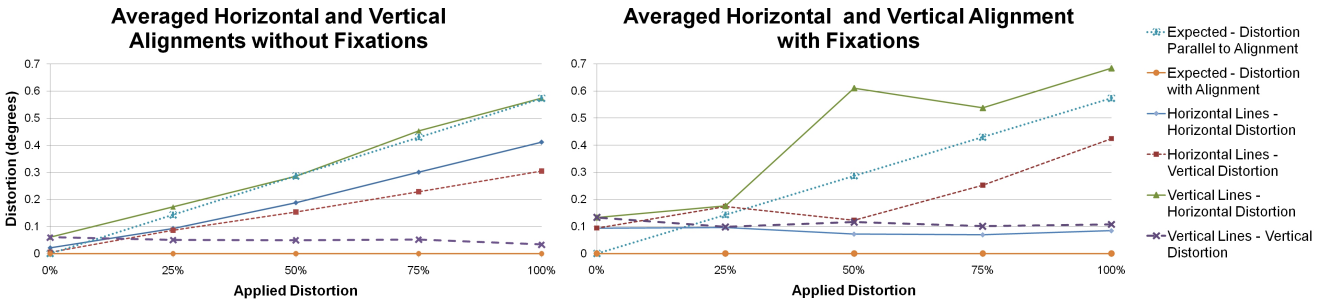


Figure 4. Averaged horizontal and vertical estimations

can be explained by the fact that in addition to the simulated visual defect, the participant's performance was affected by the natural degradation with eccentricity, which is even more pronounced in vernier acuity tasks. Our results are consistent with previous findings [5] that have reported that there is a decline in precision with increasing eccentricity. The ratios between participants' variability in performance between free viewed tasks and fixation tasks was 3.36 for vertical alignment and 3.7 for horizontal alignment. This is an indication that the performance accuracy was less affected by the eccentricity but rather by the stimulus manipulation. Nevertheless, our results show the importance of the area of fixation on the ability to perform a fine detail task, such as alignment. We found a close match between the predicted and obtained alignments. Vernier acuity tasks are precise and repeatable [5]. Thus these tasks are well suited to the measurement of perceived distortion and to the validation of simulated distortion in gaze contingent displays, both in the central visual field and in the near periphery. In the far periphery precision will be an issue and analogous technique based on apparent motion might be preferable.

CONCLUSIONS & FUTURE WORK

The results from the experiments confirmed a strong functional relation between the intended simulation distortion and psychophysical results. These types of evaluation techniques are necessary for ensuring realistic non-subjective clinical simulations when the degree of imposed visual defect is important. As well they provide a methodology to compare the desired visual field to the achieved one that the user is actually perceiving. Furthermore, incorporating these techniques into evaluation practices allows an easy comparison between different systems and algorithms. In the future, we intend to work on similar techniques that can be used to quantify and evaluate other visual defects, so that it will be possible to model a wide range of visual disorders.

REFERENCES

1. Z. Ai, B. K. Gupta, M. Rasmussen, Y. J. Lin, F. Dech, W. Panko, and J. C. Silverstein. Simulation of eye diseases in a virtual environment. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, page 5, 2000.

2. A. T. Duchowski and T. D. Eaddy. A gaze-contingent display compensating for scotoma. *EUROGRAPHICS 2009*, 2009.

3. F. C. Fortenbaugh, J. C. Hicks, L. Hao, and K. A. Turano. A technique for simulating visual field losses in virtual environments to study human navigation. *Behavior Research Methods*, 39(3):552, 2007.

4. W. S. Geisler and H. L. Webb. A foveated imaging system to reduced transmission bandwidth of video images from remote camera system. Technical Report 19990025482, AD-A358811, AFRL-SR-BL-TR-98-0858, NASA, Austin TX, 1998.

5. D. M. Levi, S. A. Klein, and A. P. Aitsebaomo. Vernier acuity, crowding and cortical magnification. *Vision Research*, 25(7):963–77, 1985.

6. H. Levitt. Transformed Up-Down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, 49:467– 447, 1971.

7. L. C. Loschky, G. W. McConkie, J. Yang, and M. E. Miller. Perceptual effects of a gaze-contingent multi-resolution display based on a model of visual sensitivity. In *the ARL Federated Laboratory 5th Annual Symposium-ADID Consortium Proceedings*, page 5358, 2001.

8. J. H. Prince and G. A. Fry. Correction for the guessing bias in the landolt ring test. *Am. J. Ophrhal.*, 46:77–85, 1958.

9. Y. Saito, Y. Hirata, A. Hayashi, T. Fujikado, M. Ohji, and Y. Tano. The visual performance and metamorphopsia of patients with macular holes. *Archives of Ophthalmology*, 118(1):41, 2000.

10. M. Vinnikov and R. S. Allison. Contingency evaluation of gaze-contingent displays for real-time visual field simulations. In *Proceedings of the 2010 symposium on Eye Tracking Research & Applications (ETRA)*, pages 263–266, 2010.

11. M. Vinnikov, R. S. Allison, and D. Swierad. Real-time simulation of visual defects with gaze-contingent display. In *Proceedings of the 2008 symposium on Eye Tracking Research & Applications (ETRA)*, pages 127–130, Savannah, Georgia, 2008. ACM.

Investigations of the Role of Gaze in Mixed-Reality Personal Computing

Thomas Pederson, Dan Witzner Hansen, and Diako Mardanbegi
IT University of Copenhagen
Rued Langgaards Vej 7
2300 Copenhagen, Denmark
{tped, witzner, dima}@itu.dk

ABSTRACT
This short paper constitutes our first investigation of how eye tracking and gaze estimation can help create better mixed-reality personal computing systems involving both physical (real world) and virtual (digital) objects. The role of gaze is discussed in the light of the situative space model (SSM) which determines the set of objects a given human agent can perceive, and act on, in any given moment in time. As a result, we propose to extend the SSM in order to better incorporate the role of gaze, and for taking advantage of emerging mobile eye tracking technology.

Author Keywords
Interaction paradigm, gaze tracking.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
The design of interactive systems that involve more than one computer device and also a range of everyday physical objects, demands us to extend the classical user-centered approach in HCI [3]. One challenge is that both system and human needs to continuously establish an understanding of what parts of the physical and virtual worlds that currently make up the "user interface" as devices and interaction modalities change with context. The egocentric interaction paradigm [5] proposes a change in view of a) the role of digital interactive devices in relation to the information they provide access to, and b) to generalize the HCI input/output concept to make room for multiple parallell interaction channels as well as interaction with objects in the real world (physical objects).

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

Virtual Objects and Mediators Instead of Interactive Devices
Input and output devices embedded in digital appliances are viewed as *mediators* through which virtual objects are accessed. Virtual objects are assumed to be dynamically assigned to mediators by an *interaction manager* software component residing on body-worn hardware. The purpose and function of mediators is that of expanding the *action space* and *perception space* of a human agent (Fig. 2).

Action and Perception Instead of Input and Output
In the egocentric interaction paradigm, the modeled human individual is an agent moving about in a mixed-reality environment, not a "user" interacting with a computer. Also the HCI concepts input and output are reconsidered: (device) "input" and "output" are replaced with (human agent) "action" and "perception". Note that object manipulation and perception are processes that can take place in any modality: tactile, visual, aural, etc. In this paper, we focus on *visual* modalities for perception and action.

HUMAN ACTIVITY AND GAZE
Eye movements are versatile and play an important role in everyday activities [2]. It is well known that human eye movements are governed by our interests and intentions [6], and

humans tend to look at the object that they want to act on prior to any motor control. The sequences of fixations, trackable by emerging mobile tracking technology [1] in some cases provide enough data for making predictions [2].

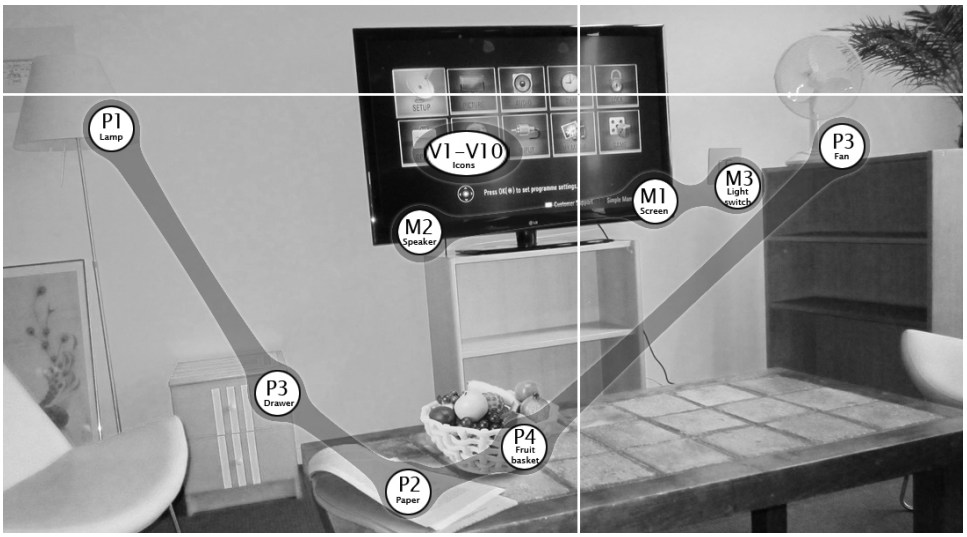


Fig. 1. A living room environment as seen by a human agent. Some physical objects (P1-P5), virtual objects (V1-V10) and mediators (M1 and M2) are labelled for illustrative purposes. The gaze direction of the human agent is indicated by the hair cross.

A SITUATIVE SPACE MODEL

The situative space model (SSM) [4] is intended to model what a specific human agent can perceive, reach and operate, at any given moment in time. This model is intended to be the emerging egocentric interaction paradigm equivalent of what the virtual desktop is for the PC/WIMP (Window, Icon, Menu, Pointing device) interaction paradigm: more or less everything of interest to a specific human agent is assumed to, and supposed to, happen here. Fig. 1. shows a typical situation which the SSM is intended to formalise and capture: a living room environment inhabited by a human agent.

In the following, we will discuss the role of gaze in the light of SSM definition excerpts from [5].

Perception Space (PS)

The part of the space around the agent that can be perceived at each moment. Like all the spaces and sets defined below, it is agent-centered, varying continuously with the agent‘s movements of body and body parts. Different senses have differently shaped PS, with different operating requirements, range, and spatial and directional resolution with regard to the perceived sources of the sense data. Compare vision and hearing, e.g.

Within PS, an object may be too far away to be possible to recognize and identify. As the agent and the object come closer to each other (either by object movement, agent movement, or both) the agent will be able to identify it as X, where X is a certain *type* of object, or possibly a unique individual. For each type X, the predicate “perceptible-as-X” will cut out a sector of PS, the distance to the farthest part of which will be called *recognition distance*. [5]

Naturally, gaze direction plays a fundamental role in defining the visual PS for a given human agent. Any object directly hit by the vector anchored in the fovea and passing through the center of the lense of an eye (that is, the line of sight, LoS) is a top candidate member of the PS since it is only along this vector human agents literally see clearly. However, other components of the human visual perception system “expands” this single vector of visual impression so that visual attention in practice typically is directed to a larger area than just a point in 3D space. Let us call this 2-dimensional expanded area – with the LoS hitting its center – the field of view (FoV). Then, very simplified, the 3D space created by the union of the two eye’s FoV, let us call it the 3DFoV, forms the basis for the visual PS (again, with the help of complementary parts of the human perception system, dealing with angular calculations and objects obstructing each other, etc.). All objects in the 3DFoV (not just the object in LoS) should be included in PS for a given agent.

