ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1069

Modeling Gaze Mechanisms for Grounding in HRI

Gregor Mehlmann [1] and **Kathrin Janowski** [1] and **Tobias Baur** [1] and **Markus Häring** [1] and **Elisabeth André** [1] and **Patrick Gebhard** [2]

Abstract. Grounding is essential in human interaction and crucial for social robots collaborating with humans. Gaze plays versatile roles for establishing, maintaining and repairing the common ground. It is combined with parallel modalities and involved in several processes for behavior generation and recognition. We present a uniform modeling approach focusing on the multi-modal, parallel and bidirectional aspects of gaze and their interleaving with the dialog logic.

1 INTRODUCTION

Participants of a human interaction constantly establish, maintain and repair the *common ground* [1]. Disruptions of the common ground mainly arise from misunderstandings, ambiguous utterances, missing joint attention or whenever one of the participants presumes sensory, perceptive or cognitive abilities that the other cannot serve with.

Humans try to ensure the grounding of their information states with the least collaborative effort by exploiting multiple parallel modalities [1, 2]. Gaze is involved in a variety of parallel and bidirectional processes for the generation and recognition of multi-modal behavior. It is aligned with other modalities to ground the speaker and listener roles [3], to elicit and recognize feedback signals [4], to follow and direct the partner's focus of visual attention [5] and to repair disruptions of the common ground by disambiguating ambiguous verbal references regarding the partner's gaze direction [2].

Embedding these manifold roles of gaze in a computational dialog model and synchronizing them with each other and with the dialog management is a complex task. That's why previous research merely studied the roles of gaze for individual sub-concepts of grounding, such as *joint attention* [6], *engagement* [7] and *turn-taking* [8]. Other work investigated modeling languages for multi-modal fusion and dialog logic [9] without specifically focusing on grounding at all.

We go beyond this work with a uniform modeling approach coping with the multi-modal, parallel and bidirectional aspects of gaze that have so far been tackled in isolation. Our approach combines the flexibility and re-usability of hierarchical and concurrent state charts with the expressiveness and declarative nature of logic programming.

2 GROUNDING

We illustrate the roles of gaze for grounding by a collaborative shared workspace application in which a robot instructs the user to move objects with certain shapes, sizes and colors to slots on a table. The user may ask back whenever an instruction is ambiguous or incomprehensible. Both may use gaze, touch and speech to refer or to draw the other's attention to an object. All objects carry markers to identify those the user is looking at with eye-tracking glasses.

[1] Human Centered Multimedia, Augsburg University, Augsburg, Germany.
[2] German Research Center for Artificial Intelligence, Saarbrücken, Germany.

Humans distribute information across multiple modalities, based on the effort and the expressiveness of each channel [2]. They rely on their partners' ability to combine this information to resolve ambiguities. Our robot is able to disambiguate the user's verbal references that caused a disruption of the common ground by regarding his gaze direction. Figure 1 shows an example in which the robot combines the user's gaze direction to disambiguate an ambiguous question.

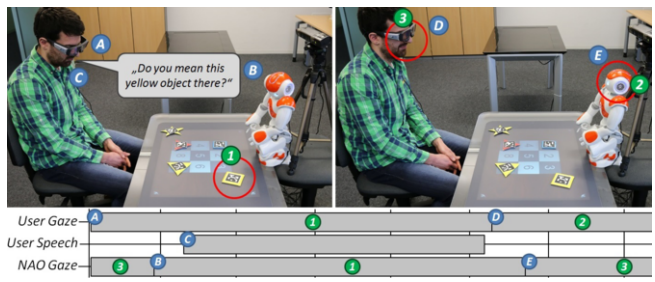

Figure 1. The robot disambiguates the user's speech by considering gaze.

Humans use gaze together with verbal references and pointing gestures to direct their partners' attention to objects or themselves. They follow their partners' gaze to share the point of reference which results in directed gaze [3] or mutual gaze [5]. Both mechanisms are essential for signaling engagement in the joint activity and maintaining the common ground. Our robot is able to draw the user's attention to an object or to himself using any combination of gaze, gestures and speech. As shown in Figure 2, it can follow the user's gaze and touch in order to focus on the same objects and to answer mutual gaze.

