Modeling of Human Visual Attention in Multiparty Open-World Dialogues

KALIN STEFANOV, KTH Royal Institute of Technology, Sweden
GIAMPIERO SALVI, KTH Royal Institute of Technology, Sweden
DIMOSTHENIS KONTOGIORGOS, KTH Royal Institute of Technology, Sweden
HEDVIG KJELLSTRÖM, KTH Royal Institute of Technology, Sweden
JONAS BESKOW, KTH Royal Institute of Technology, Sweden

This study proposes, develops and evaluates methods for modeling the eye-gaze direction and head orientation of a person in multiparty open-world dialogues, as a function of low-level communicative signals generated by his/hers interlocutors. These signals include, speech activity, eye-gaze direction and head orientation, all of which can be estimated in real-time during the interaction. By utilizing these signals and novel data representations suitable for the task and context, the developed methods can generate plausible candidate gaze targets in real-time. The methods are based on Feedforward Neural Networks and Long Short-Term Memory Networks. The proposed methods are developed using several hours of unrestricted interaction data and their performance is compared with a heuristic baseline method. The study offers an extensive evaluation of the proposed methods that investigates the contribution of different predictors to the accurate generation of candidate gaze targets. The results show that the methods can accurately generate candidate gaze targets when the person being modeled is in a listening state. However, when the person being modeled is in a speaking state the proposed methods yield significantly lower performance.

CCS Concepts: • **Computing methodologies → Spatial and physical reasoning**; **Supervised learning by classification**; **Supervised learning by regression**; **Neural networks**.

Additional Key Words and Phrases: human-human interaction, open-world dialogue, eye-gaze direction, head orientation, multiparty

ACM Reference Format:
Kalin Stefanov, Giampiero Salvi, Dimosthenis Kontogiorgos, Hedvig Kjellström, and Jonas Beskow. 2019. Modeling of Human Visual Attention in Multiparty Open-World Dialogues. *ACM Trans. Hum.-Robot Interact.* 1, 1, Article 1 (January 2019), 22 pages. https://doi.org/10.1145/3323231

1 INTRODUCTION

Human-robot interaction is a field undergoing a rapid transformation. From being primarily used in specialized tasks such as industrial manufacturing, robots are increasingly starting to take place in society, which brings the need for robots to understand and exhibit human-like social signals. One signal of particular importance in multiparty human face-to-face interactions is the eye-gaze

Authors' addresses: Kalin Stefanov, KTH Royal Institute of Technology, Lindstedtsvägen 24, Stockholm, Sweden, kalins@kth.se; Giampiero Salvi, KTH Royal Institute of Technology, Lindstedtsvägen 24, Stockholm, Sweden, giampi@kth.se; Dimosthenis Kontogiorgos, KTH Royal Institute of Technology, Lindstedtsvägen 24, Stockholm, Sweden, diko@kth.se; Hedvig Kjellström, KTH Royal Institute of Technology, Teknikringen 14, Stockholm, Sweden, hedvig@kth.se; Jonas Beskow, KTH Royal Institute of Technology, Lindstedtsvägen 24, Stockholm, Sweden, beskow@kth.se.

2573-9522/2019/1-ART1 $15.00
https://doi.org/10.1145/3323231

direction or its coarser representation, the head orientation. In this type of interactions, eye-gaze direction and head orientation signal attention to people and objects in the environment and play important roles in controlling and regulating the conversational floor.

Robots lack computational methods for understanding the nuances of multiparty human face-to-face interactions. In order for robots to engage in natural and effective multiparty face-to-face interactions with humans, there is a need for computational methods that can interpret humans' visual attention and generate appropriate gaze behavior for the robot. Therefore, in this study, we are concerned with the visual attention signaling function of eye-gaze direction and head orientation, and more precisely how we can generate gaze behavior for a robot in a way that mimics human gaze behavior. We consider interactions in an *open-world* domain, i.e., a setting that is dynamic, multiparty and situated in a physical context [Bohus and Horvitz 2010].

This study proposes, develops and evaluates methods following a data-driven modeling approach. We model the eye-gaze direction and head orientation of a person in three-party open-world dialogues, as a function of low-level multimodal signals generated by the two interlocutors. These signals include, speech activity, eye-gaze direction and head orientation which can be automatically estimated in real-time during the interaction. By utilizing these signals, the described methods can generate candidate gaze targets in real-time. To develop and evaluate the methods, we use several hours of multimodal interaction data collected from synchronized eye-gaze, head and audio recordings.

The article is organized as follows. First we examine previous research related to the current study in Section 2, then we describe the proposed methods in Section 3. The experiments we conducted are described in Section 4 and the results of these experiments are presented in Section 5. Discussion on the limitations and contributions of the developed methods can be found in Section 6. We conclude with Section 7.

2 RELATED WORK

The literature offers many different approaches to modeling of human visual attention. These approaches fall under three main categories: biologically-inspired, data-driven, and heuristic. Biologically inspired approaches mimic *bottom-up* neurological responses to the visual input from the environment or are *top-down* cognitive architectures that derive context to detect visual saliency. Data-driven approaches model behavioral aspects of visual attention in human conversations with measurements such as eye-gaze timings, frequencies and locations. Heuristic approaches design rule-based methods that use observations from human interactions in order to mimic visual attention behavior [Admoni and Scassellati 2017].

2.1 Biologically-inspired Approaches

The methods in this domain rely on the fundamental brain mechanisms that realize how humans select visual stimuli to attend to. Humans have the ability to rapidly orient their attention to visually salient locations and only a small fraction of humans' visual input is registered and processed [Itti and Koch 2000].

One of the first attempts in computationally describing humans' visual attention was proposed in [Koch and Ullman 1987]. In this bottom-up method the researchers described that visual attention is centered around a *saliency map*, which is a two-dimensional map that encodes the most salient stimuli of the visual input. Since then, several methods have developed similar structures, where the saliencies of several signals are computed in parallel and combined [Frintrop et al. 2010; Itti and Koch 2001].