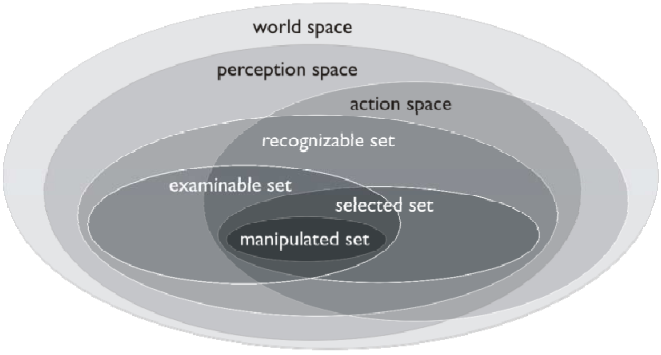


Fig. 2. A Situative Space Model. [4]

Recognizable Set (RS)
The set of objects currently within PS that are within their recognition distances.

The kind of object types we are particularly interested in here are object types that can be directly associated with activities of the agent – ongoing activities, and activities potentially interesting to start up – which is related to what in folk-taxonomy studies is known as the basic level.

To perceive the status of a designed object with regard to its relevant (perceivable) states (operations and functions as defined by the designer of the artifact) it will often have to be closer to the agent than its recognition distance: the outer limit will be called *examination distance*. [5]

Examinable Set (ES)
The set of objects currently within PS that are within examination distances. [5]

The visual RS and ES in the SSM (motivated by the potential value for an egocentric interaction system to know in what detail objects can be analysed by a human agent) raises gaze tracking questions. Can gaze estimation be used for determining whether an object is examinable, recognizable or just perceivable? Eye movement pattern categorization over time and object types could, potentially, help determining whether a visually perceivable object belongs to RS or ES.

Action Space (AS)
The part of the space around the agent that is currently accessible to the agent's physical actions. Objects within this space can be directly acted on. The outer range limit is less dependent on object type than PS, RS and ES, and is basically determined by the physical reach of the agent, but obviously depends qualitatively also on the type of action and the physical properties of objects involved; e.g., an object may be too heavy to handle with outstretched arms. Since many actions require perception to be efficient or even effective at all, AS is qualitatively affected also by the current shape of PS.

From the point of view of what can be relatively easily automatically tracked on a finer time scale, it will be useful to introduce a couple of narrowly focused and highly dynamic sets within AS (real and mediated). [5]

The visual AS is limited: Few actions that change the state of physical or virtual objects can be performed using eyes alone. However, gaze activity is often part of actions executed using other parts of the body such as the hands.

Selected Set (SdS)
The set of objects currently being physically or virtually handled (touched, gripped; or selected in the virtual sense) by the agent.

Physical selection is almost always preceded by visual selection: before grabbing anything, we visually fixate the object. Without dwelling into the reasons, this fact means that by tracking gaze, computer systems can do heuristical guesses for what object, among all the objects in AS, that is about to get manipulated next.

Manipulated Set (MdS)
The set of objects whose states (external as well as internal) are currently in the process of being changed by the agent. [5]

All these spaces and sets, with the obvious exception of the SdS and the MdS, primarily provide data on what is *potentially* involved in the agent's current activities. Cf. the virtual desktop in the PC/WIMP interaction paradigm.

Like object selection, also object manipulation can involve gaze. While visual feedback is crucial for certain kinds of physical object manipulation (e.g. hand writing), it is probably less important for most. For manipulation of virtual objects, the situation is different. One of the most prevailing critisisms of today's user interfaces is in fact the heavy reliance on visual feedback. Contrary to actions in the real world, most user interfaces *rely* on continuous visual attention also during object manipulation.

EXAMPLE SITUATION
Fig. 1. shows a living room environment. If we assume that the area covered by the photo approximately corresponds to the field of view of a given human agent, objects in the photo can be categorized using the SSM as follows:

Physical objects
The physical object P1 (the floor lamp) belongs to the examinable set since the human agent can determine whether the lamp is on or off. The paper document P2 is not in the examinable set because from this position, the human agent can not likely determine what the document is about, see what page that is on top, let alone read the text of it. P2 is however in the recognizable set because it is indeed clear that the object is a paper document. The drawer P3 belongs to the examinable set because it is possible to see whether it is open or closed. The fruit basket P4 is examinable: it is possible to determine whether it is empty or full and even the kind of fruit that it contains. The desk fan P5 is also

examinable – it is possible to see whether its rotor blades are turning or if they are still.

Mediators

The TV embeds two mediators: The screen (M1) and the speaker (M2). The screen M1 is in the examinable set since the human agent can determine what is shown on it, i.e. the virtual objects that it currently mediates. The TV speaker M2 is not in the visual perception space at all since the case design of the TV hides its presence. (It is true that it is in the aural perception space – virtual objects can be sufficiently sonified from this distance – but we limit our analysis to the visual perception space.) The light switch M3 is in the perception space but not examinable: the human agent cannot determine its state from this distance.

Virtual objects

The icons shown on the screen M1, modeled as virtual objects V1-V10, are all examinable because their state (selected/not selected) can be determined from the position of the h. agent.

Action space

With respect to action space, most of the objects labelled in Fig. 1. are outside of that space. The human agent cannot, from her/his current positon manipulate them. The exception might be the paper document P2 or the fruit basket P4 which might be just about reachable. If we imagine the human agent to hold the TV remote control in her/his hands (a physical object embedding mediator buttons) however, also the 10 icons V1-V10 enter action space since that would allow her/him to manipulate them.

The hair cross in the picture simulates the gaze direction of the human agent, currently examining one of the 10 icons on the TV.

CONCLUSION

In this paper we have taken our initial steps in modeling gaze within the situative space model (SSM). Gaze turns out to be a defining factor for to which space an object belongs, potentially altering an object's location within the model rapidly. To fully exploit the information in eye and gaze movements, the SSM might benefit from the incorporation of something like an "attended-to" set of objects (Fig. 3.), including objects across several existing SSM spaces and sets that the given human agent is attending to. Among many open issues related to gaze and human attention is that a person may attend to objects that they can see but not recognize. At the same time, an object may be recognizable but not really attended to.

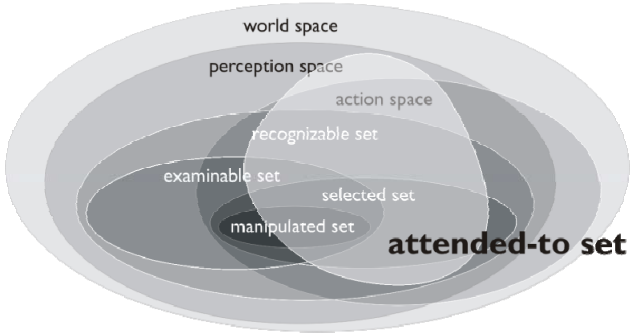


Fig. 3. Future work: extending the situative space model with an "attended-to" set.

REFERENCES

1. Hansen, D. W., Ji, Q., In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 3*, 2010,478–500.
2. Land M.F., Tatler B.W., *Looking and Acting: Vision and eye movements in natural behaviou*r. Oxford; New York: Oxford University Press, 2009.
3. Norman, D. & Draper, S. (Eds.) *User centered system design*. Erlbaum, Hillsdale, NJ, 1986.
4. Pederson, T., Janlert, L-E., Surie, D., Setting the Stage for Mobile Mixed-Reality Computing - A Situative Space Model based on Human Perception. *IEEE Pervasive Computing Magazine* (to appear), 2011.
5. Pederson, T., Janlert, L-E., Surie, D., Towards a Model for Egocentric Interaction with Physical and Virtual Objects. *Proceedings of NordiCHI'10*, ACM Press, 2010, 755-758.
6. Yarbus, A. L., *Eye Movements and Vision*. New York: Plenum Press, 1967.

Emotional Text Tagging

Farida Ismail
German University in Cairo, Egypt
farida.ismail@gmail.com

Ralf Biedert, Andreas Dengel
German Research Center For Artificial Intelligence
firstname.lastname@dfki.de

Georg Buscher
Microsoft
One Microsoft Way
Redmond, WA 98052, USA
georg@gbuscher.com

ABSTRACT
We created and evaluated a system capable of observing the reader's emotions and tag the perused text. By using either a web camera or an Emotiv neuroheadset displayed emotions like happiness, interest, boredom and doubt can be recorded. At the same time an eye tracker analyzes the reader's progress. According to the reader's current reading position in the text and the displayed emotions, the text part is automatically tagged with the reader's emotional state. The reading-interface is able to facilitate the emotional information in realtime, the user can also access the recorded eye tracking sessions and perceived emotions later on. We evaluated the system's ability to accurately tag emotions, conclude that joy is detected best, boredom is barely recognizable, and highlight some key issues we encountered.

Author Keywords
Eye Tracking, Emotions, EEG, Reading

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Interaction styles - Gaze*; H.5.2 Information Interfaces and Presentation: User Interfaces—*Interaction styles - EEG*; H.5.4 Information Interfaces and Presentation: Hypertext/ Hypermedia

INTRODUCTION
Some texts are boring, some make us laugh, and some texts raise our interest. Unfortunately, almost all of these evoked emotions are lost. They are not recorded, let alone stored or searched for, except through user's manual interaction (rename *document* to *interesting document*). Even worse, labels given on a document level are crude and reduce the document's content to at most a few tags.

Also there is a number of web services like `slashdot.org` or `dailyme.com` that enable users to rate comments according to their evoked emotions. Currently, however, these services require manual interaction as well and, again, reduce the evoked emotions to a single statement.

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

Figure 1. The experimental setup. The user sits in front of an eye tracking device while reading text. A web cam or neuroheadset track displayed emotions which are then annotated in the text.

In contrast to the websites and methods mentioned, we assume that the emotions experienced and displayed during reading are diverse. We acknowledge that one part of a text can be funny while another part of the same text may be sad, and thus we try to investigate what it would need to record and store these emotions in their whole complexity without any manual interaction.

Very little research has been conducted in the field of assigning emotions to text in real time. To date, a number of systems have been considering characteristic eye behaviors to recognize emotions while reading in real-time using an eye tracking device.They exploit variations in pupil size, blink rate and saccade length to identify the user's emotional state such as increased workload. Tiredness and attention are other emotions that can be detected by referring to changes in eye parameters [5]. However, emotions relying heavily on facial expressions such as joy and doubt cannot be detected by an eye tracker.

The section *Detecting Emotions* describes our methods used (also see Figure 1), in the section *From Emotions to Text* we report how we merged the detected emotions with eye tracking and in the *Evaluation*-part we report and discuss our findings. The paper concludes with an outlook on future work.

DETECTING EMOTIONS
The definition of emotions is rather complex. While some definitions are based on affective behaviors, others analyze cognitive and physiological reactions and again others combine all approaches and view emotions as an interaction between the mentioned factors[3]. We focus on a definition

which considers what is externally *perceivable* and classify emotions according to the typically visible facial expressions. Thus, our working hypothesis is the existence of four commonly observable, mostly mutually exclusive emotions during reading: Joy, boredom, doubt and interest. These are chosen out of the belief that they are likewise of significance in human-document-interaction as well as that there is a reasonable chance for computer-aided recognition. With that in mind we investigate two methods to detect them.

