Exploring a Model of Gaze for Grounding in Multimodal HRI

Gregor Mehlmann
Human Centered Multimedia
Augsburg University, Germany
mehlmann@hcm-lab.de

Kathrin Janowski
Human Centered Multimedia
Augsburg University, Germany
janowski@hcm-lab.de

Markus Häring
Human Centered Multimedia
Augsburg University, Germany
haering@hcm-lab.de

Tobias Baur
Human Centered Multimedia
Augsburg University, Germany
baur@hcm-lab.de

Patrick Gebhard
German Research Center for
Artificial Intelligence
patrick.gebhard@dfki.de

Elisabeth André
Human Centered Multimedia
Augsburg University, Germany
andre@hcm-lab.de

ABSTRACT

Grounding is an important process that underlies all human interaction. Hence, it is crucial for building social robots that are expected to collaborate effectively with humans. Gaze behavior plays versatile roles in establishing, maintaining and repairing the common ground. Integrating all these roles in a computational dialog model is a complex task since gaze is generally combined with multiple parallel information modalities and involved in multiple processes for the generation and recognition of behavior. Going beyond related work, we present a modeling approach focusing on these multi-modal, parallel and bi-directional aspects of gaze that need to be considered for grounding and their interleaving with the dialog and task management. We illustrate and discuss the different roles of gaze as well as advantages and drawbacks of our modeling approach based on a first user study with a technically sophisticated shared workspace application with a social humanoid robot.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; D.2.11 [**Software Engineering**]: Software Architectures

General Terms

Algorithms, Languages, Design

Keywords

Human-Robot Interaction; Multi-Modal Fusion; Dialog Modeling

1. INTRODUCTION

The participants of a human interaction constantly establish, main- tain and repair the *common ground*, which Herbert Clark defines as "the set of knowledge, beliefs and suppositions that the partici- pants believe they share" [7]. Disruptions of this common ground mainly arise from misunderstandings, ambiguous utterances, miss- ing attention or whenever one of the participants presumes sensory, perceptive or cognitive abilities that the other cannot serve with.

ICMI '14 November 12 - 16 2014, Istanbul, Turkey

http://dx.doi.org/10.1145/2663204.2663275.

Humans seek to ensure the grounding of their information states with the *least collaborative effort* [9, 7, 8]. To achieve this, they usually exploit multiple parallel information modalities. Among those, gaze and its interplay with the other modalities plays versatile important roles for grounding. Rendering account of and serving these roles is essential for social humanoid robots that interact naturally with humans in collaborative shared workspaces, as in the application that we describe and investigate in this paper.

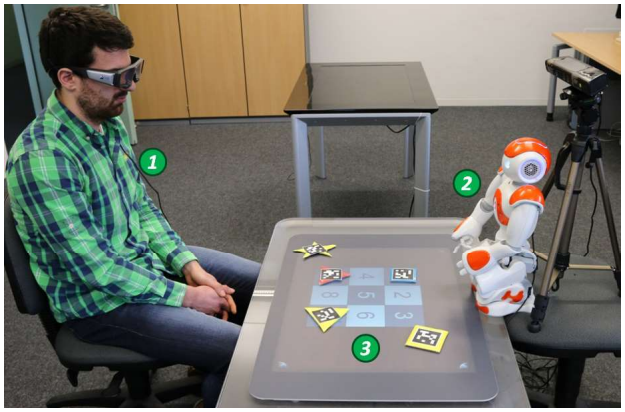

Figure 1: Our shared workspace setting with the NAO[1] robot.

Figure 1 shows the scenario of our application, a collaborative game between a human user (Fig. 1 ①) and the humanoid *NAO*[1] robot *Nali* (Fig. 1 ②) on a shared workspace realized with a *Microsoft*[2] surface table (Fig. 1 ③). The user wears *SMI* eye tracking glasses[3] and a microphone for speech recognition. The puzzle pieces have distinguishable features such as a shape, size, color and position. They are marked on both sides, to track their position and to recognize the pieces the user is looking at via marker tracking on the video of the glasses. The robot is supposed to facilitate a sorting task by instructing the user to move puzzle pieces into certain puzzle slots. Both interaction partners may use any combination of gaze, pointing gestures and speech to multi-modally refer to the objects, to regulate the speaker and listener roles and to draw the other participant's attention to the objects or even themselves.

Resembling human interaction, both partners exploit a variety of gaze mechanisms during this collaboration. Thereby, gaze is com-

[1]http://www.aldebaran-robotics.com/
[2]http://www.microsoft.com/
[3]http://www.smivision.com/

bined with multiple parallel information modalities and involved in a variety of parallel and bi-directional processes for the generation and recognition of multi-modal behavior. Gaze cues are aligned with verbal contributions to ground the speaker and listener roles [22, 15, 10, 28] in a fluent conversation. Both partners use gaze to continually give and elicit feedback signals [35, 5] and to follow and direct the other's attention to objects in the environment or to themselves [3, 1, 19]. Finally, disruptions of the common ground caused by ambiguous verbal referring expressions are disambiguated by considering the partner's gaze [23, 24, 30].

Embedding all these parallel and bi-directional roles of gaze and synchronizing them with each other and the task management in a uniform dialog model is a complex task. Previous research on dialog models in HRI considered these roles of gaze mainly in isolation. We present a modeling approach focusing on the multi-modal, parallel and bi-directional aspects of gaze and validate it in our shared workspace application. We evaluate our approach in an experimental study, discuss the results in detail and formulate recommendations for future efforts in this research area.

2. RELATED WORK

During the past years, a number of researchers studied several roles of gaze as conversational coordination mechanism that only represent individual aspects of the grounding process in HRI .