In top-down methods, a visually salient stimulus can be decided according to its relevance to the context and can be determined by cognitive phenomena such as knowledge, expectations and

current goals [Borji and Itti 2013]. Visual attention can be encoded into cognitive architectures, such as the ACT-R model [Anderson et al. 1997], that explain high level cognitive processes on how humans choose to process environmental input and attend to visual stimuli.

Some studies combine bottom-up perceptions with top-down behavioral and motivational influences to develop attentional methods [Breazeal and Scassellati 1999; Hoffman et al. 2006]. Other approaches use as input both auditory and visual saliency maps to develop multimodal bottom-up attentional methods [Ruesch et al. 2008; Trafton et al. 2008].

2.2 Data-driven Approaches

Data-driven approaches focus on gathering empirical behavioral data used in methods that model and generate visual attention behavior. The model parameters are typically extracted by analyzing video data of human *dyadic* interactions.

Data-driven studies in [Andrist et al. 2013] and [Andrist et al. 2014] proposed computational methods for generation of gaze aversions in relation to conversational functions and speech. Rich et al. [2010] and Holroyd et al. [2011] presented computational methods for recognizing and maintaining engagement between a human and a humanoid robot. The studies in [Liu et al. 2012] and [Ishi et al. 2010] described computational methods for generation of head motions in relation to dialogue acts. The model parameters here were obtained by analyzing motion capture data. Admoni and Scassellati [2014] presented a computational method for generation of nonverbal behavior. The method, was both predictive (by recognizing the context of new nonverbal behaviors) and generative (by creating new nonverbal behavior based on a desired context) and used both *speaker* and *listener* information to derive context.

Mutlu et al. [2006] presented a computational method for coordination of gaze in narration. The model parameters here were obtained by analyzing video data of a human storyteller in *multiparty* settings. The study in [Huang and Mutlu 2014] presented a computational model for coordination of speech, gaze and gestures in narration.

2.3 Heuristic Approaches

Heuristic approaches focus on using prior knowledge of human behavior from psychology in order to design methods that model and generate visual attention behavior.

Some studies have investigated heuristic methods for dyadic interactions. Colburn et al. [2000] described behavior methods for eye-gaze patterns in the context of real-time verbal communication. Peters et al. [2005] presented a method for an embodied conversational agent that is able to establish, maintain and end the conversation based on its perception of the level of interest of its interlocutor. In this work the speaker and the listener are modeled separately. The study in [Zhang et al. 2017] described an interactive gaze method implemented on a humanoid robot. The system's usability for establishing mutual gaze with a user was also tested.

Heuristic methods have also been investigated in multiparty interactions. The study in [Khullar and Badler 2001] proposed a computational framework for generating visual attending behavior in an embodied simulated human. Gu and Badler [2006] presented a computational method for predicting visual attention behavior for an embodied conversational agent in the presence of distractions. The study in [Bennewitz et al. 2005] presented a humanoid museum guide robot and the proposed system was able to interact with people in multiparty scenarios using attention shifts among other modalities. Spexard et al. [2007] used a humanoid robot which integrated different interaction concepts and perception capabilities to achieve human-oriented interaction.

2.4 Proposed Approach

This study contributes to data-driven approaches to modeling of human visual attention. While most of the data-driven approaches described previously address this problem in dyadic interactions, we propose methods that attempt to model human visual attention in multiparty conversational settings. The limitations and contributions of the study are further discussed in Section 6.

3 METHODS

This section is divided in three parts. The first part presents a formal definition of the problem investigated in the study. In the second part we propose different data representations suitable for modeling spatial relations in the context of multiparty open-world dialogues. In the third part of this section we outline methods for non-temporal and temporal modeling of eye-gaze direction and head orientation using the proposed data representations.

3.1 Problem Definition

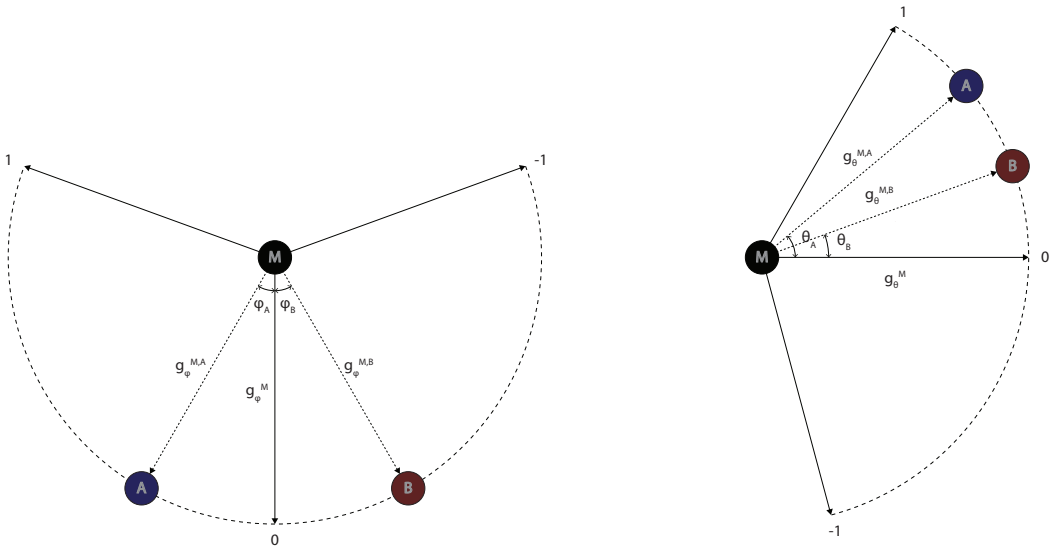
The goal of the proposed methods is to predict visual attention timings and directions generated by a human in multiparty open-world dialogues with a computational model. As a consequence, the model can be used for generation of visual attention behavior for a robot engaging in similar interactions with humans. Specifically, the goal is to build a *person-dependent* model that takes as input 1) the eye-gaze direction or head orientation of all interlocutors, 2) the contextual information from the interaction (i.e., the eye-gaze direction or head orientation that would make each participant look at or face all other participants), and 3) the speech activity of all participants. Given this information (referred to as *predictors* in the text), the model output (referred to as *predictions* in the text) is the eye-gaze direction or head orientation of the person that is being modeled. Next we give a formal description of the model inputs and outputs. For each participant $p \in [1, P]$ in the interaction, we consider the following quantities measured at regular time intervals (we have omitted the time index for simplicity),