First, we examine the use of a web cam mounted on the computer screen to record the reader's facial expressions while reading, see Figure 2. Every time a new image is captured, facial features like eyes and mouth are detected using Haar Cascade Classifiers in OpenCV. Eye and mouth regions are considered in particular because they characterize facial expressions uniquely. The located regions are extracted from the image and compared against an already classified training set of facial images by applying a trained SVM. The emotion of the closest match is then returned.

In our second approach, we replace the webcam by an Emotiv[1] neuroheadset. It measures facial expressions and fine muscular pulses to provide information about the subject's current emotional state. This muscular information can be combined with its tracking capability of brainwave signals, allowing for a detection of states even if they are not visible. We build upon the device's Emotiv API, which delivers processed EEG data in the form of different channels, each representing a subjective state or facial expression. We store all the brain specific data, and classify the emotions when needed, after taking the average values of each collected signal. The signals we consider for emotion classification are engagement (reflecting boredom and interest), furrowing (doubt), smiling and laughing (joy). We compare these values to the emotion characteristic values we collected in a calibration run and return the most likely emotion match. In case the measurements satisfy no known emotional pattern, a *neutral* state is returned, signifying that neither of the emotions is present.

Both methods have their strengths and weaknesses. While the webcam depends on visible factors, the neuroheadset has more of an *insight* into the human brain, providing more sophisticated information. During initial experiments we observed that, in contrast to the emotional expressiveness of human-to-human interaction (e.g., people laughing loudly about a joke told), the expressiveness when a reader interacts with text yields rather low levels of emotion (e.g., people may just show a faint smile if reading a joke). In that respect the neuroheadset is more sensitive since it detects also slight facial muscular movements.

The webcam on the other hand requires the emotion to be expressed in a high level of intensity in order to be detected in a satisfactory manner as shown in [2] with a low-cost webcam and in [4] with more sophisticated algorithms and apparatus. Also, the webcam does not allow any obstacle (e.g. a hand partially obscuring the reader's face) for the face to be analyzed, while the neuroheadset depends only on the direct connectivity between sensor and scalp to provide accurate brainwave data. However, the neuroheadset requires a more lengthy initial preparation, the sensors (and therefore the reader's scalp) have to be dampened and due to its applied pressure for sensor connectivity some users find it uncomfortable to wear after some time.

[1] `http://www.emotiv.com/`

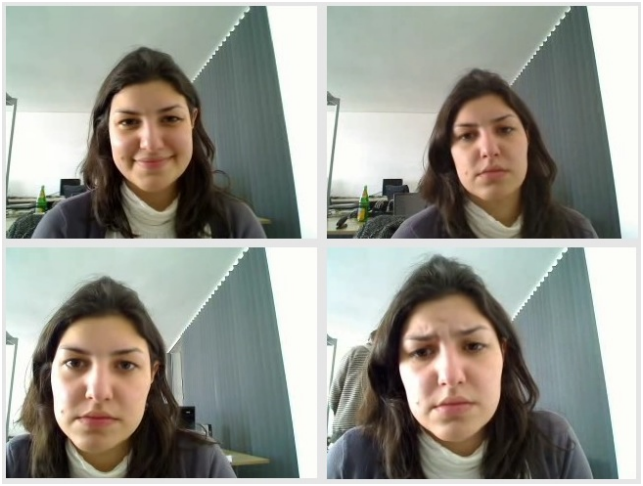


Figure 2. Samples captures of four emotions displayed during explicit web cam training. The four emotions are, clockwise starting from the upper left, joy, boredom, doubt and interest. It can be seen that boredom and interest are visually almost indistinguishable. In this respect EEG devices are likely to be of advantage as they are capable of detecting more subtle muscular activity as well as capable of evaluating brain waves.

Still, for the final prototype application, we preferred the EEG solution, as it provided more detailed emotional information and insight, specifically when emotions are not displayed visibly such as when reading.

FROM EMOTIONS TO TEXT

The rest of the setup consists of an eye tracker, a web browser and one of the emotion detection methods described above, compare Figure 1. The user opens an HTML document she wants to read in the browser, it is then loaded and segmented using the Text 2.0 framework[1]. After the page is fully initialized, evaluation of the gaze data and measured emotions begins.

The framework delivers fixation events to the document level and we use these to follow the reader's gaze over the text. By splitting the text into a set of span-elements, each containing a single word (compare [1] for details), raw fixation and emotion data can be assigned on word-level. This granularity allows for a detailed insight into the reader's emotional reading experience and it serves as a foundation to deduce the prevalent emotion on the paragraph or article level.

Each span element is augmented with a JavaScript `onGazeOut` event handler such that, whenever the reader's gaze leaves a word, a callback on that word is triggered. Next, the callback queries the present emotion and the currently focussed word is then tagged with an additional attribute `emotion`. The value assigned to it represents the emotional state that was evoked at this particular document position. At the same

time the emotion values, along with the gaze data and the pages's structure, are written into a session log for later processing and retrieval.

In order to make sure that words are only augmented with emotions if they are actively read within their context and not only skimmed or unintentionally looked at, the global reading behavior needs to be kept track of as well. As long as the distance between the words read is within a certain bound the emotional tagging is active, otherwise the tagging stops until a *consistent* reading behavior is detected again. For the purpose of this study this was defined as reading any three of seven consecutive words in an incremental order, thus, from *left to right*.

If the reader moves his eyes back over previously considered words, i.e., when performing a regression, possibly different emotions are evoked than those detected during the first pass. In this case we do not overwrite the emotional information in order to keep the initial reading experience.

Also, while reading, not every single word in a sentence triggers an `onGazeOut` event, either due to normal reading behavior or eye tracking inaccuracies, leading to untagged words within a read sentence. So in case words are perceived but no fixations fall upon them, all the emotions evoked by those words will be assigned to the next element gazed upon. To avoid gaps between two fixations upon later retrieval and visualization emotions are also spread backwards. This means that on each new emotion assignment we proceed through the read text from back to front, and everytime an untagged word is encountered, it is assigned the emotion of the last word considered.

Interaction

In addition to the process of recording emotions we also investigated how they could be facilitated in real time. In this respect we followed the paradigm we already implemented in the Text 2.0 framework: gaze active handlers. In an initial approach we defined a set of attributes: `onSmile`, `onFurrow`, `onBoredom` and `onInterest`. These attributes can be used in parts of the web site's DOM tree. If then the eye tracker detects that the user's gaze is within the element's screen position and the emotion detection finds the corresponding emotions above a certain threshold the code within the handler is being executed. In this way, web designers can define event handlers for elements that are triggered if the user shows a specific emotion while looking at an element.

EVALUATION

We evaluated the overall system performance with respect to its overall accuracy of the tagged text. Based on a number of initial test runs comparing both emotion detectors we decided to use the EEG method throughout the experiment, as it proved to be less sensitive to head movement or rotation and provided more detailed insights in terms of sensor data.

The participants were seated in front of a desktop mounted Tobii X120 eye tracker and wore an Emotiv neuroheadset for the EEG emotion detection method as shown . Their task was to read five different articles in a browser. The articles were, for example, about bizarre scientific news, lengthy political discussions and funny jokes. Others were articles and threads retreived from `slashdot.com` and `dailyme.com` which were already classified by their users and we used this classification as a pre-selection to evenly distribute documents falling into different emotional classes. Upon completing each task, our users were then asked to mark the interesting, boring, doubtful, and joyful parts on the screen. Afterwards, the algorithmically computed emotions were compared against this feedback. Figure 3 shows a piece of text after tagging and coloring words according to the emotions evoked.

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

Table 1. The system performance with respect to the four emotions of interest. Joy and doubt are significantly better to detect than boredom.

In total, nine undergraduate students participated in the experiment and gave feedback on five articles. We considered 32 tagged texts for evaluation, 13 had to be discarded due to missing or low quality eye tracking or emotional tagging data. 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 adjoining 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 1. While the rather expressive emotions joy and doubt were often detected when they occurred, boredom was almost imperceivable. 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. Based on the participants's oral feedback, we also found that the definitions of interest and boredom when related to text are ambiguous: Three participants mentioned that to them the opposite of an interesting reading experience was the neutral emotional state instead of boredom. And two participants defined an interesting text to be any text that handled a favoured topic, without considering their actual reading be-

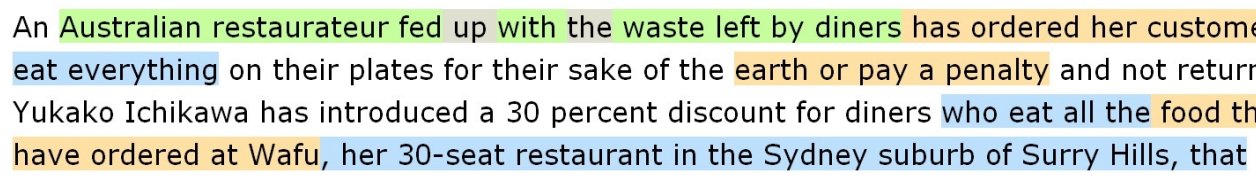


Figure 3. Sample image displaying a tagged text fragment and a number of issues. The participant read the text inside a browser while the emotions were recorded. After the reader finished, she was asked about her judgment with respect to her *real* emotions and the text was algorithmically colored. Blue denotes *interest*, orange stands for *joy*, green *doubt* and gray equals *neutral*. For the white parts no reading behavior was detected because the reader was laughing intensely which resulted in closed eyes and heavy body movement.

havior while going through the article. Thus, the evaluation results for these two emotions have to be considered as being rather coarse. It should also be noted that, due to the granularity applied, the reported number for precision and recall are likely higher than the device is actually capable of achieving in a frame-per-frame analysis of recorded emotions.

DISCUSSION

Assessing implicit emotions during reading is delicate. They are not as expressive as during human to human interaction, rendering them hard to detect. Likewise is their calibration and eventual evaluation, and many open questions remain.

Is text *interesting*, as one of our participants reported, when it adds new information about a topic which is of one's concern (while the text itself may be merely practical)? Or is it to be considered interesting only when one reads with a certain level of engagement and feeling of suspense? Does the *neutral* emotional state while reading exist, or is an article either interesting or boring? Should such emotions be defined subjectively and, if not, what would be a feasible set of emotions in terms of detectability and (semantical) expressiveness? How do we deal with emotions that take time to rise or decline and what are the sensible temporal and spacial limits of assigning them to text?

We also learned some lessons during our experiments. First of all, training and recording of emotions should happen implicitly. Recording explicitly evoked emotions might result in well divisible data sets, however, actual observable reactions show much less intensity than those displayed explicitly and thus, can not be categorized based on these. Also the generation of ground truth by having participants manually annotate their emotions, even after reading only a single document, proved to be unreliable as people often could not remember all evoked emotions anymore or were not sure about the specific place of evocation.

CONCLUSION & OUTLOOK

We presented a framework and a case study for a system that is capable of recording the emotional state of a user while reading text, and automatically assigns these detected emotions to the text. The evoked emotions and reading information can be used in real time or stored in a database for later retrieval.

Our results lead us to the conclusion that, given the emotion detection methods we employed, an algorithmic evaluation of emotions should happen well above word level. On this level a strict distinction into mutually exclusive emotions will also not hold anymore and should rather be expressed in terms of relative tendencies.

Another area in need of improvement are heuristics to deal with missing or inaccurate eye tracking data. For example, when laughing readers sometimes close their eyes or shake their bodies. In these cases neither can the eye tracker detect the readers eye and deliver fixation information, nor is a consistent reading behavior observed, thus resulting in missing emotional information about the text.