Most of this research studied the roles of gaze for *joint attention*, the "ability to follow the direction of the gaze and gestures of others in order to share a common point of reference" and the "use of gestures and eye contact to direct others' attention to objects, to events, and to themselves" [19]. Huang and Thomaz found that a robot that continually ensures the user's joint attention is perceived as more natural [14] while Staudte and Crocker found that exploiting the speaker's focus of visual attention via gaze-following has a positive influence on utterance comprehension to anticipate, ground, and disambiguate spoken references [30]. Mutlu *et al.* investigated the importance of joint attention for collaborative settings in which human and robot are performing joint actions [21] on a shared workspace. Mutlu *et al.* also studied gaze mechanisms that are used for the regulation of the dialog structure and for grounding different speaker and listener roles [20] in human-robot dialog.

Rich *et al.* studied computational models of recognizing and generating *engagement* [25, 13], a process "by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" [29]. This work captures the role of directed gaze and mutual gaze for the connectivity status of interaction partners but not the role of gaze for the disambiguation of speech or the use of gaze for eliciting feedback.

Other related research focuses on generic modeling languages for multi-modal fusion or dialog logic for multi-modal user interfaces [16], embodied conversational agents [6, 32] and human-robot interaction [31] but does not focus on grounding or similar concepts or sub-concepts, such as joint attention or engagement.

The related efforts either studied individual roles of gaze for grounding in human-robot interaction in isolation or presented a generic modeling approach without focusing on grounding at all. The work in this paper draws on valuable ideas from this related work but goes beyond it because it combines those individual aspects of grounding in a novel uniform modeling approach.

3. GROUNDING IN HRI

In this section we systematically review the roles of gaze that contribute to grounding in human-robot interaction and illustrate them based on our representative application shown in Figure 1.

3.1 Disambiguating Speech

Humans distribute information across different modalities, depending on the effort and the expressive power of each channel and rely on their partners' ability to combine this information in order to resolve ambiguities [23, 24]. Isolated verbal referring expressions can be ambiguous and thus cause a disruption of the common ground. In this case, a listener usually tries to combine the speaker's verbal statement with his gestures and eye gaze into an unambiguous interpretation before asking for a clarification. Referential gaze in speech typically precedes the corresponding linguistic reference by approximately 800-1000 milliseconds and people look at what they hear after about 200 milliseconds [18, 12].

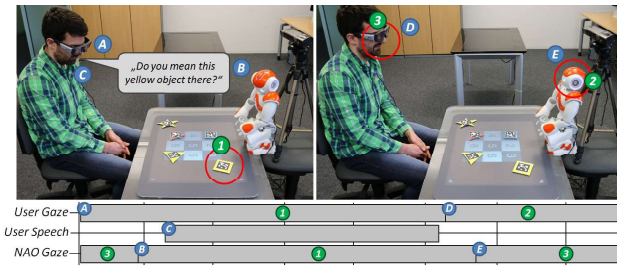

Figure 2: Producing an object reference with speech and gaze.

In our application, the robot is able to combine an ambiguous verbal referring expression with the user's eye gaze during this time window. For example, in Figure 2 the user starts looking (Fig. 2 Ⓐ) at an object (Fig. 2 ①) and the robot follows the user's gaze (Fig. 2 Ⓑ). The user asks the clarification question *"do you mean this yellow object there?"* (Fig. 2 Ⓒ) and afterwards looks (Fig. 2 Ⓓ) to the robot (Fig. 2 ②) in order to yield the floor. Although there are several objects on the table that match the verbal referring expression, the robot infers that the user referred to the large yellow triangle by considering the user's gaze direction at the right time.

3.2 Joint Visual Attention

During collaborative activities, humans seek to direct their partners' attention to objects in the environment or to themselves [8]. Beside verbal references and pointing gestures they often use gaze and a combination of these modalities to reach that goal [9, 4, 26]. They also follow their partners' gaze to share their point of reference [19] which usually results in directed gaze and joint attention to an object [15]. In this way they signal their partners that they are engaged in the joint activity and able to identify the referred objects. When the partners pay attention to each other they perform mutual gaze [1]. Both gaze mechanisms for joint attention are essential for maintaining the common ground.

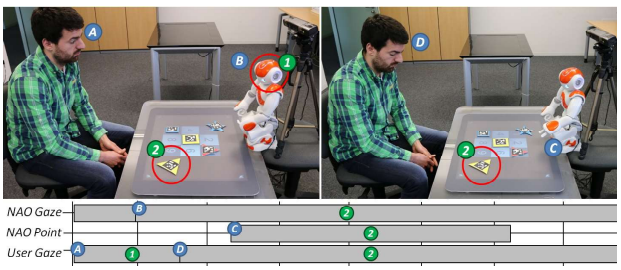

Figure 3: Drawing the user's attention to an object using gaze.

The robot in our application is able to draw the user's attention to

an object which is relevant for the next step in the task or to itself whenever it starts an instruction or clarification. Figure 3 shows such a situation in which the user is looking (Fig. 3 Ⓐ) at the robot (Fig. 3 ①) expecting an instruction. The robot starts looking (Fig. 3 Ⓑ) at an object (Fig. 3 ②) and points to it (Fig. 3 Ⓒ) a few moments later. The user immediately follows the robot's gaze (Fig.3 Ⓓ) which leads to directed gaze and attention to the object.

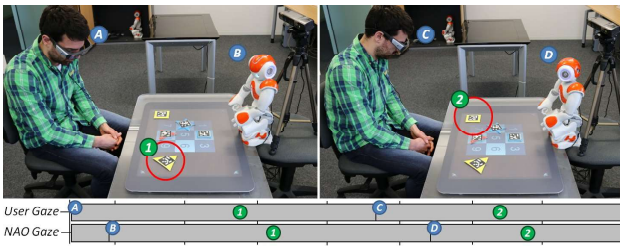

Figure 4: Following the user's gaze to focus to the same objects.