- The eye-gaze direction in azimuthal and polar angle $\boldsymbol{g}^p = \{g^p_\phi, g^p_\theta\} \in \mathbb{R}^2$.
- The head orientation in azimuthal and polar angle $\boldsymbol{h}^p = \{h^p_\phi, h^p_\theta\} \in \mathbb{R}^2$.
- The eye-gaze direction in azimuthal and polar angle $\boldsymbol{g}^{p,q} = \{g^{p,q}_\phi, g^{p,q}_\theta \mid p \neq q\} \in \mathbb{R}^2$, that makes participant p look at participant q.
- The head orientation in azimuthal and polar angle $\boldsymbol{h}^{p,q} = \{h^{p,q}_\phi, h^{p,q}_\theta \mid p \neq q\} \in \mathbb{R}^2$, that makes participant p face participant q.
- The binary speech activity $v^p \in \{0, 1\}$.

The problem is then to predict either the eye-gaze direction $\boldsymbol{g}^M$ or the head orientation $\boldsymbol{h}^M$ of a person ($p = M$), based on 1) the eye-gaze direction or head orientation of all interlocutors $\{\boldsymbol{g}^p, \boldsymbol{h}^p \mid p \neq M\}$, 2) all participant-directed eye-gaze directions or head orientations $\{\boldsymbol{g}^{p,q}, \boldsymbol{h}^{p,q} \mid p \neq q\}$, and 3) the binary speech activity v^p of all participants. The following section describes the data representations we have introduced in the study to encode $\boldsymbol{g}^p$, $\boldsymbol{h}^p$, $\boldsymbol{g}^{p,q}$, and $\boldsymbol{h}^{p,q}$.

3.2 Data Representations

A raw three-dimensional representation of eye-gaze direction or head orientation does not capture the dynamic relation between the candidate gaze targets and the current spatial state of the interaction. One of the contributions of this study is the introduction of data representations suitable for modeling spatial relations in the context of multiparty open-world dialogues. The first step in all proposed data representations is transforming the spatial inputs from an interaction-centered three-dimensional Cartesian coordinate system to a participant-centered three-dimensional spherical



(a) Azimuthal angles range.

(b) Polar angles range.

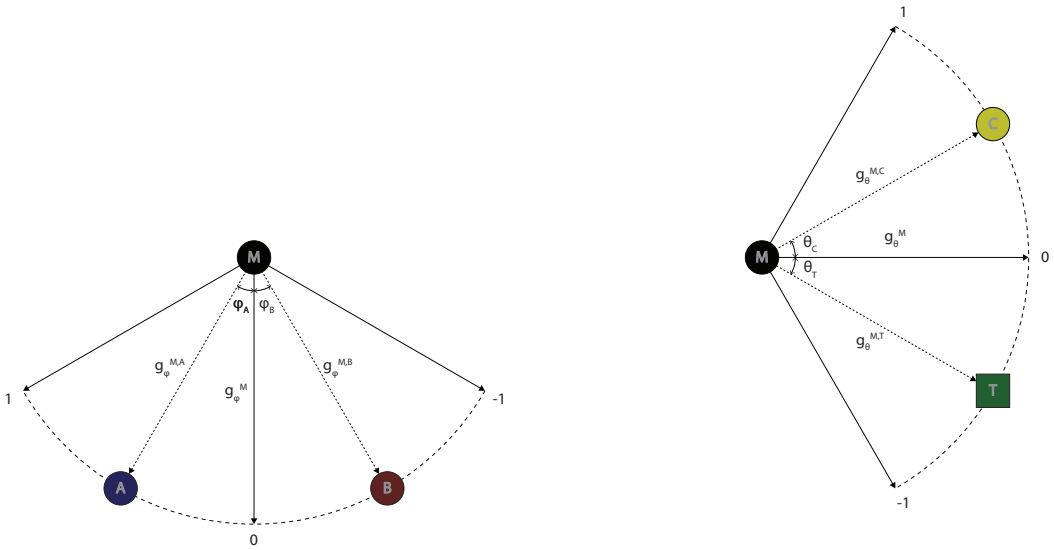
Fig. 1. Continuous and passive data representation. In the figures, M is the person being encoded and for simplicity, we have drawn only the eye-gaze directions. The ranges $[-110°, 110°]$ (azimuthal angles) and $[-75°, 60°]$ (polar angles), are used to create the data representation ranges $[-1, 1]$. All azimuthal and polar angles of the eye-gaze direction or head orientation of M are encoded in these ranges. Then, the current state of the interaction from the perspective of M is expressed in terms of 1) M's eye-gaze direction or head orientation, 2) an eye-gaze direction or head orientation of M that intersects A (interlocutor A), and 3) an eye-gaze direction or head orientation of M that intersects B (interlocutor B). This process of encoding is used to encode the current state of the interaction from the perspective of all participants.

coordinate system. In this way the eye-gaze direction or head orientation are expressed in terms of aforementioned azimuthal angle ϕ and polar angle θ.

Once the spherical coordinates are computed, we consider four different representations for the data illustrated by Figures 1 through 4. The first two representations are continuous and differ by the way angles are mapped to the $[-1, 1]$ range. The last two representations, instead, are obtained by quantizing the space into a finite number of regions, and are, therefore, discrete. In the following sections we will give details on each representation.