For the future we anticipate a number of applications where the emotional tagging system could be integrated to enhance the user experience by emotionally interacting with text. Possible applications include searching for text parts or articles that evoked certain emotions either by the reader himself while reading an own document or emotions which were displayed by other readers as a form of emotional feedback. Authors could also make use of the automatic real-time tagging and retrieve information about how readers respond to writings and help them analyze texts emotionally. Besides the emotional tagging, applications could actively interact with the user and react to the emotions displayed by using the emotional event listeners introduced to the Text 2.0 framework[1]. Finally, the tagging system can be easily expanded such that it can be applied on non-textual elements such as images and videos as well.

REFERENCES

1. R. Biedert, G. Buscher, S. Schwarz, M. Moeller, A. Dengel, and T. Lottermann. The Text 2.0 Framework. Presented at International IUI 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction, 2010.
2. E. Cerezo, I. Hupont, C. Manresa-Yee, X. Varona, S. Baldassarri, F. J. P. Lpez, and F. J. Sern. Real-Time Facial Expression Recognition for Natural Interaction. In J. Mart, J.-M. Bened, A. M. Mendona, and J. Serrat, editors, *IbPRIA (2)*, volume 4478 of *Lecture Notes in Computer Science*, pages 40–47. Springer, 2007.
3. P. R. Kleinginna and A. M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, December 1981.
4. M. Pantic and L. J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1424–1445, 2000.
5. M. Porta. Implementing Eye-Based User-Aware E-Learning. In *CHI 2008 Proceedings*, pages 3087–3092, 2008.

The eyePad - Tom Riddle in the 21st Century

Mostafa El Hosseiny
German University in Cairo, Egypt
es.mostafa@gmail.com

Ralf Biedert, Andreas Dengel
German Research Center for Artificial Intelligence
Germany
fistname.lastname@dfki.de

Georg Buscher
Microsoft
One Microsoft Way
Redmond, WA 98052, USA
georg@gbuscher.com

ABSTRACT
We created a multimodal book reader combining eye tracking, handwriting and speech I/O in a novel storytelling concept. We present a number of scenarios integrated in an ad-hoc story to demonstrate new human-text interaction techniques and reading assists, and report on our user study conducted to evaluate the prototype in the real world. Our results show that the new reading assists were invariably reported as being helpful and entertaining.

Author Keywords
Multimodal interfaces, Eye tracking, Handwriting, Speech I/O, Reading, Storytelling

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Input devices and strategies*; J.5 Computer Applications: Arts and Humanities—*Literature*

General Terms
Human factors, Languages, Experimentation

INTRODUCTION
Considering the ongoing miniaturization of eye tracking devices we assume that their large scale integration into tablets and eReaders is a realistic possibility. These devices usually also include touch screens with the capability of handwriting recognition and speech input and output facilities. Given this scenario we want to explore how a user's interaction with such a future reading device can be enhanced in a natural manner, fusing several input modalities to form new reading assists on the one hand, and to provide new authoring capabilities on the other.

Built on top of our idea of the *eyeBook*[2] and various other prototypes[1, 5, 6, 7], it raised our special interest how new ideas of gaze aware interactive reading can be constructed and facilitated. We present a prototype implementing a number of these ideas. It borrows its main ideas from Tom Riddle's Diary first described in the novel *Harry Potter and the Chamber of Secrets* by J. K. Rowling. Tom Riddle's Diary was a blank journal which Tom Riddle transformed into a magical object. The diary allowed a writer to communicate with the memory of a younger Tom Riddle merely through writing on the journal's blank pages.

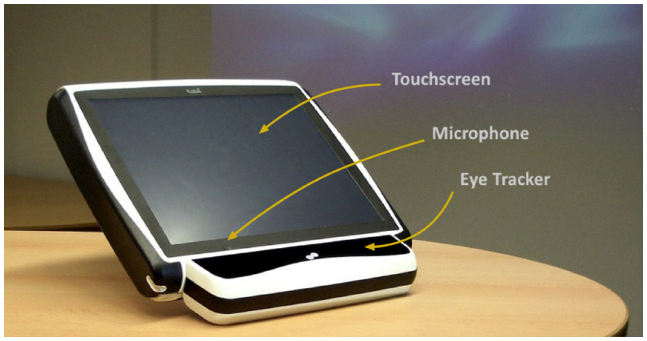


Figure 1. For our interaction research we used a Tobii C12 device. Its main input channels are a touch screen which we use for handwriting recognition, a microphone for speech interaction and an eye tracker to record the user's point of regard.

Numerous work has been done to augment paper books electronically. *Listen Reader*[1] explores sound integration with traditional books. *The mixed reality book*[5] augments book content by adding background music, narrator's voice over, sounds matching pictorial content, animations and augmented surroundings. *The Haunted Book*[7] is an electronic book augmented with animated illustrations of ghost creatures.

However, little work has been done to enhance reading using multimodal interaction techniques. *Novella*[6] is an electronic book reader which combines mouse and speech to navigate and annotate book content. The *eyeBook*[2] is an augmented multimedia book where illustrations, sound effects and background music are adjusted to match the story setting. It also uses gaze control for application interaction (e.g. scrolling).

OVERVIEW AND ARCHITECTURE
We created what we call *eyePad*, a gaze aware, hand writing sensitive, speech responsive prototype that is able to deliver

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

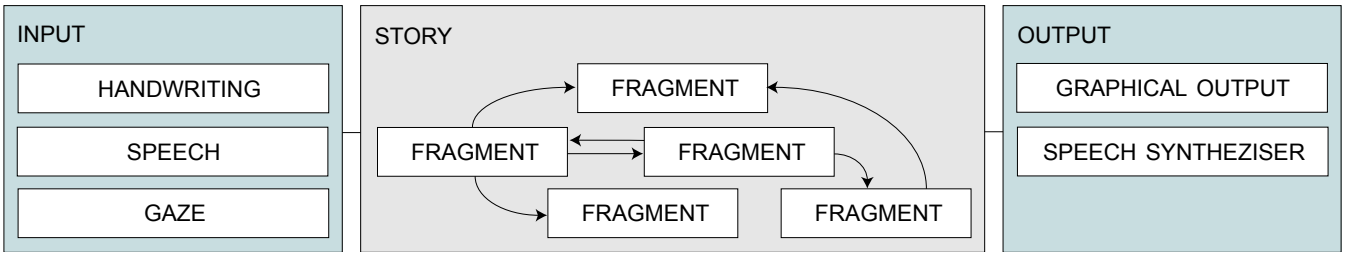


Figure 2. The input evaluators capture the multi-modal input, which is then used by the story module to update the current fragment or to render the next fragment with the help of the rendering/speech facilities.

highly interactive stories[1] to its reader.

The hardware platform is a Tobii C12, see Figure 1, an augmentative and alternative communication (AAC) device intended to be used as an assistive technology for individuals with communication disabilities. The Tobii CEye module is an eye control unit which can be attached to the Tobii C12. The gaze data rate of the Tobii CEye is 30 Hz, and the gaze accuracy is reported to be within 0.8 degrees.

The actual software system is based on a plugin architecture (Figure 2) and fuses the system's main components (compare [4]): input evaluators, the story module, and the rendering and speech output facilities. It is written in Java with the help of Processing[2] and uses parts of the Text 2.0 framework[3].

Input evaluators

The input evaluators monitor the hardware channels and convert the measured *raw* data into high level events which can be facilitated by the story module. In order to convert gaze, handwriting and speech data, we used the following input evaluators:

- Handwriting recognition - Based on an SMO classifier a set of pen coordinates is obtained by observing pen strokes and recording pixel positions for each observation point while it is pressed. A feature vector is then constructed and passed to a previously trained model that is able to perform the classification and emits a list of classified characters. The recognition result is afterwards matched against a set of globally or contextually expected handwriting commands.
- Speech recognition - We built this upon the operating system's inbuilt speech recognition system. Similar to the handwriting recognition the system is primed with a set of expected verbal commands and a callback is executed upon their detection.
- Gaze evaluation - The gaze evaluation module uses the gaze data from the CEye module to keep track of the reader's progress in the text (i.e., which part of the text is currently being read) and if there are regions on the screen that readers can interact with, it keeps track of those that are currently active (i.e., the reader is looking at them). It uses the renderer's layout information and reports input in the form of elements gazed upon.

The Story Module

The story module is the application's heart and drives the story based on the user's input by evaluating story-bundles. A *story bundle* consists of *story fragments*, each containing text, a number of images, expected speech and handwriting commands, and their respective output.

Story fragments have dependencies, that is, only a subset of the story is available to the reader at any moment (of which one story fragment is displayed). Once a fragment has been read or interacted with, the subset changes and the story proceeds according to the user's input and the story's structure.

Due to the nature of story fragments the user's reading behavior is characterized by a high degree of non-linearity, that is, there are several reading paths to choose from and the version of the story that is delivered depends on the specific interactions that took place.

In this respect, the process of reading or interacting with the eyePad is similar to the user's interaction with a *gamebook* or computer game. The player/reader advances in the story by making decisions while the pad dynamically reacts to these choices. The reader may take the decision to perform a certain action, thus triggering a progression in the storyline that is usually irreversible and affects future fragments as well. For example, the pad may ask the reader, "Should X live or die?" causing the reader to respond by writing, "X should die.". The death of X may in turn cause X's mate to seek revenge from the player character or cause the player character to feel remorse for killing X for the rest of the story.

Rendering/Speech facilities

Once the input is evaluated with respect to the current story fragment the output module is employed to display the selected text or synthesize spoken responses. It contains generic renderers for plain text, handwritten text, and images. In addition, special renderers can be implemented as plugins and used for specific types of story fragments.

DEMO SCENARIOS

Using the architecture described above we implemented a story which is explorable by the reader through the afore-

[1]See `http://media.text20.net` for an interaction video with the eyePad.
[2]`http://processing.org`

mentioned kinds of interaction. In addition we integrated a number of special fragments containing novel human-text interaction techniques. The most notable special fragments included:

Interactive map

In the map scenario the reader can follow the protagonist's travel through a wilderness when reading his diary. At the same time, a map (see Figure 3) is updated with suitable icons based on the reader's real time progress in the text. The icons are added transparently and only slightly fade in while his focus is still on the text to minimize distraction. The map served as an interactive visualization and constantly up-to-date remainder of character's location, and the last place of interest is highlighted upon a glance to the left. The reader can also look at any other previously displayed icon and verbally ask 'what is this (again)?' or 'tell me more about this place' to listen to a short abstract. For non-fictional places, like Berlin, the abstract is extracted from Wikipedia, for fictional places, a database has to be populated.

Speaker reminder

Another scenario we looked into was the optimization of dialogs. Many novels contain lengthy conversations between various characters, and frequently the speaker names are omitted due to visual and linguistic elegance, making them hard to follow if one lost track in between. We address this by semantically augmenting dialogs with the actual information about the speaking character. If the reader has no trouble following the text, the book does not trigger any assistance. If the reader encounters, however, any problems during conversations he can gaze on the side of the screen where information about the current speaker is displayed, e.g., an image. In addition, by looking at the image and saying, for example, 'who is she?' the reader can also listen to a brief character profile.