Analogous to the direction of attention, the robot can also pay constant attention to the user's gaze shifts and manipulations of the objects on the workspace. It is able to follow the user's gaze and to focus on the objects that the user looks or points to while it answers with mutual gaze whenever the user looks at the robot. In Figure 4 the user looks (Fig. 4 Ⓐ) at an object (Fig. 4 ①) and a few moments later the robot follows the user's gaze by looking (Fig. 4 Ⓑ) at the same object. Next, the user moves his gaze (Fig. 4 Ⓒ) to another object (Fig. 4 ②) and again the robot follows the user's gaze (Fig. 4 Ⓓ) to the second object for joint attention.

3.3 Regulating Turn-Taking

A fluent conversation requires successful regulation mechanisms for grounding the speaker and listener roles [15, 10]. Thereby, gaze direction serves as a key signal in managing or inhibiting the exchanges of these roles. Speakers usually look away from their addressees to indicate that they want to keep the floor and look at one of their partners to pass the floor [22]. The exchange of turns can be delayed if a contribution does not end with gaze at a partner [15].

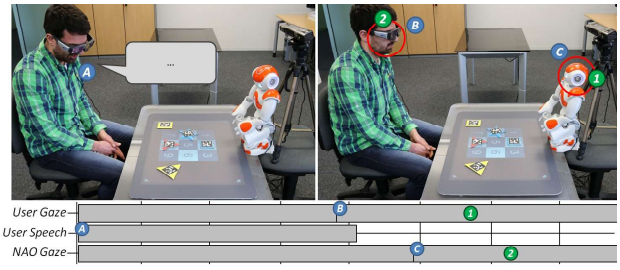

Figure 5: Passing the floor to the robot after the contribution.

In our application, the robot avoids mutual gaze while giving an instruction or clarification and looks to the human partner to pass the floor when finishing its turn. In return, it is able to detect the user's end of speech and whenever the user looks to the robot via a marker on its chest. Both events in close temporal alignment are the signal for the robot that the user wants to pass the floor. Such a situation is shown in Figure 5 in which the user is asking a question (Fig. 5 Ⓐ) and afterwards looks (Fig. 5 Ⓑ) at the robot (Fig. 5 ①) to yield the turn. The robot (Fig. 5 Ⓒ) then answers with mutual gaze (Fig. 5 ②) to signal that it is going to take the turn now.

The interaction partners continually produce back-channel signals to let the partner know that they are still engaged and understand what has been said and done [35] to ground their information states. This feedback is provided by the listener via nonverbal cues such as head nods or short verbal statements while the partner is speaking or performing an activity. In return, speakers occasionally perform a short glance of mutual gaze to the listener with the aim to elicit feedback cues at specific points in time [15, 5]. Although the gaze cues for feedback eliciting and turn-yielding might look very similar at first sight, it is very important not to confuse these signals and to handle them differently in the dialog model.

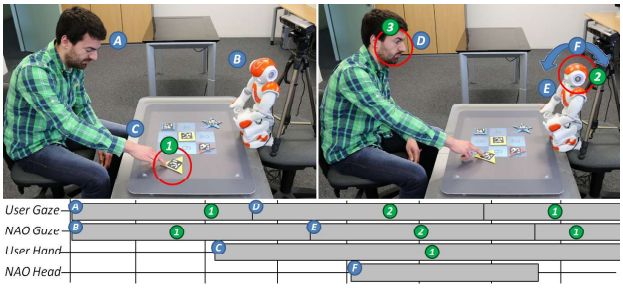

Figure 6: The user elicits a head nod feedback from the robot.

In this work we are not further investigating the role of feedback eliciting for grounding but of turn-regulating gaze signals. Therefore, in our application, the robot is able to recognize the user's behavior pattern that are supposed to elicit feedback, to distinguish them from turn-yielding cues and to react with an adequate feedback signal. Figure 6 shows such a situation in which both the user (Fig. 6 Ⓐ) and the robot (Fig. 6 Ⓑ) look at an object (Fig. 6 ①) before a few moments later the user starts moving the object (Fig. 6 Ⓒ) and looking at the robot (Fig. 6 Ⓓ) nearly simultaneously. The robot reacts to this behavior pattern with mutual gaze (Fig. 6 Ⓔ) and a feedback signal in form of a head nod (Fig. 6 Ⓕ).

4. REALIZATION

The dialog flow and interaction logic in our application are modeled with a *Sceneflow* [11, 17]. This is a hierarchical and concurrent state chart which is used to control and synchronize multiple parallel processes modeling the robot's behavior and input processing on different abstraction levels. This model decomposition divides the responsibilities in the model so that each process resembles a single role of gaze, a fusion process or a dialog phase. The robot's context knowledge and user inputs are represented as feature structures in a *Prolog* [34] fact base. The fact base defines domain specific logic predicates that are called from within the Sceneflow. This logic is used for multi-modal fusion, knowledge reasoning and for the exchange of information between parallel processes. This highly modular modeling approach makes already modeled state chart components and predicates easily reusable and adaptable.

4.1 Input Processing

The movement of puzzle pieces and other touch events are directly forwarded from the surface table to the fact base. All other user input events are first processed and interpreted by the *SSI* framework [33]. The resulting interpretations are then asserted as events to the fact base. They carry modality-independent features, such as timestamps and confidence values, and modality-specific semantic information, such as gaze distributions and dialog acts.