3.2.1 Continuous and passive data representation. This representation, illustrated by Figure 1, is based on a fixed mapping from the azimuthal and polar angles to the $[-1, 1]$ ranges of the representation. We use the visual field of humans [Walker et al. 1990] to define the limits of the observed range. Without head movements, this range corresponds to $\sim 220°$ in azimuthal angles (Figure 1a) and $\sim 135°$ in polar angles (Figure 1b). In this representation the eye-gaze directions $\boldsymbol{g}^p$, head orientations $\boldsymbol{h}^p$, participant-directed eye-gaze directions $\boldsymbol{g}^{p,q}$, and participant-directed head orientations $\boldsymbol{h}^{p,q}$, all vary for every time step. When using this data representation, we include all these quantities in the encoding of the current state of the interaction.

3.2.2 Continuous and active data representation. In this representation, illustrated by Figure 2, we perform a dynamic mapping of the physical angles to the $[-1, 1]$ ranges, that depends on the relative position of the participant with respect to the other participants. We map the azimuthal angles in such a way that the participant-directed eye-gaze directions or participant-directed head orientations is always either -0.5 or 0.5. For example, for $p = M$, $\boldsymbol{g}_{\phi}^{M,A} = \boldsymbol{h}_{\phi}^{M,A} = 0.5$ and $\boldsymbol{g}_{\phi}^{M,B} = \boldsymbol{h}_{\phi}^{M,B} = -0.5$ (Figure 2a). Similarly, we assign the value 0.5 to the polar angles of the eye-gaze directions or head orientations that intersect the current mean position of two of the other participants and the value -0.5 to the polar angles of the eye-gaze directions or head orientations that intersect a static object in the environment. For example, for $p = M$, $\boldsymbol{g}_{\theta}^{M,C} = \boldsymbol{h}_{\theta}^{M,C} = 0.5$ and $\boldsymbol{g}_{\theta}^{M,T} = \boldsymbol{h}_{\theta}^{M,T} = -0.5$ (Figure 2b). In this representation only the eye-gaze directions $\boldsymbol{g}^{p}$ and head orientations $\boldsymbol{h}^{p}$ vary for every time step.



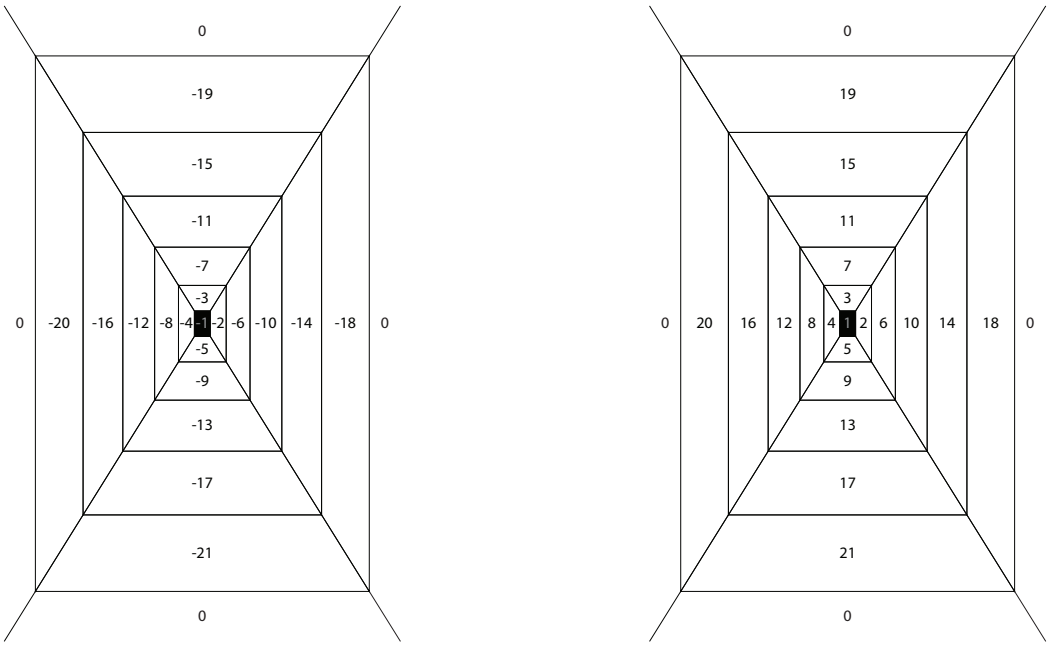
(a) Azimuthal angles range.

(b) Polar angles range.

Fig. 2. Continuous and active data representation. In the figures, M is the person being encoded and for simplicity, we have drawn only the eye-gaze directions. The azimuthal angles of the eye-gaze directions or head orientations of M that pass through the current position of A (interlocutor A) and B (interlocutor B) are used to create a data representation range $[-0.5, 0.5]$ for the azimuthal angles. The polar angles of the eye-gaze directions or head orientations of M that pass through C (the current mean position of A and B) and the position of T (a static object) are used to create a data representation range $[-0.5, 0.5]$ for the polar angles. In the final step, both data representation ranges are extended to $[-1, 1]$. Then, the current state of the interaction from the perspective of M is expressed in terms of 1) M's eye-gaze direction or head orientation. This process of encoding is used to encode the current state of the interaction from the perspective of all participants.

3.2.3 Discrete and high-resolution data representation. The high-resolution representation partitions the space around two participants into 42 (+1) regions, Figure 3. We use average male head size (AMH) as reference for the size of the regions. Regions ±1 to ±5 have sizes of 1× AMH, ±6 to

±9 have sizes of 1.5× AMH, ±10 to ±13 have sizes of 2× AMH, ±14 to ±17 have sizes of 2.5× AMH, and ±18 to ±21 have sizes of 3× AMH. The rest of the space belongs to region 0. The incremental increase of the regions' size is based on the assumption that regions that are spatially close to the position of the head are more important than those spatially distant. Then, the number of regions defined here covers the social space as described in [Hall 1990]. The regions (one for interlocutor A and one for interlocutor B) intersected by the current eye-gaze direction or head orientation of the participant under consideration are used for encoding.