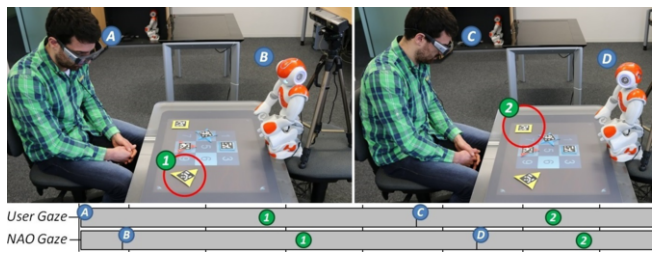

Figure 2. The robot follows the user's gaze resulting in directed gaze.

Gaze serves as a key signal in grounding the exchanges of the speaker and listener roles [3]. Speakers usually look away from their addressees to indicate that they want to keep the floor and look at their partners to pass the floor. Listeners continually produce back-channel signals to signal engagement and understanding [4] using nonverbal cues and verbal remarks. Speakers occasionally initiate mutual gaze to the listener with the aim to elicit feedback [3]. Gaze cues for turn-regulation and feedback eliciting are both essential for grounding but may not be confused and are handled separately by our robot's computational dialog model, as shown in Figure 3.

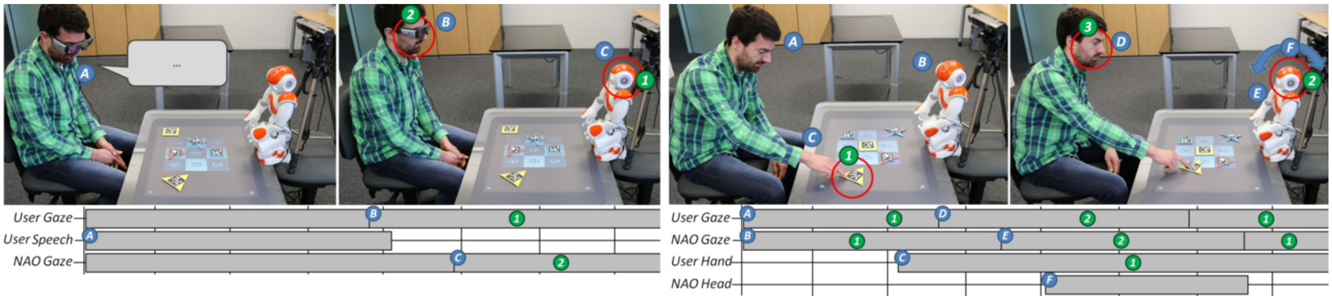

Figure 3. The user is passing the floor to the robot by looking to it after the turn (*left*). The user is eliciting a feedback by the robot during his activity (*right*).

3 REALIZATION

Our realization follows a highly modular approach which is easily reusable and adaptable [10]. Dialog flow and interaction logic are modeled with hierarchical and concurrent state charts that control and synchronize parallel processes for behavior generation and input processing. Knowledge reasoning and multi-modal fusion are realized with a domain specific logic language written in *Prolog*. The robot's expressive behavior is specified in a scripting language which aligns the robot's nonverbal behavior with its verbal statements.

```
[ type: event                                   ]   [ type: event                        ]
[ ...                                           ]   [ sent: track                        ]
[        [ type: dialact                      ] ]   [ mode: gaze                         ]
[        [ cat: seeking                       ] ]   [ from: 121                          ]
[        [ fun: choice                        ] ]   [ life: 43                           ]
[ data:  [        [[ type: object          ]] ] ]   [ time: 62946                        ]
[        [ data:  [[ data: [ col: green  ] ]] ] ]   [ conf: 1.0                          ]
[        [        [[       [ shp: square ] ]] ] ]   [        [[ type: mark   ]]          ]
[        [        [[ type: object          ]] ] ]   [        [[ name: 1      ]]          ]
[        [        [[ data: [ col: yellow ] ]] ] ]   [        [[ conf: 0.88   ]]          ]
[        [        [[       [ loc: [ ]    ] ]] ] ]   [ data:  [ ...            ]          ]
                                                    [        [[ type: mark   ]]          ]
                                                    [        [[ name: 9      ]]          ]
                                                    [        [[ conf: 0.001  ]]          ]
```

Figure 4. A dialog act (*left*) and a gaze event (*right*) as feature structures.