For gaze recognition we rely on Algorithm 1 that reduces the influence of recognition errors that arise whenever the user blinks,

Algorithm 1 Compute Fixation Events from Input Streams

procedure TRANSFORM($Video\ v$, $Point\ g$, $List\ l$)
  $\Delta_{max} \leftarrow \sqrt{v_w^2 + v_h^2}$
  if $g_x \geq 0 \wedge g_x \leq v_w \wedge g_y \geq 0 \wedge g_y \leq v_h$ **then**
    for all $m \in l$ **do**
      if $visible(m)$ **then**
        $\Delta_m \leftarrow \sqrt{(g_x - m_x)^2 + (g_y - m_y)^2}$
      else
        $\Delta_m \leftarrow min(\delta \cdot \Delta_m, \Delta_{max})$
      end if
      $\Phi^*_m \leftarrow (1 - (\Delta_m / \Delta_{max}))^\phi$
      $\Phi_m \leftarrow \sigma \cdot \Phi_m + (1 - \sigma)\Phi^*_m$
    end for
  end if
end procedure

rolls his eyes or is shortly distracted. In each frame we compute the distances of all recognized markers on the puzzle pieces and the robot's chest to the user's gaze position (Alg. 1 ⑥). Then we compute a fixation confidence for each marker based on this distance and the respective confidences in the past few frames (Alg. 1 ⑪). We assume the user is looking at the environment if all confidences are below a certain threshold. Every couple of frames we produce an event carrying this fixation confidence distribution and assert it to the fact base where it is processed by the logic.

```
type:  event, # This feature structure is an event
sent:  ssiv2, # The sender is the SSI framework
mode:  eyegaze, # A marker probability distribution
dist:  121, # The event has started 121 ms ago
life:  43, # The event's total lifetime is 43 ms
time:  62946, # The VSM framework arrival time
conf:  1.0, # The event's confidence value is 1.0
data:  ①  [ [ type: marker, name: 1, conf: 0.88 ], ..., [ type: marker, name: 9, conf: 0.001 ] ]
```

Figure 7: A gaze distribution event as a feature structure.

Figure 7 shows the representation of a gaze distribution event containing the confidence values for all markers (Fig. 7 ①). Within our interaction model these events are then used to realize gaze following, recognize when the user yields the turn and elicits feedback and for the disambiguation of speech. For this reason, they are hold in the fact base for some seconds so that they can be combined with the user's speech input that usually needs longer to be interpreted.

The user's speech is processed by a semantic parser plug-in of SSI which is based on the *Microsoft*® speech platform[4]. Utterances are translated into abstract dialog acts with communicative functions from the *information seeking* category of the *DiAML* classification scheme[5]. *Set questions* ask for a decision between a number of puzzle pieces, *choice questions* prompt a decision between a list of candidates, while *propositional questions* seek confirmation for exactly one proposed puzzle piece. Figure 8 shows the representation of the choice (Fig. 8 ①) question *"the green triangle or the yellow object there?"* with the included location referent (Fig. 8 ④). Our robot can resolve unimodal referring expressions such as *"the green triangle"* (Fig. 8 ②) using an algorithm similar to that described by Ros *et al.* [27]. It determines a set of puzzle pieces on the table that match the list of features from the user's verbal description and computes the optimal set of discriminating features to correct the user with an unambiguous answer.

[4]http://msdn.microsoft.com/library/hh361572.aspx
[5]http://semantic-annotation.uvt.nl

```
type:  event,
...,   # General event properties, i.e. timestamp, modality, etc.
data:  [ type: dialog # A dialog act feature structure
         category: seeking # Information seeking category
         function: choice ① # Choice question function
         contents: [ ② [ type: description # The 1st object
                         data: [ color: green # is green and
                                 shape: triangle # a triangle ] ],
                     ③ [ type: description # The 2nd object
                         data: [ color: yellow # is yellow
                                 location: reference ④ ] ] ] ]
```

Figure 8: A choice question dialog act as a feature structure.

```
1  fuseWithGaze(SpeechEvent, FusedAct) :-
2    // Get The Speech Act Of The Speech Event
3    getSpeechAct(SpeechEvent, SpeechAct),
4    // Get The Speech Act's Object Description
5    getDescription(SpeechAct, Description),
6    // Get The Gaze Fixation During The Speech
7    getBestFixation(SpeechEvent, GazeFixation),
8    // Update The Object Description With Gaze
9    has(Description, 'location', 'reference'),
10   set(Description, 'fixation', GazeFixation).
11   // Update The Object Description Of The Speech
12   // Act And Return The Fused Multi-Modal Action
13   update(SpeechAct, Description, FusedAct).
14
15 fuseWithGaze(SpeechEvent, SpeechAct) :-
16   getSpeechAct(SpeechEvent, SpeechAct).
```

Figure 9: The predicate combining user's speech and gaze.

If the utterance contains location referents such as *"there"* or *"here"* (Fig. 8 ③), then we additionally consider the user's referential gaze to resolve this multi-modal referring expression using the logic predicates shown in Figures 9 and 10. We determine the focused object based on the gaze events during a time window relative to the user's utterance (Fig. 9 ③-⑦ and Fig. 10 ①-⑨) and add the reference to the object description (Fig. 9 ⑨-⑬).

```
1  getBestFixation(SpeechEvent, GazeFixation) :-
2    // Get The Time Interval Of The SpeechEvent
3    start(SpeechEvent, Start),
4    end(SpeechEvent, End),
5    // Infer All Gaze Events In The Time Window
6    findall(EyeGaze, (during(EyeGaze, SpeechEvent),
7      mode(EyeGaze, 'eyegaze')), GazeEventList),
8    // Compute The Object With The Most Fixations
9    fixationMajority(GazeEventList, GazeFixation).
```

Figure 10: Predicate inferring eye gaze aligned to speech.