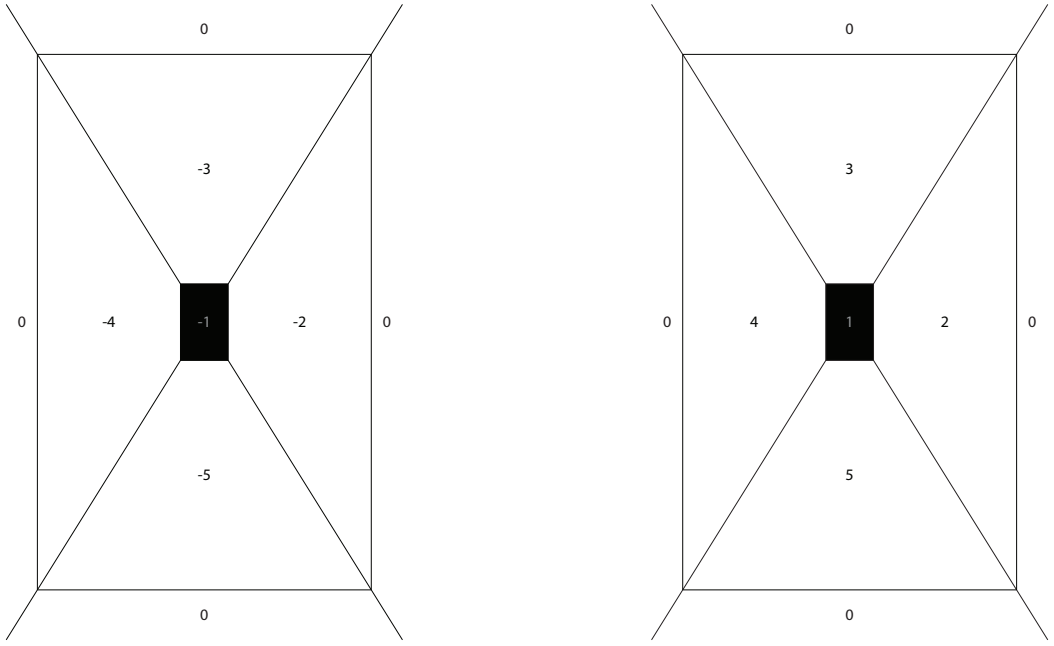


(a) Interlocutor B.

(b) Interlocutor A.

Fig. 3. Discrete and high-resolution data representation. As seen from the perspective of M, given the current position of interlocutor A, a grid centered at interlocutor A is defined. The grid is defined in such a way that it partitions the space around the position of interlocutor A into 21 regions. The same type of partitioning is applied to interlocutor B. The regions (one for interlocutor A and one for interlocutor B) intersected by the current eye-gaze direction or head orientation of M are used as encoding. This process of encoding is used to encode the current state of the interaction from the perspective of all participants.

3.2.4 Discrete and low-resolution data representation. The low-resolution representation partitions the space around two participants into 10 (+1) regions, Figure 4. Given the current position of two participants, we follow the same process as in the high-resolution representation. In this case however, regions ±1 to ±5 are merged into regions ±1, regions ±6, ±10, ±14, and ±18, are merged into regions ±2, regions ±7, ±11, ±15, and ±19, are merged into regions ±3, regions ±8, ±12, ±16, and ±20, are merged into regions ±4, and regions ±9, ±13, ±17, and ±21, are merged into regions ±5. The rest of the space belongs to region 0. As in the previous representation, the regions (one for interlocutor A and one for interlocutor B) intersected by the current eye-gaze direction or head orientation of the participant under consideration are used for encoding.



(a) Interlocutor B. (b) Interlocutor A.

Fig. 4. Discrete and low-resolution data representation. As seen from the perspective of M, given the current position of interlocutor A, a grid centered at interlocutor A is defined. The grid is defined in such a way that it partitions the space around the position of interlocutor A into 5 regions. The same type of partitioning is applied to interlocutor B. The regions (one for interlocutor A and one for interlocutor B) intersected by the current eye-gaze direction or head orientation of M are used as encoding. This process of encoding is used to encode the current state of the interaction from the perspective of all participants.

3.3 Models

Given the proposed data representations, we use a supervised learning approach for estimation/prediction of a participant's eye-gaze direction or head orientation based on the current eye-gaze directions or head orientations of two interlocutors and the current speech activity. For the continuous data representations, the problem of modeling the eye-gaze direction or head orientation is a regression task. In the discrete data representations, the problem is a classification task. The regression models output estimated continuous values of the azimuthal and polar angles in the ranges defined by the continuous data representations. The classification models output predicted discrete values of the most likely regions defined by the discrete data representations. The evaluation of the models is performed by computing the error/accuracy of the estimations/predictions with respect to a ground truth (the estimates of the recording devices), on a frame-by-frame basis. We use two types of models: non-temporal and temporal.

3.3.1 Non-temporal models. We train a Feedforward Neural Network (ANN) model to perform both classification and regression tasks. In the case of discrete data representations, the goal of the ANN classifier is to predict the most likely regions (one per interlocutor) for the participant's eye-gaze direction or head orientation. In the case of continuous data representations, the goal of the ANN regressor is to estimate the azimuthal and polar angles for the participant's eye-gaze

direction or head orientation. These models work on a frame-by-frame basis and have no memory of past frames.

3.3.2 Temporal models. We also train a Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber 1997] model to perform both classification and regression tasks. In the case of discrete data representations, the goal of the LSTM classifier is the same as the one of the ANN classifier: to predict regions (one per interlocutor) for the participant's eye-gaze direction or head orientation. Similarly, in the case of continuous data representations, the goal of the LSTM regressor is the same as the one of the ANN regressor: to estimate the azimuthal and polar angles for the participant's eye-gaze direction or head orientation. However, the LSTM models maintain memory of past frames, so that temporal continuity is taken into account in the estimations/predictions.