Character reminder

The more general idea of the *speaker reminder* mentioned above was the introduction of a character reminder. The longer a story runs, the more characters are usually introduced and back references can become a source of confusion, especially if they suddenly reappear after some time. Thus we tagged not only the dialog lines with their respective speaker, but we stored a database for every character name and their synonyms. This enables the reader to inquire, for example, 'who was that again?' and likewise receive a brief summary for the character name that the reader momentarily focuses on with his eyes..

EVALUATION

We performed a preliminary analysis of the prototype's capabilities. We designed a user study in which participants interacted with the book and had to perform a number of tasks, the impressions were reported in a questionnaire afterwards. The tasks included:

- responding by handwriting to the book's offers to disclose some parts of its fictitious history.
- reading and interacting with a travel journal, augmented with the map system described above.
- reading a special part of the story that referred to an unknown character which was not introduced before, thus simulating forgetfulness.

Participants

Eight Participants performed all tasks, i.e., four males and four females, with an average age of 21.5 years. They were university graduate and undergraduate students majoring in computer science and engineering. The completion of all four tasks took around 30 minutes (including calibration and training).

Results

Our first question was if the integrated modalities enhanced the reading experience (see Table 1). Although 63% of the participants thought that the pen interaction enhanced their reading experience, one participant mentioned that he would rather tap buttons on the screen than ask/answer questions. 88% of the participants were pleased with the gaze interaction describing it as 'useful' and 75% of the participants thought that the speech interaction enhanced their reading experience.

Figure 3. A disoriented reader looks at the map to find out the last place of interest.

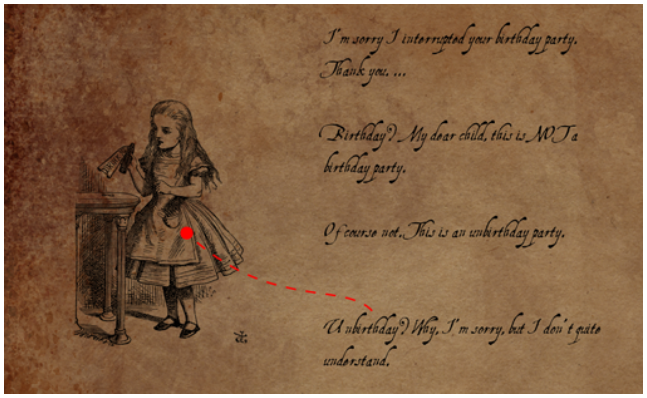


Figure 4. A reader observes an illustration of the unknown speaker after looking to the left.

Table 1. Does integrating pen, speech and gaze input modalities enhance the reading experience?

	Yes	Neutral	No
Gaze interaction	88%	0%	13%
Speech interaction	75%	25%	0%
Pen interaction	63%	25%	13%

Table 2. How do readers find the quality of the eye tracking, the handwriting recognition and the speech recognition/synthesis? Results were reported on a five point Likert scale, 1 equals very bad, 5 equals very good.

	Rating
Handwriting recognition	3.9
Speech recognition	3.9
Speech synthesis	3.5
Eye tracking	3.4

In order to distinguish whether disfavoring ratings were caused by principal interaction flaws or rather caused by a poor implementation of a particular subsystem we also asked the participants for explicit ratings of their perceived performance and accuracy of the individual subsystems. They were asked four questions of the form 'How would you rate the quality of X?' where X is replaced by the subsystem under consideration. The ratings went from 1 (very poor) to 5 (very good), a value of 3 was considered acceptable, see Table 2 for the results.

Additional open feedback could be given as well, in here the most notable points were related to the speech input and output. Some users usually expected their utterances to be recognized, even if they did not speak loudly and in a clear voice. For others the speech synthesis voice was too fast. One participant mentioned that it was difficult to understand, and that it needs a very silent environment. Regarding the eye tracking performance it could be observed that the device's accuracy degraded due to the participants' urge to move during the experiment, resulting in a shifted head position relative to the calibrated position.

We also researched how the readers reacted to the presented scenarios. All participants reported that all of the special assists happened to be helpful. However, although all participants agreed that the interactive map makes the text easier to visualize, one participant mentioned that he did not notice when the icon was placed on the map because he was concentrating too much on the text. Another participant said that it was not clear whether they should look at the map while reading or after reading the text.

The speaker and character reminders were thoroughly received positively, they were reported to be uniquely distinctive to ordinary (e-)books and would really 'add something extra'.

OUTLOOK & CONCLUSION

We presented a multi modal gaze aware, hand writing sensitive, speech responsive prototype on a tablet computer. We implemented a demo story and integrated story fragments containing novel human-text interaction techniques. We also evaluated the prototype's capabilities in a user study that addressed various issues of the interface and the implementation.

Initial results with regards to the overall reading experience and the helpfulness of the proposed reading assists were very satisfactory considering the prototype status. Improving the robustness and quality of the eye tracking, speech recognition/synthesis and handwriting recognition is crucial to the eventual acceptance of the prototype by real users as evident from the results of our user study.

The usage of non-linear story fragments and their implicit control through gaze, or explicit control through handwriting and speech allow for exciting possibilities. Book authors are given an interesting set of plot and interaction means to shape a story according to their own imagination and the reader's progress. We see also plenty of use-cases for our prototype in various domains including e-learning and entertainment.

REFERENCES

1. M. Back, J. Cohen, R. Gold, S. Harrison, and S. Minneman. Listen reader: an electronically augmented paper-based book. In *Proc. SIGCHI conference on Human factors in computing systems*, page 29, 2001.

2. R. Biedert, G. Buscher, and A. Dengel. The eyebook, using eye tracking to enhance the reading experience. *Informatik Spektrum*, 2010.

3. R. Biedert, G. Buscher, S. Schwarz, M. Moeller, A. Dengel, and T. Lottermann. The Text 2.0 Framework. Workshop on Eye Gaze in Intelligent Human Machine Interaction, 2010.

4. A. Gourdol, L. Nigay, D. Salber, J. Coutaz, and L. de Génie Informatique. Two case studies of software architecture for multimodal interactive systems: Voicepaint and a voice-enabled graphical notebook. In *Proc. IFIP TC2/WG2*, volume 7, pages 271–284.

5. R. Grasset, A. Duenser, H. Seichter, and M. Billinghurst. The mixed reality book: a new multimedia reading experience. In *CHI '07 extended abstracts on Human factors in computing systems*, page 1958, 2007.

6. J. Hodas, N. Sundaresan, J. Jackson, B. Duncan, W. Nissen, and J. Battista. NOVeLLA: A multi-modal electronic-book reader with visual and auditory interfaces. *International Journal of Speech Technology*, 4(3):269–284, 2001.

7. C. Scherrer, J. Pilet, P. Fua, and V. Lepetit. The haunted book. In *Proc. 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, volume 0, pages 163–164, 2008.

Challenges and Limits of Gaze-including Interaction

Sandra Trösterer
Chair of Human-Machine Systems, TU Berlin
Franklinstr. 28-29, 10587 Berlin, Germany
str@mms.tu-berlin.de
+49 30 314-79522

Jeronimo Dzaack
Chair of Human-Machine Systems, TU Berlin
Franklinstr. 28-29, 10587 Berlin, Germany
jdz@mms.tu-berlin.de
+49 30 314-79519

ABSTRACT
Interacting with computing systems can be demanding for the user due to factors such as the complexity of the task or time pressure. So far it is unclear how new interaction techniques that include gaze as input modality are affected. In this paper we present the results of a study in which we investigated the influence of mental workload and visual distraction on gaze-including interaction. Gaze-based interaction with two different dwell times, a combination of gaze and key presses, and for control reasons mouse interaction were investigated. We found that gaze-including interaction reaches its limit in situations of increased workload, especially with visual distraction added.

Author Keywords
Gaze-based interaction, gaze-including interaction, mental workload, visual distraction.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION AND THEORETICAL BACKGROUND

Using gaze data to control computing system gains in importance in modern human-computer interaction (HCI) as it offers a way of contactless and fast interaction. Nevertheless we have to keep in mind that in a working context the requirements that are put on the user can be challenging. Time pressure, task complexity or the need for multitasking might stress the user. Therefore, designing gaze-including interaction in that context is a challenge, especially because the gaze has special features that might hinder interaction efficiency. One general problem is the so-called "Midas touch" problem, i.e. the gaze is always online, and therefore we have to be aware that everywhere we look, we might activate a command [4]. Another problem lies in the fact that, according to the theory of guided search, our visual search is influenced by task-related knowledge (top down) but also salient features of our environment (bottom up, [7]), i.e. our gaze is much more distraction-prone compared to a mouse movement. Additionally the kind of task, task know-how and task difficulty can lead to changes in eye movement behavior (see [6] for an overview). Regarding mental workload, i.e. the total amount of mental activity imposed on working memory at an instance in time [1], e.g. [5] found that imposed cognitive workload leads to a more data-driven behavior, i.e. one reacts even more to salient features. Now the question is, if our gaze is that sensitive, what will happen if we use it to interact with a computer system in situations of increased workload? How will this affect the performance and efficiency? Does a gaze-including kind of interaction reach its limits in such a context? In order to gain more insight and to derive design notes, we conducted an explorative study investigating the influence of mental workload and visual distraction on different kinds of interaction.

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

EXPERIMENTAL STUDY

Method

Participants

39 subjects (19 female, 20 male) between the ages of 21 and 38 years (mean age 26, *s*=3.7) participated in the study. All subjects were European and PC-users who spend approximately 4.5 hours a day working with a PC and mouse. 20 subjects had already experience with eye tracking, 9 subjects even with gaze-based interaction.

Experimental design

The study was realized as a 2x3x4 design, where the kind of workload (*between*), the level of workload (*within*) and the kind of interaction (*within*) have been varied. Basically subjects had to perform a simple searching task: Four capital letters (one target *X* and three *O*s as distractors) were presented at the corners of the computer screen and four buttons were displayed at the center of the screen. The goal was to find the target letter *X* in one of the corners of the display and to confirm its position with the corresponding button on the screen. If the target *X* was e.g. in the lower left corner, subjects had to press the lower left button. This had to be done within a time interval of 4000 ms. After a button-press or a time-out, a fixation cross was presented for 1000 ms before the next trial (see Figure 1). In every trial the *X* appeared randomly at another corner.

For one group of subjects workload was externally induced, i.e. subjects had to perform a (1) secondary acoustic task simultaneously with the simple searching task. The acoustic task was an adapted acoustic version of the Sternberg task, i.e. five random numbers between 1 and 9 were presented orally to the subject, followed by a sound and one number. Subjects then had to indicate orally whether this number was part of the previously mentioned numbers or not. A second sound indicated the beginning of the next trial. The acoustic task was independent of the searching task, i.e. the tasks were not clocked.

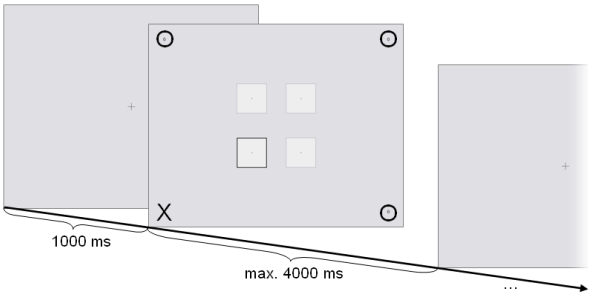


Figure 1. Searching task (*X* and *O*s are enlarged for better visibility and corresponding button is marked)

For the other group of subjects workload was induced by a (2) higher complexity of the searching task. Therefore, the fixation cross was substituted by a small arrow showing a rotation to a certain direction (right or left) and a certain amount of rotation (quarter, half, three quarters, full). Subjects had to keep in mind that rotation as it indicated the position of the correct button based on the corresponding button, i.e. if the *X* was in the lower left corner, but a quarter right rotation was shown before, the correct button to press was the upper left button.