4.2 Interaction Modeling

Figure 11 shows an overview of our application's interaction model consisting of three parallel state charts that are exchanging information via the fact base. The first state chart (Fig. 11 Ⓐ) is used for monitoring the system time and for retracting outdated events that are not needed any more. The second state chart (Fig. 11 Ⓑ) is processing touch, speech and gaze distribution events in three parallel processes. They are updating the fact base positions

of objects that have been moved on the surface table and produce higher-level events carrying the user's last speech act and discrete fixation change events from the continuous gaze distribution events. These events are consumed by the processes in a third state chart (Fig. 11 Ⓒ) that is modeling higher-level event processing. It enfolds a process for the multi-modal fusion of gaze and speech as well as processes for the detection of turn-taking and feedback eliciting signals. A last state chart (Fig. 11 Ⓓ) is modeling the dialog flow in two separate states for the user's and the robot's turn.

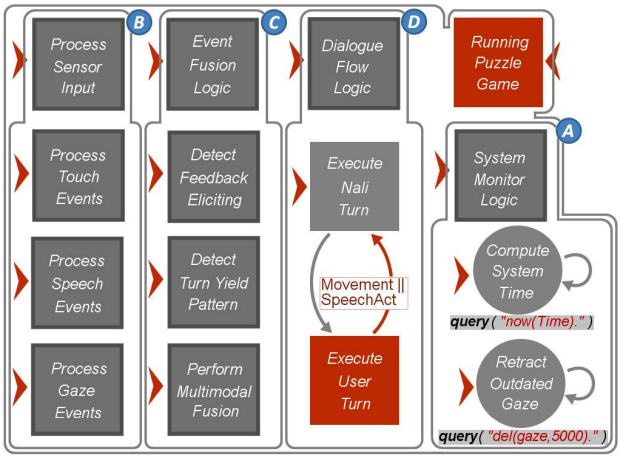

Figure 11: The high-level view of the application's sceneflow.

Figure 12 shows these processes modeling the dialog flow in more detail. When the user has the turn, the robot either follows the user's gaze fixations and object placements or performs feedback while waiting for a turn-yielding event from the process detecting the relevant behavior pattern. When the user finishes his turn by moving an object to a field or speaking an utterance and looking to the robot, then these processes are immediately interrupted and the turn is assigned to the robot. The robot then checks the type of the user's contribution and performs an adequate reaction, such as answering a question or giving the next instruction.

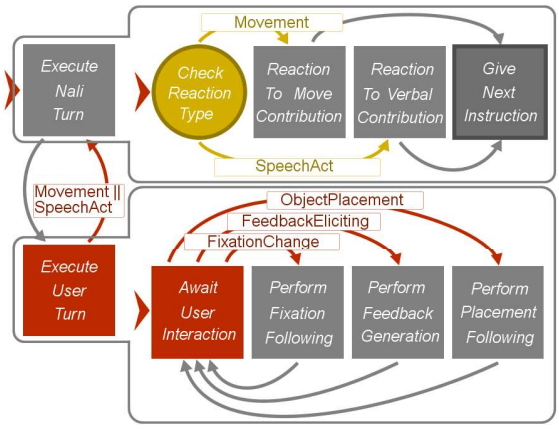

Figure 12: The sceneflow modeling the dialog flow logic.

The robot's verbal and nonverbal behavior is specified in a multimodal *Scenescript*, resembling a movie script with dialog utterances and stage directions for gestures, postures, and facial expressions. Individual *Scenes* from this script are executed from the nodes in the Sceneflow and then scheduled on the robot to generate the behavior. Figure 13 shows an example of such a script and the corresponding scheduling. It shows that nonverbal behavior can be executed either in parallel with speech or sequentially. This specification method allows us to generate the robot's gaze behavior for directing and following the user's gaze and yielding the speaker role. In addition, is allows us to easily align the robot's gaze and other nonverbal behavior such as feedback cues and pointing gestures with its speech in a very intuitive but also effective way.

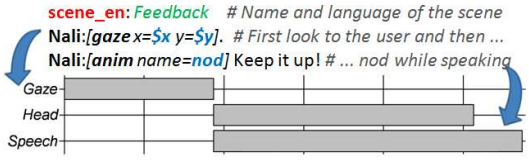

Figure 13: A feedback scene and the corresponding schedule.

5. EVALUATION

We evaluated our application and the underlying computational dialog model in order to answer two questions. First, we wanted to validate our modeling approach as described in Section 4 and check to what extent our computational model covers the aspects of gaze for grounding presented in Section 3. Second, we wanted to find out how the different roles of gaze influence the subjective perception and the efficiency of the interaction in our application. Thereby, we were interested in the effect of two independent variables that were manipulated in a $2x2$ within-subjects design:

- *Referential gaze* with the levels O^+ and O^-: In condition O^+, the robot always looked at the puzzle pieces it was referencing during the instructions and was able to disambiguate the user's ambiguous clarification requests by considering the user's referential gaze. In condition O^-, the robot did not look at any particular puzzle piece on the shared workspace to facilitate object grounding during instructions. Furthermore, it was not able to resolve the user's multi-modal references by considering the user's referential gaze during the user's clarification requests.

- *Social gaze* with the levels U^+ and U^-: In condition U^+, the robot tried to establish mutual gaze with the user, reacted to turn-yielding gaze cues and followed the user's gaze or hand movements to signal attentiveness. It reacted to speech acts and placement attempts after the user looked at it in order to pass the floor, or after a maximum delay of 3.1 seconds [25]. In condition U^-, the robot did not react to the user's gaze for turn-taking or attention following. Instead, it simulated ideal gaze behavior by looking at the shared workspace for 70% of the time and at the user otherwise as suggested by the literature [2]. It waited for 1 second before taking the turn and answering a question [25].