4 EXPERIMENTS

This section is divided in two parts. In the first part we describe the datasets used to develop and evaluate the proposed methods. In the second part we describe the specific experiments used to evaluate the proposed methods and general experimental setup settings shared across experiments.

4.1 Datasets

The methods proposed in Section 3 are developed and evaluated using two multiparty interaction datasets. The main purpose of these datasets is to explore visual attention patterns of humans in different contexts. In this study we consider only data which is generated in the context of dynamic multiparty situated interactions without visual/auditory distractions, that [Bohus and Horvitz 2010] define as an open-world dialogues.

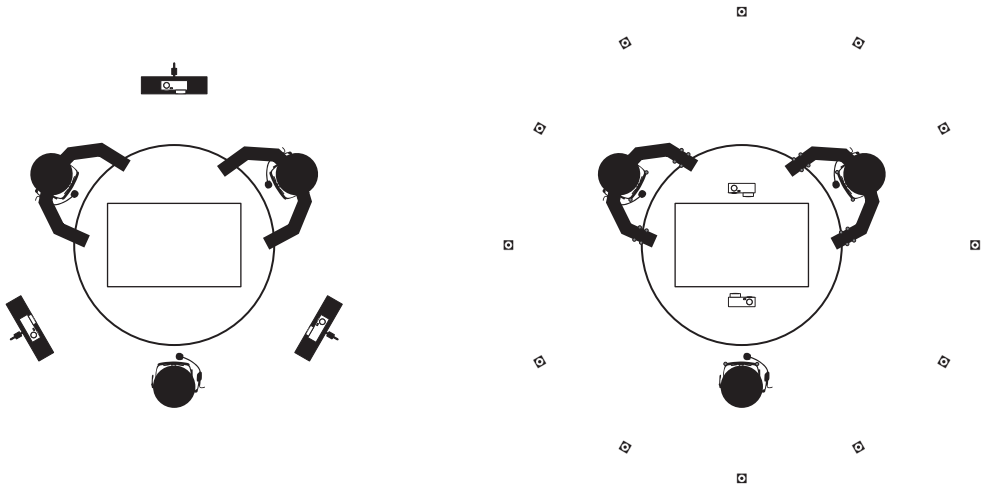
4.1.1 Kinect dataset. This dataset consists of a total of 15 sessions, each with duration of ∼30 minutes, resulting in ∼7.5 hours of data per recording device [Stefanov and Beskow 2016]. Three participants take part in each session where a pair of participants is new in every session, and one participant takes the role of moderator for all sessions. All interactions are in English and all data streams are spatially and temporally synchronized and aligned. The interactions occur around a round table and the participants are seated. There are in total 23 unique participants and 1 moderator. Figure 5a illustrates the spatial configuration of the setup during the recordings. The following data streams are included in the dataset,

- 3 streams of audio data.
- 3 streams of eye-gaze data.
- 3 streams of color, depth, infrared, body and face data.
- 3 streams of video data.
- 1 stream of game state data.
- 1 stream of robot state data.

4.1.2 OptiTrack dataset. This dataset consists of a total of 15 sessions, each with duration of ∼1 hour, resulting in ∼15 hours of data per recording device [Kontogiorgos et al. 2018]. Three participants take part in each session where two of them are new in every session, and one participant takes the role of a moderator for all sessions. All interactions are in English and all data streams are spatially and temporally synchronized and aligned. The interactions occur around a round table and the participants are seated. There are in total 30 unique participants and 1 moderator. Figure 5b illustrates the spatial configuration of the setup during the recordings. The following data streams are included in the dataset,

- 3 streams of audio data.
- 3 streams of eye-gaze data.

- 2 streams of video data.
- 1 stream of game state data.
- 1 stream of motion capture data.



(a) Kinect dataset. (b) OptiTrack dataset.

Fig. 5. Spatial configuration of the sensors and the participants in the datasets.

In this study, the eye-gaze directions are generated by Tobii Pro Glasses 2 eye trackers used in both datasets. The head orientations are generated by Kinect v2 sensors used in the Kinect dataset and by an OptiTrack motion capture system used in the OptiTrack dataset. The audio used for speech activity is generated by close-talking microphones used in both datasets.

4.2 Setup

We report results based on random sampling of the data, where ~80% of the data is used for training, ~15% is used for testing, and ~5% is used for validation. Next we outline the parameters used in all of the presented models.

4.2.1 Non-temporal models. The non-temporal models (ANN) are trained and evaluated with isolated frames and include one fully-connected layer of size 64 (regression) or 128 (classification) with rectifier activations, followed by an output layer with linear activations (regression) or sigmoid activations (classification).

4.2.2 Temporal models. The temporal models (LSTM) are trained and evaluated with 25 frame (1 second) long segments without overlaps and include one long short-term memory layer of size 64 (regression) or 128 (classification) with hyperbolic tangent activations, followed by an output layer with linear activations (regression) or sigmoid activations (classification).

4.2.3 Heuristic models. To compare the performance of the proposed methods we defined and developed one heuristic baseline model for prediction of the participant's eye-gaze direction or head orientation. In this model the participant's eye-gaze direction or head orientation is always towards the currently speaking interlocutor (if exists), and in the center (between the interlocutors) otherwise.

4.2.4 Model hyperparameters. For training the models (ANN and LSTM) we use Adam optimizer, [Kingma and Ba 2014], with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$). The models that perform a regression task use a mean absolute error loss function as an optimization criteria, and the models that perform a classification task use a binary crossentropy loss function as an optimization criteria. All models are trained for 100 epochs. The model corresponding to the best validation performance is selected and evaluated on the test data. For development we use Keras [Chollet et al. 2015], with TensorFlow backend [Abadi et al. 2015]. We have summarized all hyperparameters in Table 1.