User input events are interpreted and then asserted as typed feature structures to the fact base. As shown in Figure 4, they carry modality-independent features, such as time stamps and confidence values, as well as modality-specific information, such as gaze distributions and dialog acts. Eye-tracking errors when the user blinks or rolls the eyes are reduced by computing the gaze distributions from a number of past frames. The user's speech is processed by a semantic parser and translated into dialog acts of the *DiAML* scheme, such as propositional or choice questions. The parser is able to extract the objects' features and spatial relations from the users' object descriptions.

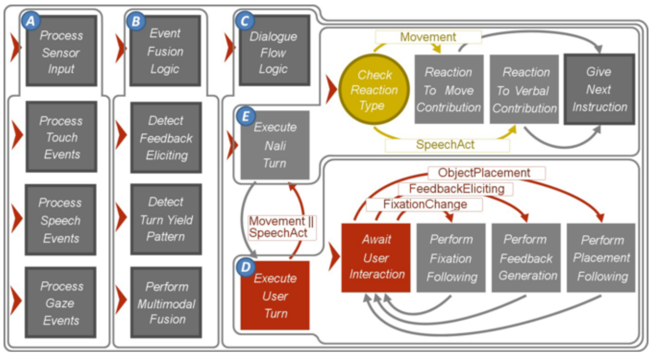

Figure 5. The parallel and hierarchical main state charts of the scenario.

Figure 5 shows an overview of the interaction model consisting of several hierarchical and parallel state charts synchronized via the fact base. A first state chart is processing touch, speech and gaze distribution events in three parallel processes (Fig. 5 Ⓐ). Multi-modal fusion and the detection of turn-taking and feedback eliciting behavior takes place in a second state chart (Fig. 5 Ⓑ). A third state chart is modeling the dialog flow for the user's and the robot's turn (Fig. 5 Ⓒ). In the user's turn (Fig. 5 Ⓓ), the robot either follows the user's gaze fixations and object placements or performs feedback while waiting for a contribution of the user. When the user moves an object to a field or speaks an utterance and gives a turn-yielding signal, these processes are immediately interrupted and the turn is assigned to the robot (Fig. 5 Ⓔ), in which it checks the type of the user's contribution and performs an adequate reaction before giving the next instruction.

4 SUMMARY

Our uniform modeling approach combines the flexibility and reusability of hierarchical and parallel state charts with the expressiveness and declarative nature of logic and is mastering the multi-modal, parallel and bidirectional aspects of gaze that have so far been tackled in isolation. First tests with users revealed that the implemented gaze mechanisms for speech disambiguation and joint visual attention enable fluent and pleasant human-robot dialogs demonstrating the potential of our bidirectional gaze model for grounding. In a next step will use the current system for a systematic evaluation of our modeling approach and a profound investigation of the roles of gaze mechanisms for grounding and their interplay in different scenarios.

REFERENCES

[1] H. H. Clark, *Using Language*. Cambridge University Press, 1996.
[2] S. Oviatt, *The Human-Computer Interaction Handbook*, 2008, ch. Multimodal Interfaces, pp. 413–432.
[3] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.
[4] V. H. Yngve, "On getting a word in edgewise," in *Regional Meeting of the Chicago Linguistic Society*, 1970, pp. 657–677.
[5] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge University Press, 1976.
[6] C.-M. Huang and A. Thomaz, "Effects of responding to, initiating and ensuring joint attention in human-robot interaction," in *ROM-AN*, 2011, pp. 65–71.
[7] A. Holroyd, C. Rich, C. Sidner, and B. Ponsler, "Generating connection events for human-robot collaboration," in *RO-MAN*, 2011, pp. 241–246.
[8] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *TIIS*, vol. 1, no. 2, pp. 12:1–12:33, 2012.
[9] D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J. F. Ladry, "Fusion engines for multimodal input: A survey," in *ICMI*, 2009, pp. 153–160.
[10] G. Mehlmann and E. André, "Modeling multimodal integration with event logic charts," in *ICMI*, 2012, pp. 125–132.