Accordingly, we implemented four variations of the dialog model and conducted a within-subjects experiment with each subject participating in all four conditions in a randomized order. A total of 13 subjects, 3 female and 10 male, in age from 19 to 34 years ($M = 26.4$, $SD = 4.41$) participated in the experiment. Most of them were students or researchers in Computer Science and Multimedia. For each participant and condition, we recorded the speech and gaze fixation events, the raw input streams from which they were derived and the video and audio from an external camera which captured the whole scene and synchronized these data streams prior to the annotation. In total, we collected 3 hours

and 8 minutes of material per channel. Afterwards, multivariate analysis of variance (*MANOVA*) and univariate analysis of variance (*ANOVA*) tests were conducted to assess differences on mean user scores across the two categories of referential and social gaze.

5.1 Objective Measures

In order to measure the efficiency of the human-robot dialog, the system automatically logged the duration as well as the number of placement attempts and clarification questions that the users asked. Results of a *Two-Way Repeated Measure MANOVA* revealed no significant multivariate effect of referential gaze or social gaze on the three dependent variables we investigated as objective measurements. However, we observed a univariate effect for referential gaze on all three dependent variables. Overall, the efficiency was significantly improved by adding gaze for disambiguation. The participants asked fewer clarification questions ($F = 11.636$, $p = 0.005$), the dialog was shorter ($F = 11.128$, $p = 0.006$) and less placement attempts were made ($F = 9.794$, $p = 0.009$) by the users when referential gaze was recognized by the robot.

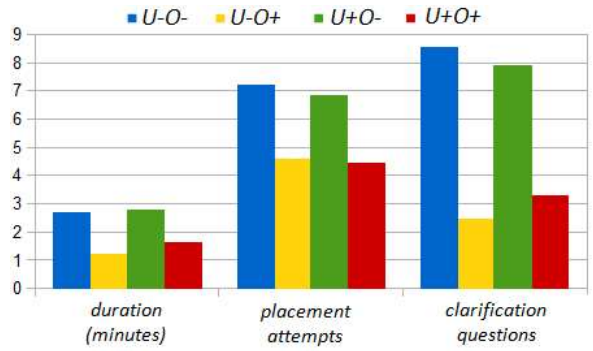

Figure 14: Results of the objective task performance measures.

As shown in Figure 14, the O^+ conditions took only about half of the time that was used in the O^- conditions. The number of placement attempts was reduced by about one third and the number of clarification questions even by two thirds.

5.2 Subjective Measures

After each condition the subjects filled in a questionnaire, rating several aspects of their interaction experience and the robot's behavior in the condition on 7-point Likert scales. Depending on the question, these ranged either from 1 for *never* to 7 for *always* or from 1 for *complete disagreement* to 7 for *full agreement*. As shown in Figure 15, their answers confirmed the trend that had already been observed for the objective measures, demonstrating again an advantage of the O^+ conditions over the O^- conditions.

Results of a *Two-Way Repeated Measure MANOVA* revealed that there were no significant multivariate effects for referential or social gaze on the set of dependent variables. Apart from two exceptions, all univariate effects for referential gaze were, however, significant and showed that referential gaze was associated with positive user ratings of the robot. Participants stated that the robot appeared more natural ($F = 8.348$, $p = 0.014$), more interested ($F = 13.5$, $p = 0.003$) and more understanding ($F = 14.990$, $p = 0.002$). They also thought that the dialog with the robot was more fluent ($F = 8.794$, $p = 0.012$). Furthermore, they believed more often that the robot was following their hand movements ($F = 6,464$, $p = 0.026$) as well as their gaze movements ($F = 14.495$, $p < 0.002$) and that it was more often trying to establish eye contact with them ($F = 27.307$, $p < 0.0005$).

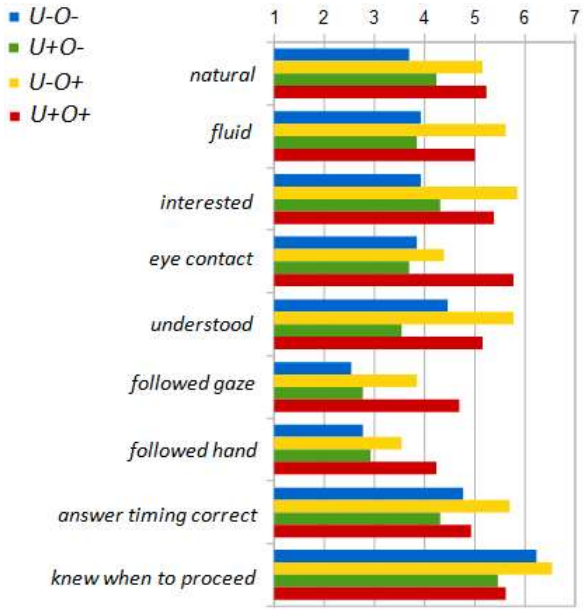

Figure 15: Results of the subjective questionnaire ratings.

Unfortunately, we obtained two significant univariate effects for social gaze that reflected some synchronization issues between the user and the robot. In both cases, U^- scored significantly better than U^+. For U^+, participants stated that they thought less often that the robot responded to their questions at the right point in time ($F = 7.870$, $p = 0.016$). They also said that they knew less often when they were supposed to continue ($F = 4.697$, $p = 0.051$).

There was also a significant interaction effect between referential and social gaze for "The robot tried to establish eye contact with me." ($F = 6.936$, $p = 0.022$). The use of referential gaze had a positive effect on the rating of this item. The additional use of social gaze then led to a further increase of this improvement.

6. DISCUSSION

Our evaluation revealed that the robot's dialog behavior was rated better when it made use of referential gaze and considered the user's referential gaze. In line with the subjective assessments, the objective measurements demonstrated that task performance was improved by referential gaze as well. The users indeed noticed that the robot was following their gaze towards the focused objects and also believed that the robot understood them better when it was able to disambiguate speech with gaze. The results of both subjective and objective measures confirmed our expectations that the gaze mechanisms for speech disambiguation and joint visual attention, realized with our modeling approach as described in Section 4, had a positive influence on user experience and performance.

