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 [1]. It plays an important part in regulating face-to-face communication, including conversational turn-taking in adults [12]. 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 [11]. Mutual gaze behavior in humans is the basis of and precursor to more complex task-oriented gaze behaviors such as visual joint attention [10].

Compared to other primate species humans have very visible eyes [13, 14]. 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 [5]. Human infants on the other hand are able to follow eye movements from around 18 months of age [6].

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 [2, 4, 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 [4, 3] 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 and processing their output. The interprocess communication was implemented using YARP [17].

IUI'11, February 13–16, 2011, Palo Alto, California, USA.

ASL MobileEye gaze tracking systems were used to collect the gaze direction data [16]. 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 from the study. The nature of these errors were: loss of gaze tracker calibration due to the glasses with the camera mount slipping or being moved by the participant, failure of the face tracker to acquire and track the face of a participant, and failure of the firewire connection that was used to transmit the video data to the computers for analysis. These failures reflect the difficulty of deploying a real-time system for mutual gaze tracking due to the complexity of the necessary hardware and software components. The five remaining pairs of participants for whom complete face and gaze tracking data were available 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

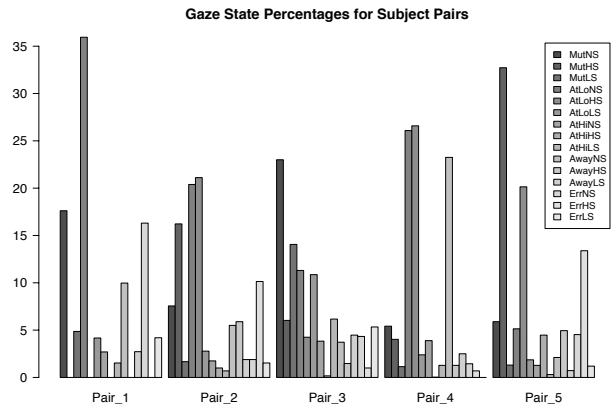

Figure 1. The percentage of time spent in each gaze state by each conversational pair.

- 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 detection errors. While the sound recording levels for the microphones were adjusted for each speaker at the start of an experiment, the microphones still sometimes failed to detect quiet speech. These results most likely have classified some parts of both speakers' conversational turns as times when neither are speaking. 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 additional state information about which participant in the pair was speaking as follows: 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 detection errors. While the sound recording levels for the microphones were adjusted for each speaker at the start of an experiment, the microphones still sometimes failed to detect quiet speech. These results most likely have classified some parts of both speakers' conversational turns as times when neither are speaking. We intend to address this weakness in our data collection equipment in future experiments.

- 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 find 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

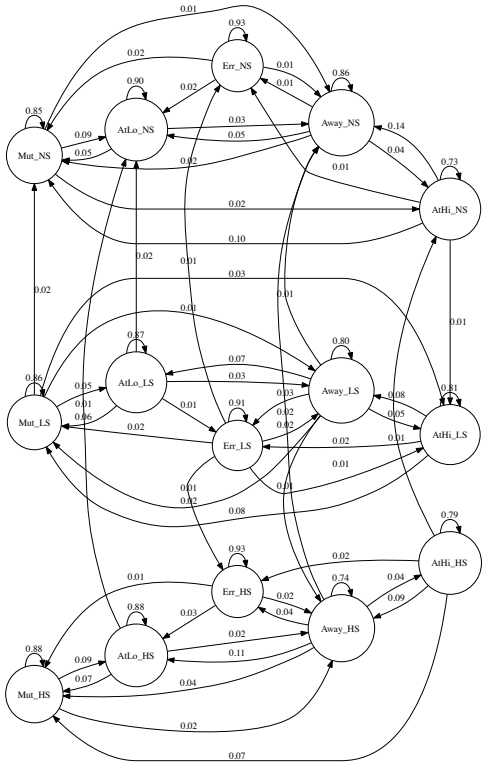

Figure 2. Markov model of the gaze state transitions for all of the conversational pairs.

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.

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 [15], 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 [8, 9, 7] 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 automa-

ton 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 to little data to determine whether this may play a role in the behavior differences observed, but it is an area for further investigation.

CONCLUSIONS

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 this human-human gaze data to produce a robotic gaze controller, we created a Markov model from the data collected and discussed how it captures the gaze behavior dynamics of the human conversational pairs. Additionally, we present preliminary results from an algebraic analysis of the structure of the resulting Markov model and discuss how this type of analysis may be used to computationally investigate qualities of the gaze interaction.

REFERENCES

1. M. Argyle. *Bodily communication*. Routledge, second edition, 1988.
2. 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.
3. 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.
4. 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.
5. J. Call and M. Tomasello. Social cognition. In D. Maestripieri, editor, *Primate Psychology*, pages 234–253. Harvard University Press, Cambridge, MA, 2003.
6. 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.
7. A. Egri-Nagy and C. L. Nehaniv. Sgpdec - hierarchical composition and decomposition of permutation groups and transformation semigroups. .
8. 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.
9. 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.
10. 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.
11. S. M. Hains and D. W. Muir. Infant sensitivity to adult eye direction. *Child development*, 67(5):1940–1951, October 1996.
12. C. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, 1986.
13. H. Kobayashi and S. Kohshima. Unique morphology of the human eye. *Nature*, 387:767–768, 1997.
14. 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.
15. 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.
16. A. S. Laboratories. Mobile eye gaze tracking system. .
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. I. Seeing Machines. faceAPI. .
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.