The level of induced workload was varied in three stages. In the (1) low workload (LW) condition subjects had to perform the simple searching task (in both groups). In the (2) higher workload (HW) condition subjects had to perform the searching task with either the secondary acoustic task (group 1) or the higher task complexity (group 2). In the (3) highest workload (HTW) condition the tasks were the same as in the HW condition, but visual distraction was added, i.e. every 200ms the color of one of the four buttons randomly changed to red.

Subjects had to perform the tasks with four different kinds of interaction: gaze-based interaction with (1) 500 ms and (2) 250 ms dwell time respectively, a (3) combination of gaze to point and a key press to activate the buttons on the screen, and (4) mouse interaction. The order was permutated for each subject. Within a certain kind of interaction the order in which subjects had to perform the tasks regarding the level of workload was permutated as well. For every subject reaction times, errors (false responses, time outs), subjective experienced stress indicated on the German "Skala zur Erfassung subjektiv erlebter Anstrengung" (SEA-Skala; [3]), i.e. "scale for the acquisition of subjective experienced stress", and the gaze data were recorded. Additionally the answers to a structured interview at the end of the experimental session were transcribed.

Materials and apparatus

The experiment was realized on a PC with a standard mouse and a 19" monitor with 1280x1024 pixel resolution. The displayed buttons on the monitor had a size of 150x150 pixels, letters were 17 pixels in height and 14 pixels in width. The button size was chosen this large to allow an easy gaze-interaction and is based on empirical values of pretests. The small size of the letters was chosen to evoke visual search and avoid pop-out effects. An iView-X RED (SensoMotoric Instruments) eye-tracking unit was mounted under that monitor and was controlled by a further laptop (see Figure 2). Gaze data was tracked with 50Hz resolution and was processed in real-time using iCOMMIC (integrated COntroller for MultiModal InteraCtion; [2]), in order to transform the data to corresponding pointing and manipulation commands. Each kind of interaction was set up in this framework. For the mouse interaction the cursor was displayed in the usual way, for the gaze-including interaction forms the cursor was set half-transparent in order to reduce visual distraction by the cursor itself. For the combined condition a wireless number pad (Logitech N305, 2.4 GHz wireless connection) was used to perform the key press.

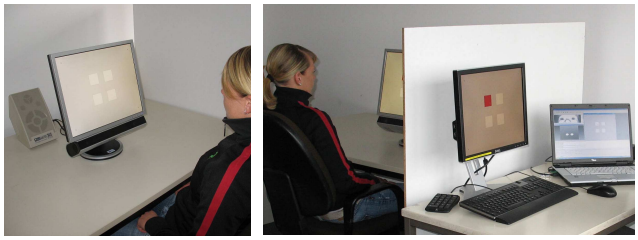


Figure 2. Experimental setup

A java applet was developed to present the secondary acoustic task in a standardized way. This applet was controlled on a further laptop with two sound speakers attached that were positioned behind the monitor in order to guarantee best sound. The demographic questionnaire, written instructions, and the SEA-scale were given to the subjects in paper form.

Procedure

Subjects were welcomed and seated at a desk to fill out a demographic questionnaire. Afterwards, they were informed that the aim of the study is to investigate different kinds of interaction. Instructions for the simple search task were given to the subject in a written form. They were instructed to perform the task as fast and as accurate as possible. After calibration of the eye-tracker, the subject had the possibility to train the task using the mouse. Therefore, 20 trials were presented to the subject. After they finished the training, subjects were given the instructions for the second task (HW) and the third task

(HTW). Subjects then had the possibility to train these tasks too with 20 trials each. After the training, subjects were calibrated again. In the main part of the study each subject fulfilled each task (40 trials) with each kind of interaction, i.e. 12 conditions. After each task subjects were asked to assess their mental stress using the SEA-scale. While filling out the SEA-scale, the experimenter configured the system for the next task and/or the next kind of interaction. After each kind of interaction, the calibration of the eye was reassessed and if necessary a new calibration was performed in order to ensure the most accurate gaze. Finally, participants were interviewed and asked what kind of interaction they preferred the most and which interaction they liked the least. Overall, the experiment lasted one hour. At the end of the experiment, subjects were compensated with 10 Euros and thanked for their participation.

Results

In the following section we present selected results regarding the most important aspects of our study. Data was preprocessed using a routine programmed in Java. Further statistical analysis was conducted with PASW Statistics 18.

Subjective experienced stress

In order to control whether workload was really induced to the subjects, we first had a look at the subjective experienced stress. A 3-factorial ANOVA with repeated measures on two factors was conducted, including the variables kind and level of workload and the kind of interaction. We found a significant interaction for the kind and level of induced workload ($p<.05$). As can be seen in Figure 3 there is no difference in the subjective experienced stress for the LW condition.

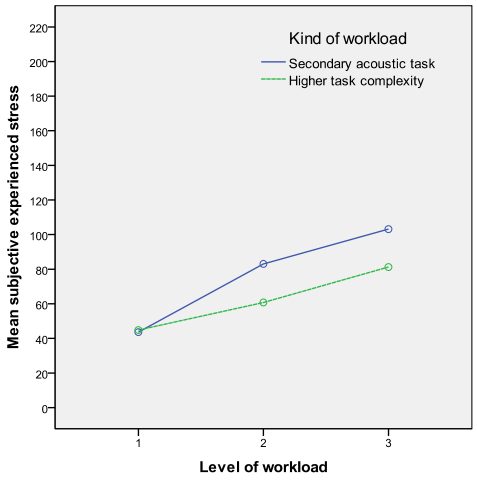


Figure 3. Mean subjective experienced stress for kind and level of workload (1=LW, 2=HW, 3= HTW)

This indicates that the groups had the same abilities to perform the task, as both groups had to perform the simple searching task on that level. For the HW and HTW level we find a significant difference between the two kinds of workload ($p<.05$). The task with the additional acoustic task is experienced more stressful (mean SEA-value 93) than the task with the higher complexity (71; $p<.05$). Generally we found a significant main effect for the level of workload ($p<.001$). In the LW condition the mean SEA-value is 44, in the HW condition it is 72, and 92 in the HTW condition (all $p<.001$). Therefore, we can conclude that workload was successfully induced. Furthermore we found a highly significant main effect ($p<.001$) for the kind of interaction. Mouse interaction is experienced significantly less stressful than all other kinds of interaction (all $p<.001$). The mean SEA-value was 55 for the mouse, 69 for the combined, 73 for the 500ms dwell and 80 for the 250ms dwell condition. We could not find a significant difference in experienced stress between the gaze-including kinds of interaction.

Reaction times for correct responses

Did the workload affect the performance of the subjects regarding their reaction times and what is the impact of the different kinds of interaction? To answer these questions we computed a 3-factorial ANOVA with repeated measures on two factors as well.

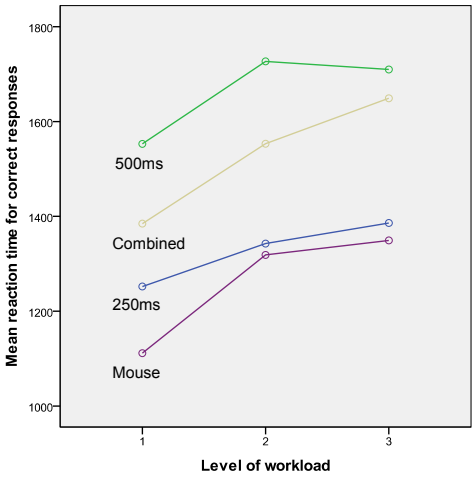


Figure 4. Mean reaction time for level of workload (1=LW, 2=HW, 3=HTW) and kind of interaction

We found a highly significant interaction for the level of workload and the kind of interaction ($p<.001$) and a highly significant main effect for the level of workload ($p<.001$). Generally, in the LW condition mean reaction times for correct responses are lower (1325ms) than in the HW condition (1485ms), but as can be seen in Figure 4 from the HW to HTW condition there is a further increase in reaction time just in the combined condition. No significant differences could be found when comparing the HW and HTW level for all other kinds of interaction. Additionally we found a highly significant main effect for the kind of interaction ($p<.001$). Mean reaction times for mouse and 250ms dwell gaze-based interaction do not significantly differ, whereas reaction times for the combined condition and the 500ms dwell gaze-based interaction are each significantly higher ($p<.001$).

False response errors

Regarding the total number of false response errors, we found, that the number is significantly lower in the mouse (67 errors) and combined condition (77) compared to the 250ms (637) and 500ms dwell time gaze-based condition (276). Chi-Tests revealed no significant difference between the mouse and combined condition but highly significant differences to and between all other conditions (all $p<.001$). Regarding the impact of the level of workload (see Figure 5), we found for the mouse and combined condition, that there is a significant increase in number of errors from the LW to the HW and HTW condition (all $p<.001$), but there is no difference between HW and HTW. In the gaze-based 500ms dwell condition we found a successive increase in number of errors for each level of workload (all $p<.001$). In the 250ms dwell condition we found no significant difference between the LW and HW condition, but a significant difference to the HTW condition ($p<.05$).

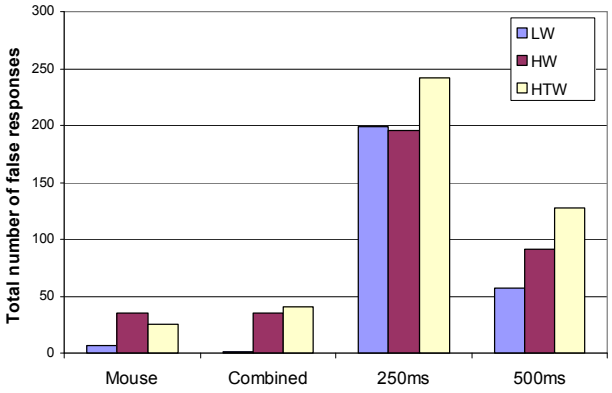


Figure 5. Total number of false response errors for kind of interaction and level of workload

CONCLUSION

In this experimental study we investigated the influence of imposed mental workload and visual distraction on gaze-including and mouse interaction. We found that mouse interaction was experienced as least stressful overall. This might be due to the "mouse-friendly" interface design we used in this investigation, but also to a lack of long-term experience with gaze-including kinds of interaction. It has to be pointed out, that even the combined condition was experienced as stressful as dwell time based gaze-interaction, although subjects had more control in this condition.

We found that the combined interaction led to comparably high reaction times for correct responses and was surprisingly sensitive to workload. Although error rates were comparable to mouse interaction, the necessity of precise eye-hand coordination in the combined condition led to a sort of reassurement behavior. Although subjects already looked at the correct button very quickly, they did not press immediately the key, but reassured themselves that it is really the correct button before pressing. In the HTW condition this reassurement behavior was obviously disturbed by the added visual distraction. From that point of view, it has to be taken into consideration that the coordination of different modalities leads to strong requirements for feedback design. In our study the borders of the button changed to indicate that it is chosen. More distinct feedback might have supported the combined interaction.

Finally we could show that visual distraction in combination with increased workload can lead to severe problems in gaze-based interaction. In both dwell time conditions error rates were highest in the condition where visual distraction was added. Although, as opposed to the combined condition, reaction times for correct responses were not affected in the same way.