However, our expectations for the effect of social gaze were disappointed. The key difference between the conditions with and without social gaze, and also the main difference in the users' assessment of those conditions, lied in the robot's turn-taking behavior, or more precisely in the robot's reaction to the users' gaze signals for yielding the turn. We had hoped that responding to the user's turn-regulating gaze cues in the U^+ conditions would improve the interaction compared to the idealized turn-taking in the U^- conditions. However, the opposite was the case, as the conditions without social gaze were often perceived as equal to or even better than their counterparts. Although the robot was not adapting

to any gaze cues the user employed for turn-taking, the timing of the robot's responses was considered to be more appropriate.

6.1 Corpus Analysis

In order to explain the results of our experiment, we analyzed our video recordings of the three conditions U^-O^+, U^+O^- and U^+O^+ for all thirteen participants. We annotated the robot's and users' speech behavior as well as the users' gaze behavior which accompanied a total of 209 clarification requests during 2 hours and 14 minutes of video corpus. This analysis revealed that the users yielded their turn using mutual gaze after 78% of the questions. For this purpose, they looked at the robot shortly before (60%) or after (18%) they finished speaking. For about 20% of the questions, they did not look towards the robot at all. While this observation is mainly in line with previous studies and our own expectations, we found that 83% of the questions which were not accompanied by a turn-yielding gaze at the robot contained location referents which required additional information to be resolved. For example, referential gaze had to be considered to disambiguate questions such as *"This one here?"* or *"Do you mean the one over there?"*.

In contrast to that, for the great majority of unambiguous questions, the users gazed towards the robot as a turn-yielding signal as we had expected. A second interesting observation was that those users who made use of pointing gestures tended to employ gaze for the purpose of turn-regulation more frequently. Obviously, the issues with social gaze were not caused by technical limitations of the modeling approach. Rather they can be ascribed to some misconceptions included in the robot's behavior model. These issues could be resolved by giving up specific assumptions regarding the use of gaze cues for turn-regulation that did not hold for all users.

6.2 Interpretation

A reason for the minor role of social gaze might be the emphasis on the task-oriented nature of the dialog. The users knew that they were supposed to collaboratively solve a task as opposed to engaging in a purely social interaction. They were also aware that the robot was able to disambiguate verbal utterances by considering their gaze behavior. Therefore, the users might have devoted a lot of their attention to the production of non-ambiguous references using both speech and gaze, thus focusing on the functional aspect of gaze rather than making use of gaze for turn-regulation. The effect might have been reinforced by the structure of the dialog that did not require sophisticated turn-taking behaviors. An interesting topic for a follow-up experiment would be to investigate the effect of social gaze during interactions with a much more social component, such as discussing photos displayed on the surface table.

Another factor to consider is the robot's inability to detect pointing gestures. To fill the role of pointing gestures, some users might have relied more heavily on gaze behaviors. As described in [25], a common pattern for directing a listener's gaze is to start looking at the target object, pointing at it shortly afterwards and then looking back up to monitor the partner's behavior while continuing the gesture. In the absence of gesture recognition, the gaze fixation becomes the only modality capable of maintaining the focus this way. Thus, it appears plausible that gaze was employed in a similar manner as gestures. Four users indeed tried placing their finger on the object or picking the object up and then displayed the regular gaze behavior, looking at the robot near the end of their utterance. Consequently, it might be worthwhile to explore the effect of gesture recognition on the user's gaze behavior. Our generic modeling approach would easily allow for the addition of a gestural input device. This extension would enable us to investigate the interplay of gaze and gestures in both social and task-oriented types of dialogs.

6.3 Technical Issues

The evaluation also revealed some technical issues with our current implementation that we would like to share at this point.

6.3.1 Eye-Tracking Approach

Our eye tracking solution worked well with most of the participants, but we also observed some effects which might cause problems in the long run. For instance, identifying the user's gaze target using *ARTK*$^+$ markers only works if they are clearly recognizable in the input video. It will be hard to prevent them from being occluded when we are going to add pointing gestures. In the current study, two participants already made the markers unreadable by placing their fingers on the puzzle pieces during a pointing gesture.

Another problem is the limited field of view of the *SMI* glasses. If participants move only their eyes rather than their head, they are able to look at a target outside these boundaries which can not be detected by the system. For 21 of the 209 annotated turns, it was impossible to detect the turn-yielding gaze because the marker on the robot's chest was partially or even fully outside the video image. This might depend on individual preferences since it only happened with 4 of the 13 participants, but also on other factors, such as the distance between the referent and the partner's face.

6.3.2 Speech Understanding

To avoid recognition problems, we had limited the grammar to questions that could be answered with "yes" or "no" and gave the participants example sentences that were understood by our speech recognition. However, about 44% of the users' questions were still not covered by our grammar. For example, our system was able to understand phrasings like *"Do you mean the red one?"*, but not *"Is the color red?"*. About half of these utterances were still matched with the correct attributes, allowing the interaction to continue without any problems. The others, however, were either ignored or interpreted as referring to a wrong object. The former led to additional delays because the users spent several seconds waiting for a reaction, whereas the latter led to answers which appeared to be inappropriate. Both might have interfered with the effect of the robot's gaze behavior. To counter these problems, we are going to expand our grammar based on the utterances we collected in our experiment. We are also thinking of adding a feedback signal when an utterance is being processed, such as the flashing ear LEDs described by Huang and Thomaz [14], so that users can repeat or rephrase an utterance which was not understood by the robot.