Table 1. Model hyperparameters. The first column lists all of the presented model configurations; the data representation is specified in the parenthesis using the notations introduced in Section 3.2 and Section 3.3. In this column CP stands for continuous and passive, CA stands for continuous and active, DH stands for discrete and high-resolution, and DL stands for discrete and low-resolution. The second column lists the specific model parameters. In this column the data flow from the Input layer is represented by arrows and the layer size and activation function is specified in the parenthesis. The size of the Input layer depends on the type of predictors used by the model. Here Dense stands for Fully Connected layer, LSTM stands for Long Short-Term Memory layer, and TD_Dense stands for time distributed Fully Connected layer.

Models	Parameters
ANN (CP)	Input ($7 \vee 16 \vee 19$,) $\rightarrow$ Dense (64, ReLU) $\rightarrow$ Dense (2, Linear)
ANN (CA)	Input ($3 \vee 4 \vee 7$,) $\rightarrow$ Dense (64, ReLU) $\rightarrow$ Dense (2, Linear)
LSTM (CP)	Input ($7 \vee 16 \vee 19$,) $\rightarrow$ LSTM (64, Tanh) $\rightarrow$ TD_Dense (2, Linear)
LSTM (CA)	Input ($3 \vee 4 \vee 7$,) $\rightarrow$ LSTM (64, Tanh) $\rightarrow$ TD_Dense (2, Linear)
ANN (DH)	Input ($3 \vee 84 \vee 87$,) $\rightarrow$ Dense (128, ReLU) $\rightarrow$ Dense (42, Sigmoid)
ANN (DL)	Input ($3 \vee 20 \vee 23$,) $\rightarrow$ Dense (128, ReLU) $\rightarrow$ Dense (10, Sigmoid)
LSTM (DH)	Input ($3 \vee 84 \vee 87$,) $\rightarrow$ LSTM (128, Tanh) $\rightarrow$ TD_Dense (42, Sigmoid)
LSTM (DL)	Input ($3 \vee 20 \vee 23$,) $\rightarrow$ LSTM (128, Tanh) $\rightarrow$ TD_Dense (10, Sigmoid)

4.2.5 Predictors and predictions. The goal of the experiments in this study is to investigate the utility of different predictors for prediction of eye-gaze direction or head orientation. Therefore, we built a model for each moderator and tested it on independent data from the same moderator, i.e., the models are person-dependent as described in Section 3.1. To test the contribution of different predictors to the prediction of eye-gaze direction or head orientation we conducted the following experiments,

- Interlocutors' eye-gaze direction and all speech activity as predictors of moderator's eye-gaze direction (GSa-G).
- Interlocutors' head orientation and all speech activity as predictors of moderator's head orientation (HSa-H).
- Interlocutors' eye-gaze direction as a predictor of moderator's eye-gaze direction (G-G).
- Interlocutors' head orientation as a predictor of moderator's head orientation (H-H).
- All speech activity as a predictor of moderator's eye-gaze direction (Sa-G).
- All speech activity as a predictor of moderator's head orientation (Sa-H).

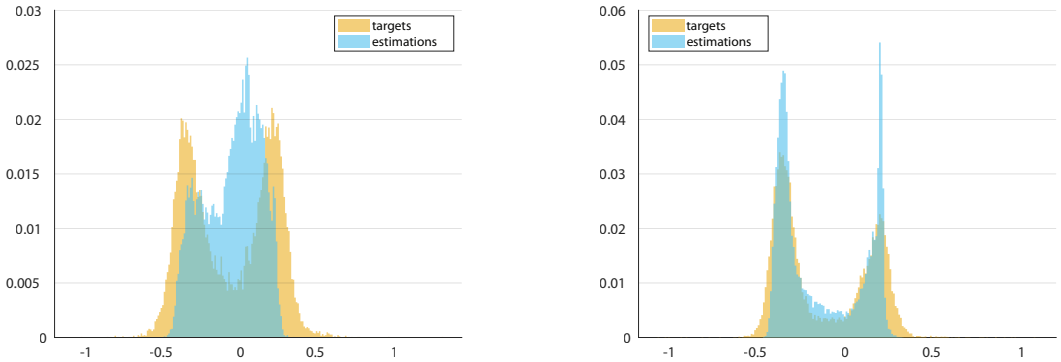
5 RESULTS

This section is divided in three parts. In the first part we compare the performance of the proposed methods when the moderators are either in *speaking* or in *listening* state. In the second part we report and discuss the results of the conducted experiments for the moderator in the Kinect dataset.

In the third part we report and discuss the results of the conducted experiments for the moderator in the OptiTrack dataset.

5.1 Speaking and Listening

Regardless of the representation, modeling method, or moderator, the performance on data in which the moderators are in speaking state is significantly lower than the performance on data in which the moderators are in listening state. We report this result with the histograms in Figure 6 and Figure 7. The figures illustrate the distribution of the estimated values for the azimuthal angles in comparison to the distribution of the target values, for both speaking and listening states. Clearly, when a moderator is in speaking state, the distribution of the estimated values is different than the underlaying target distribution. In listening state, however, the distribution of the estimated values is more similar to the underlaying target distribution. One explanation of this result is the fact that the data representations proposed in this study do not encode any information about the intentions of the moderators. Therefore, in the next two sections, we present results only from the listening state of the moderators.



(a) Speaking state. (b) Listening state.

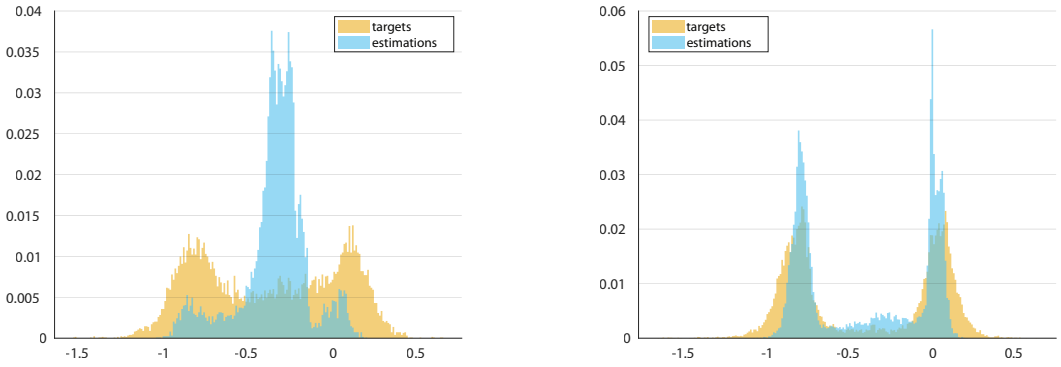
Fig. 6. Example distributions of the target and estimated azimuthal angles for the moderator in the Kinect dataset.