From our results we want to develop design guidelines that are adjusted to gaze-including interaction and take into account future demands of HCI (e.g. complex information presentation, aggregated displays). We believe that this study is a base for further research and ongoing studies.

ACKNOWLEDGMENTS

We thank Mandy Dotzauer and Jessika Reissland for preparing and conducting the experiment, Mario Lasch for his technical support, and Marius Schwalbe for his programming work.

REFERENCES

1. Cooper, G. Research into Cognitive Load Theory and Instructional Design at UNSW. http://dwb4.unl.edu/Diss/Cooper/UNSW.htm.
2. Dzaack, J., Trösterer, S, Nicolai, T. and Rötting, M. iCOMMIC: Multimodal Interaction in Computing Systems. In *Proceedings of the 17th World Congress of the International Ergonomics Association* [CD], 2009, ID 3EP0100.
3. Eilers, K., Nachreiner, F. & Hänecke, K. Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung. *Zeitschrift für Arbeitswissenschaft*, 40 (1986), 215- 224.
4. Jacob, R. J. What You Look at is What You Get: Eye Movement-Based Interaction Techniques. In *Human Factors in Computing Systems: CHI '90 Conference Proceedings*. ACM Press (1990), 11-18.
5. Myers, C.W., Gray, W.D. and Schoelles, M. The effects of stimulus configuration and cognitive workload on saccadic selectivity. *Journal of Vision 4,* 8 (2004), article 740.
6. Rötting, M. *Parametersystematik der Augen- und Blickbewegungen für arbeitswissenschaftliche Untersuchungen.* Shaker, Aachen, 2001.
7. Wolfe, J.M. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1 (1994), 202-238.

Automated Analysis of Mutual Gaze in Human Conversational Pairs

Frank Broz, Hagen Lehmann, Chrystopher L. Nehaniv, and Kerstin Dautenhahn
Computer Science
University of Hertfordshire, UK
{f.broz, h.lehmann, c.l.nehaniv, k.dautenhahn}@herts.ac.uk

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. As a first step towards this goal, a Markov model representation of human gaze data is produced. We also discuss how an algebraic analysis of the state transition structure of such models may reveal interesting properties of human gaze interaction.

INTRODUCTION

Mutual gaze is an ongoing process between two interactors jointly regulating their eye contact, rather than an atomic action by either person [2]. It plays an important part in regulating face-to-face communication, including conversational turn-taking in adults [13]. The ability to detect face-directed gaze is present from an early developmental stage; even young infants are responsive to being the object of a caretaker's gaze [12]. Mutual gaze behavior in humans is the basis of and precursor to more complex task-oriented gaze behaviors such as visual joint attention [11].

Compared to other primate species humans have very visible eyes [14, 15]. A possible explanation for this phenomenon is the evolution of a new function of the human eye in close range social interactions as additional source of information about the intention of the other [24]. In many studies it has been shown that apes and monkeys have no or only very limited abilities to follow a human experimenters eye movement to locate a hidden reward [6]. Human infants on the other hand are able to follow eye movements from around 18 months of age [7].

Humans rely heavily on gaze information from their conspecifics especially during cooperative, mutualistic social interactions. The importance of eye gaze 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 [3, 5, 22]. 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 [5, 4] enabling us to interact empathically. For these reasons, humans need eye gaze information in order to feel comfortable and to function adequately while interacting with other humans.

In order to develop artificial systems with which humans feel comfortable interacting, it is necessary to understand the mechanisms of human gaze, especially if these systems are humanoid robots. 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 [26, 27, 25] or not reactive to the human partner's gaze actions [20]. Conversational gaze behavior is an interaction, and the robot's gaze policy will have an impact on the human's gaze behavior and the impressions they form about the robot.

In order to support natural and effective gaze interaction, it is worthwhile to first look at gaze behavior in human-human pairs. By examining human gaze, we can gain insight into how to build better gaze policies for robots that interact with people. The approach presented in this paper enables us to monitor the gaze behaviour in a dyadic interaction in real time and this allows a thorough and very detailed analysis.

EXPERIMENT

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

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

and processing their output. The interprocess communication was implemented using YARP [17].

ASL MobileEye gaze tracking systems were used to collect the gaze direction data [1]. The output of the scene camera of each system was input into face-tracking software based on the faceAPI library [23]. Each participant also wore a microphone which was used to record a simple sound level (speech content was not stored). 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 hertz. 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 [19].

Setup

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. They 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.

RESULTS

Ten pairs of people participated in the study. Of these pairs, five experienced errors during data collection that resulted in their data being discarded. 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.

Data Analysis

For each pair, the contiguous two minute period of their data with the lowest number of tracking errors was selected for analysis. The data was classified into high-level behavioral states depending on where both participants were looking and who was speaking at each timestep. In all pairs observed, one participant looked at their partner noticeably more than the other. The participant with the high face-directed gaze level will be referred to as the "high" participant and the partner with the lower level of face-directed gaze will be referred to as "low". The gaze states and their descriptions are given below:

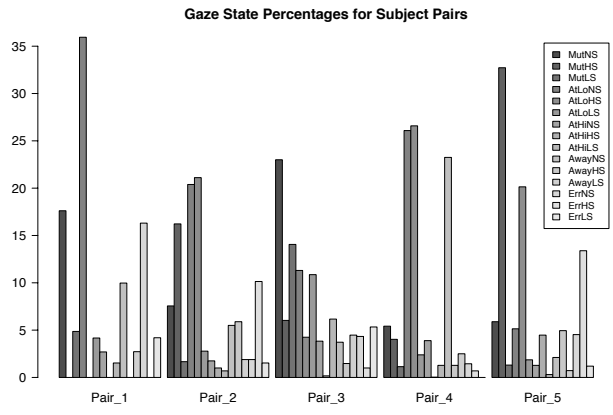


Figure 1. The percentage of time spent in each gaze state by each conversational pair.

- 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 level partner while they look elsewhere
- At High - the low gaze level partner looks at the face of the high level partner while they look elsewhere
- Away - both partners look somewhere other than their partner's face
- Err - gaze state could not be classified due to missing gaze direction or face location readings

It should be noted that the "Err" state may be caused by loss of face tracking that is due either to failures of the face tracker or to the partner's face being undetected because a person has turned their head away. This state measures a combination of system error and participant behavior that we cannot reliably distinguish between in this data set. We intend to address this in future experiments for the purpose of analysis, but this way of modeling error is consistent with how a humanoid robot using the face tracker as input to its controller would experience it.

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 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 de-

tection 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. We intend to record full speech in future experiments to allow for more accurate and detailed analysis. The high level states used for analysis were created by combining the gaze states described above with the state information about which participant in the pair was speaking as follows:

- NS - neither participant is speaking
- HS - high gaze level participant is speaking
- LS - low gaze level participant is speaking

There are fifteen behavioral states in all. The overall amount of time spent in each state by each pair is shown in Figure 1. 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 future experiments, we intend to collect data from a larger set of participants so that we can use statistical tests to identify factors that influence gaze behavior. We will also use questionnaires to measure traits (such as personality) that can't be observed directly from the behavioral data yet may have an impact on gaze behavior.

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 [18]. Data from the contiguous two minute period with the lowest error rate for each pair was combined to construct a model of their average behavior. This model is shown in Figure 2. 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 that 0.01 probability are not shown.

It can be seen in Figure 2 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.

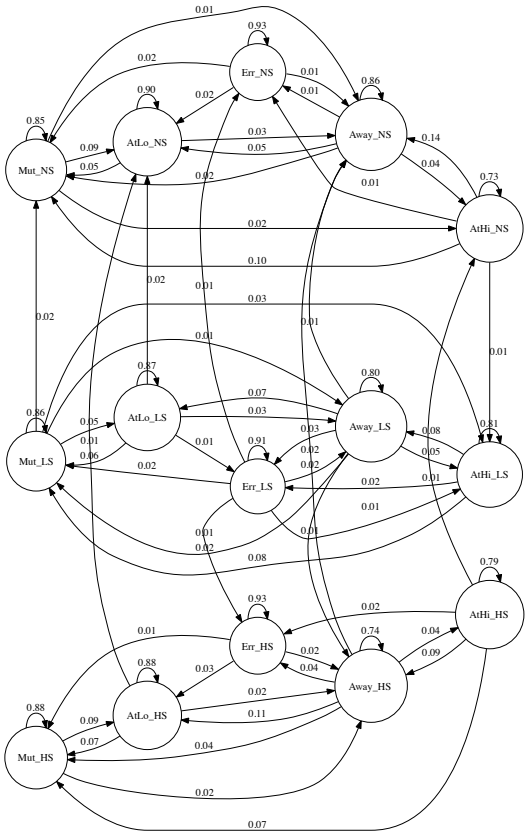


Figure 2. Markov model of the gaze state transitions for all of the conversational pairs.

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 [16], 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 analyse marriages or other interpersonal relationships [21]. 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 [9, 10, 8] 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.

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. We are currently exploring what aspects of interaction are reflected by this algebraic complexity.

The behavior of pair 4 is clearly distinct from the other pairs (as can be seen in Figure 1) 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 too little data to determine whether this may play a role in the behavioral differences observed, but it is an area for further investigation.

CONCLUSION

In this paper, a system for the automated detection of mutual gaze was described, and results were presented from natural conversational interactions between human pairs. This real time system is designed not purely for analysis, but to provide gaze information as input to a controller for a humanoid robot in the future. As a demonstration of how we intend to use human-human gaze and speech data to produce a robotic gaze controller, we created a Markov model from the data collected that captures the gaze behavior dynamics of the human conversational pairs. Additionally, we presented preliminary results from an algebraic analysis of the structure of the resulting Markov model and discussed how this type of analysis may be used to computationally investigate qualities of the gaze interaction.