7. CONCLUSION

In this paper, we reviewed the roles of gaze that are essential for grounding in collaborative and situated human-robot interaction. Going beyond previous work on computational models of human-robot dialog that covers only certain subsets of these roles, we illustrated a modeling approach which copes with all of them. Our approach combines the flexibility and re-usability of hierarchical and concurrent state charts with the expressiveness and declarative nature of logic programming. We evaluated and discussed our approach based on our collaborative shared workspace scenario.

In line with previous studies, the evaluation confirms that adding gaze tracking to enable crucial task-oriented functions of gaze, specifically gaze following to referenced objects and the disambiguation of spoken referring expressions, improves the interaction with a robot by making it more natural, pleasant and efficient.

However, we made very interesting and unexpected observations for the role of gaze as turn-yielding signal. Our findings suggest that it depends on the combination of modalities used and whether the dialog is task-oriented or social. Gaze is preferably used as a

turn-yielding cue for most of the unimodal questions and for most of those multi-modal questions in which the user decides to use a pointing gesture rather than gaze for disambiguating speech. In those utterances which require referential gaze for disambiguation, this functional aspect of gaze has priority over the use as a turn-yielding signal. In future work, we will use our modeling approach to investigate whether these behavior patterns are grounded in the task-oriented nature of our scenario and if they can still be observed in dialogs with a more pronounced social component.

Acknowledgements

The work described in this paper was partly supported by the Free State of Bavaria, in the form of a doctoral scholarship through the Innovationsfonds.

8. REFERENCES

[1] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[2] M. Argyle and J. Graham. The central europe experiment: Looking at persons and looking at objects. *Environmental psychology and nonverbal behavior*, 1(1):6–16, 1976.

[3] M. Argyle, R. Ingham, F. Alkema, and M. McCallin. The different functions of gaze. *Semiotica*, 7:19–32, 1973.

[4] A. Bangerter. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15:415–419, 2004.

[5] J. Bavelas, L. Coates, and T. Johnson. Listener responses as a collaborative process: The role of gaze. *Communication*, 52:566–580, 2002.

[6] J. Brusk, T. Lager, A. Hjalmarsson, and P. Wik. Deal: Dialogue management in scxml for believable game characters. In *Proceedings of Future Play '07*, pages 137–144. ACM, New York, NY, 2007.

[7] H. Clark. *Using Language*. Cambridge University Press, May 1996.

[8] H. H. Clark. Coordinating with each other in a material world. *Discourse Studies*, 1:507–525, 2005.

[9] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.

[10] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Personality and Social Psychology*, 23:283–292, 1972.

[11] P. Gebhard, G. Mehlmann, and M. Kipp. Visual scenemaker: A tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 6:3–11, 2012.

[12] Z. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11:274–279, 2000.

[13] A. Holroyd, C. Rich, C. Sidner, and B. Ponsler. Generating connection events for human-robot collaboration. In *RO-MAN*, pages 241–246, 2011.

[14] C.-M. Huang and A. Thomaz. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *RO-MAN*, pages 65–71, 2011.

[15] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[16] D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J. F. Ladry. Fusion engines for multimodal input: A survey. In *Proceedings of ICMI' 09*, pages 153–160. ACM, New York, NY, USA, 2009.

[17] G. U. Mehlmann and E. André. Modeling multimodal integration with event logic charts. In *ICMI*, pages 125–132, 2012.

[18] A. Meyer, A. Sleiderink, and W. Levelt. Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66:25–33, 1998.

[19] P. Mundy and L. Newell. Attention, joint attention, and social cognition. *Current Directions in Psychological Science*, 16:269–274, 2007.

[20] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM Trans. Interact. Intell. Syst.*, 1(2):12:1–12:33, Jan. 2012.

[21] B. Mutlu, A. Terrell, and C.-M. Huang. Coordination mechanisms in human-robot collaboration. In *Workshop on Collaborative Manipulation at ACM/IEEE HRI*, 2013.

[22] G. Nielsen. *Studies in Self Confrontation*. Munksgaard, Copenhagen, 1962.

[23] S. Oviatt. Advances in robust multimodal interface design. *IEEE Comput. Graph. Appl.*, 23(5):62–68, September 2003.

[24] S. Oviatt. *The Human-Computer Interaction Handbook*, chapter Multimodal Interfaces, pages 413–432. Mahwah, NJ: Lawrence Erlbaum and Associates, 2008.

[25] C. Rich, B. Ponsler, A. Holroyd, and C. Sidner. Recognizing engagement in human-robot interaction. In *HRI*, pages 375–382, 2010.

[26] D. C. Richardson and R. Dale. Looking to understand - the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29:1045–1060, 2005.

[27] R. Ros, S. Lemaignan, E. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken. Which one? grounding the referent based on efficient human-robot interaction. In *RO-MAN, 2010 IEEE*, pages 570–575, 2010.

[28] H. Sacks, E. Schlegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking. *Language*, 50:696–735, 1974.

[29] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[30] M. Staudte and M. W. Crocker. Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition*, 120:268–291, 2011.

[31] R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel. Enabling multimodal human–robot interaction for the karlsruhe humanoid robot. *Robotics, IEEE Transactions on*, 23(5):840–851, 2007.

[32] D. Traum, A. Leuski, A. Roque, S. Gandhe, D. DeVault, J. Gerten, S. Robinson, and B. Martinovski. Natural language dialogue architectures for tactical questioning characters. In *Army Science Conference*, 2008.

[33] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time. In *Proceedings of the ACM International Conference on Multimedia*, pages 831–834, New York, NY, USA, 2013.

[34] J. Wielemaker, T. Schrijvers, M. Triska, and T. Lager. SWI-Prolog. *Theory and Practice of Logic Programming*, 12(1-2):67–96, 2012.

[35] V. H. Yngve. On getting a word in edgewise. In *Regional Meeting of the Chicago Linguistic Society*, pages 657–677, 1970.