REFERENCES

1. Applied Science Laboratories. Mobile eye gaze tracking system. `http://asleyetracking.com/`.
2. M. Argyle. *Bodily communication*. Routledge, second edition, 1988.
3. S. Baron-Cohen, R. Campbell, A. Karmiloff-Smith, J. Grant, and J. Walker. Are children with autism blind to the mentalistic significance of the eyes? *Br. J. Dev. Psychol.*, 13:379–398, 1995.
4. S. Baron-Cohen, J. Wheelwright, Y. Hill, and I. RastePlumb. 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. Psychiat.*, 42:241–252, 2001.
5. S. Baron-Cohen, S. Wheelwright, and T. Jolliffe. Is there a "language of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome. *Vis. Cogn.*, 4:311–331, 1997.
6. J. Call and M. Tomasello. Social cognition. In D. Maestripieri, editor, *Primate Psychology*, pages 234–253. Harvard University Press, Cambridge, MA, 2003.
7. V. Corkum and C. Moore. Development of joint visual attention in infants. In C. Moore and P. Dunham, editors, *Joint Attention: Its Origins and Role in Development*. Erlbaum, Hillsdale, NJ, 1995.
8. A. Egri-Nagy and C. L. Nehaniv. Sgpdec - hierarchical composition and decomposition of permutation groups and transformation semigroups. `http://sgpdec.sourceforge.net/`.
9. A. Egri-Nagy and C. L. Nehaniv. Algebraic hierarchical decomposition of finite state automata: Comparison of implementations for krohn-rhodes theory. *Implementation and Application of Automata: 9th International Conference, CIAA 2004, Kingston, Canada, July 22-24, 2004, Revised Selected Papers*, 3317:315–316, 2005.
10. A. Egri-Nagy and C. L. Nehaniv. Hierarchical coordinate systems for understanding complexity and its evolution, with applications to genetic regulatory networks. *Artificial Life (Special Issue on Evolution of Complexity)*, 14(3):299–312, 2008.
11. T. Farroni. Infants perceiving and acting on the eyes: Tests of an evolutionary hypothesis. *Journal of Experimental Child Psychology*, 85(3):199–212, July 2003.
12. S. M. Hains and D. W. Muir. Infant sensitivity to adult eye direction. *Child development*, 67(5):1940–1951, October 1996.
13. C. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, 1986.
14. H. Kobayashi and S. Kohshima. Unique morphology of the human eye. *Nature*, 387:767–768, 1997.
15. H. Kobayashi and S. Kohshima. 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, 2001.
16. K. Krohn and J. Rhodes. Algebraic theory of machines. I. prime decomposition theorem for finite semigroups and machines. *Transactions of the American Mathematical Society*, 116:450–464, 1965.
17. G. Metta, P. Fitzpatrick, and L. Natale. Yarp: Yet another robot platform. *International Journal of Advanced Robotics Systems, special issue on Software Development and Integration in Robotics*, 3(1), 2006.
18. S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.
19. D. L. Mills. Improved algorithms for synchronizing computer network clocks. *SIGCOMM Computer Communication Review*, 24:317–327, October 1994.
20. B. Mutlu, J. Forlizzi, and J. Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoids*, pages 518–523, 2006.
21. J. Rhodes. *Applications of Automata Theory and Algebra via the Mathematical Theory of Complexity to Finite-State Physics, Biology, Philosophy, and Games*. World Scientific Pres, 2009.
22. J. Ristic and A. Kingstone. Taking control of reflexive social attention. *Cognition*, 94(3):B55–65, 2005.
23. Seeing Machines, Inc. faceAPI. `http://seeingmachines.com/`.
24. M. Tomasello, B. Hare, H. Lehmann, and J. Call. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, 52:314–320, 2007.
25. A.-L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wredek. People modify their tutoring behavior in robot-directed interaction for action learning. In *DEVLRN '09: Proceedings of the 2009 IEEE 8th International Conference on Development and Learning*, pages 1–6, Washington, DC, USA, 2009. IEEE Computer Society.
26. Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto. The effects of responsive eye movement and blinking behavior in a communication robot. In *IROS*, pages 4564–4569, 2006.
27. C. Yu, M. Scheutz, and P. Schermerhorn. Investigating multimodal real-time patterns of joint attention in an hri word learning task. In *HRI '10: 5th ACM/IEEE international conference on Human-robot interaction*, pages 309–316, New York, NY, USA, 2010. ACM.

The Role of Eye Tracking in Adaptive Information Visualization

Anna Flag,Mona Haraty,Guiseppe Carenini,Cristina Conati
Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, BC
aflagg,haraty,carenini,conati@cs.ubc.ca

ABSTRACT
We plan to design adaptive information visualization systems that adjust to the specific needs of each individual viewer. Our first step is to explore data sources that could help predict different levels of performance on visualization tasks, including interface interactions, eye-tracking, and physiological sensors. In this poster, we discuss how a viewer's gaze pattern could inform the design of adaptive visualization systems.

Author Keywords
Gaze, information visualization, adaptive user interface, individual differences, eye tracker.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Information visualization is a thriving area of research in the study of human/computer communication. However, attempts to measure and formalize visualization effectiveness often have led to inconclusive and conflicting results [14]. We believe this is because existing visualizations are designed mostly around the target data set and associated task model, with little consideration for individual differences. Both long term traits like cognitive abilities and short term factors like mental state have been largely overlooked in the design of information visualization systems, despite studies linking individual differences to visualization efficacy for search and navigation tasks [1,7], as well as anecdotal evidence of diverse personal visualization preferences [3]. Thus we plan to explore the possibilities of intelligent, human-centered visualizations that *understand* different users have different visualization needs and abilities, and can *adapt* to these differences.

There is already some evidence of the impact individual differences can have on visualization effectiveness. For example, Velez, Silver, and Tremaine [15] found significant correlations between individual spatial abilities and performance on identification of a 3D object from visualizations of its orthogonal projections. Conati and Maclaren [5] found that an individual's perceptual speed was a significant predictor of her ease in understanding the same data set with two different visualization types.

Although the benefits of user-adaptive interaction have been shown in a variety of tasks such as operation of menu-based interfaces, web browsing, desktop assistance and human-learning [10], these ideas have rarely been applied to data visualization. This is largely due to the fact that there is limited understanding of which combinations of user traits/goals are relevant for adaptivity. Two notable exceptions are the work by Gotz and Wen [8], and by Busilowsky et al. [4].

Gotz and Wen [8] propose a technique to automatically detect a user's changing goals during interaction with a multi-purpose visualization, and adapt the visualization accordingly. In contrast, we focus on adapting the visualizations to other relevant user-dependent factors in addition to goals. Brusilowsky et al. [4] adapt the *content* of the visualization to the user's state in an educational system, but maintain a fixed visualization technique. In contrast, we are interested in adaptation that involves both selecting *alternative visualizations* for different users, as well as providing *adaptive help* with a given visualization to accommodate changing user needs during interaction.

To achieve this objective, two research questions need to be answered: 1) given a visualization, why do some people perform better than others, and 2) how can a visualization system detect when a user is not performing well. We plan to explore two avenues to answer these questions. One is to investigate further how long term user traits (e.g, spatial/perceptual abilities, personality traits, learning styles) may impact visualization effectiveness. If such measurable features are found and are collectible before interaction, they could be given as input to the system to help it select the best visualization method for this viewer. Our second approach is to study whether user proficiency with a given visualization can be inferred from her interaction behaviors. We believe that an important window into these behaviors can be provided by eye-tracking information. The rest of the paper focuses on some preliminary ideas on how this information can be collected and utilized in the design of adaptive visualizations.

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

GAZE PATTERN: AN INPUT TO ADAPTIVE VISUALIZATIONS

Several researchers have explored eye-tracking as a source of information for real-time assessment of human/machine interaction performance. Amershi and Conati [2] used an unsupervised machine learning technique to separate effective and ineffective user behaviors during interaction with a teaching tool for math. The behaviors captured both interface actions as well as attention patterns monitored via eye-tracking. We plan to conduct similar studies to try to reproduce these results in the context of visualizations.

Iqbal and Bailey [9] found that a given task, including a reading comprehension, searching, or object manipulation, has a unique signature of eye movement. We hypothesize that the correlation between task performance and a specific class of gaze patterns might depend on the type of visualization being used. To test this hypothesis, we will analyze several visualizations such as bar graph, line plot and pie chart for a specific task. For example, users will be asked to perform the following filtering task: "Find data cases satisfying the following concrete conditions on attribute values.". We will monitor the resulting interactions and see if we can identify common "successful" gaze patterns for each visualization.

We believe the following results can help us model user performance on a visualization using gaze behavior:

1. The duration of fixations on each area of interest is an indicator of the complexity of that area [6,11]

A study by Crowe and Narayanan found that-as one might expect-an unusually long fixation on one component of a visualization indicates lack of understanding of that component [6]. Identifying these areas may make for a more focused adaptation, because it allows the system to target the specific area that is perplexing the viewer.

2. Degree of pupil dilation has been proved to be a valid and reliable measure of cognitive load [12]

We plan to investigate if pupil dilation as measured via an eye-tracker can be a reliable indication of cognitive load during visualization processing. If this is the case, detecting high cognitive load could prompt the system to take steps to simplify the data presentation or the viewer's task.

3. Users do not look at all areas of interest [9,13]

Analyzing gaze locations might be a good first step to identifying when a viewer is having trouble with a given visualization. Lohmann, et al. used this approach to compare relative effectiveness of alternative tag cloud visualizations in the context of drawing attention to the areas of greatest interest [13]. Gaze locations, and the locations that have been overlooked, can inform the design of the adaptive help. For instance, after becoming aware the viewer is not looking at an area of crucial importance, the visualization could emphasize this area to attract attention.

UTILIZING GAZE DATA FOR TWO TYPES OF ADAPTATION

We are interested in adaptation that involves both selecting different visualizations for different viewers, as well as providing adaptive help within a visualization to accommodate changing user needs during interaction. For example, given a set of alternative visualizations, our adaptive system would monitor the interaction and may switch visualizations if the current display does not appear to be working for the viewer. During the interaction itself, the system would focus more on providing explicit interactive help, such as drawing attention to certain important areas or explaining explicitly how to derive a given piece of information from the current visualization.

These proposals for adaptation must be thoroughly tested within the context of information visualization before they can be realistically applied. Thus, in addition to conducting studies to validate the use of eye-tracking data in detecting when a viewer is having difficulties, we plan to investigate the benefits and feasibility of a variety of adaptive interventions within the context of information visualizations.

REFERENCES

1. Allen, B. Individual differences and the conundrums of user-centered design: Two experiments. *Journal of the American Society for Information Science 51*, 6 (2000), 508-520.
2. Amershi, S. and Conati, C. Unsupervised and supervised machine learning in user modeling for intelligent learning environments. *Proc. of the 12th intl. conf. on Intelligent user interfaces*, (2007), 72–81.
3. Baldonado, M.Q.W., Woodruff, A., and Kuchinsky, A. Guidelines for using multiple views in information visualization. *Proc. of the working conf. on Advanced visual interfaces*, ACM (2000), 110-119.
4. Brusilovsky, P. and Su, H.D. Adaptive visualization component of a distributed web-based adaptive educational system.*Intelligent Tutoring Systems:6th Intl. Conf., ITS, Biarritz, France & San Sebastian, Spain, June 2-7, Proc. (2002)*, 309–316.
5. Conati, C. and Maclaren, H. Exploring the role of individual differences in information visualization. *Proc. working conf. on Advanced visual interfaces*, (2008), 199–206.
6. Crowe, E.C. and Narayanan, N.H. Comparing interfaces based on what users watch & do. Proc. eye tracking research & applications symp. (2000), 29-36.
7. Dillon, A. Spatial-semantics: how users derive shape from information space. *J. Am. Soc. Inf. Sci. 51*, 6 (2000), 521-528.
8. Gotz, D. and Wen, Z. Behavior-driven visualization recommendation. *Proc. 13th intl. conf. on Intelligent user interfaces*, ACM (2009), 315-324.
9. Iqbal, S.T. and Bailey, B.P. Using eye gaze patterns to identify user tasks. *The Grace Hopper Celebration of Women in Computing*, (2004).
10. Jameson, A. Adaptive interfaces and agents. In *human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. L.

Erlbaum Associates Inc., 2003, 305-330.
11. Just, M.A. and Carpenter, P.A. A theory of reading. *Psycholinguistics: critical concepts in psychology 87*, 4 (2002), 365.
12. Loewenfeld, I. and Wilhelm, H. The Pupil: Anatomy, Physiology, & Clinical Applications, Volumes I and II. *ARCHIVES OF OPHTHALMOLOGY 118*, (2000), 864–864.
13. Lohmann, S., Ziegler, J., and Tetzlaff, L. Comparison of tag cloud layouts: Task-related performance and visual exploration. *INTERACT*, (2009), 392–404.
14. Nowell, L., Schulman, R., and Hix, D. Graphical encoding for information visualization: an empirical study. *IEEE Symp. on Information Visualization, 2002. INFOVIS 2002*, (2002), 43–50.
15. Velez, M.C., Silver, D., and Tremaine, M. Understanding visualization through spatial ability differences. (2005).