5.2 Kinect Dataset

The results in this section are based on the Kinect dataset. As mentioned previously, we present results only from the listening state of the moderator. We have summarized the data distribution in different sets in Table 2.

Table 2. Data distribution in different sets in the Kinect dataset.

	Number of Frames		
Set	**Train**	**Test**	**Validation**
Eye-gaze direction	81236	15121	5094
Head orientation	138340	25784	8596



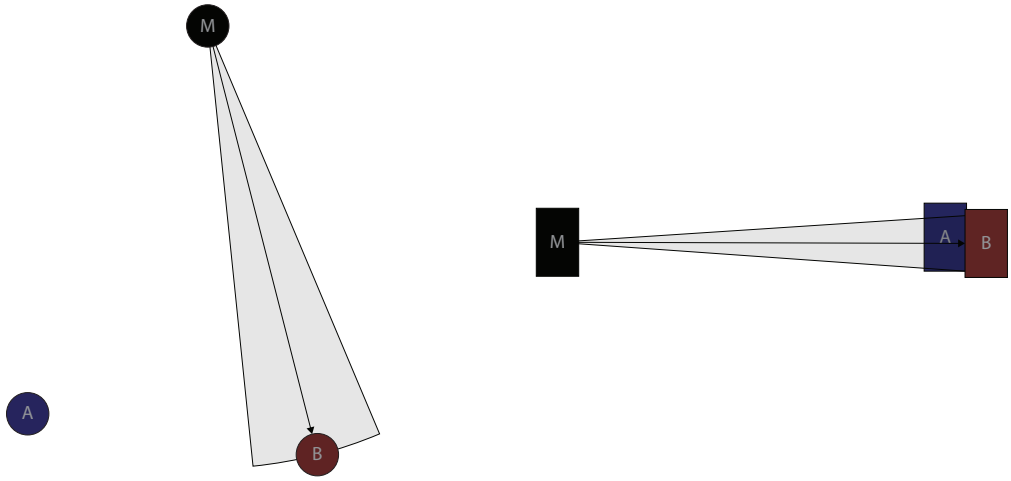
(a) Speaking state. (b) Listening state.

Fig. 7. Example distributions of the target and estimated azimuthal angles for the moderator in the OptiTrack dataset.

5.2.1 Continuous and passive data representation. The results of this experiment are summarized in Figure 10a using mean absolute error (the best performance is error of 0°). Here, the $[-1, 1]$ range corresponds to 220° in azimuthal angles and 135° in polar angles. The proposed temporal method (LSTM) reaches the lowest error of ∼10° for the azimuthal angle and ∼5° for the polar angle when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input (HSa-H). The proposed non-temporal method (ANN) reaches the lowest error in the same configuration (HSa-H) an it is ∼9° for the azimuthal angle and ∼4° for the polar angle. The baseline method also reaches the lowest error in this configuration and it is ∼20° and ∼10° for the azimuthal and the polar angle respectively. In summary, the proposed methods outperform the baseline method, and their performance is almost identical. We have visualized the mean absolute error of the best performing model in Figure 8.

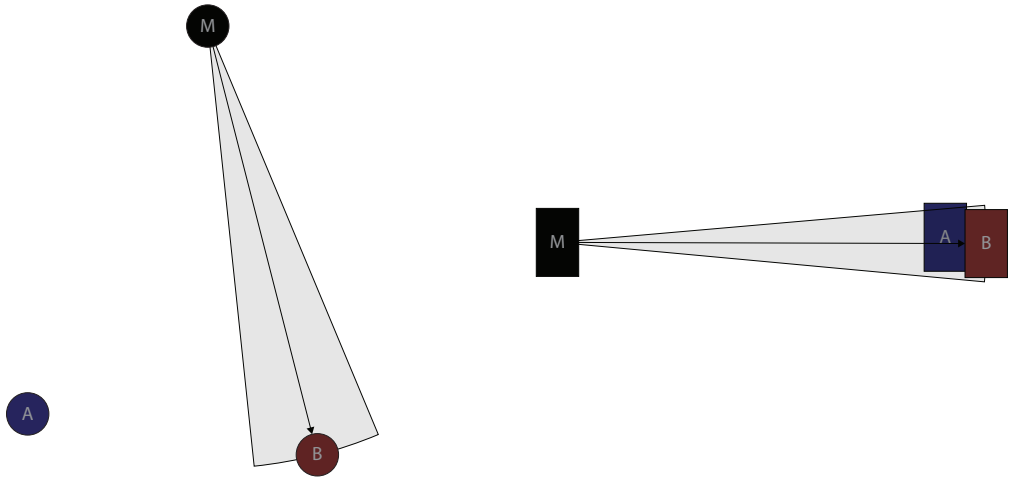
5.2.2 Continuous and active data representation. The results of this experiment are summarized in Figure 10b using mean absolute error (the best performance is error of 0°). Here, the $[-1, 1]$ range corresponds to mean of ∼78° in azimuthal angles and mean of ∼42° in polar angles. The proposed temporal method (LSTM) reaches the lowest error of ∼9° for the azimuthal angle and ∼6° for the polar angle when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input (HSa-H). The proposed non-temporal method (ANN) reaches the lowest error in the same configuration (HSa-H) an it is ∼8° for the azimuthal angle and ∼5° for the polar angle. The baseline method also reaches the lowest error in this configuration and it is ∼21° (azimuthal angle) and ∼11° (polar angle). In summary, the proposed methods outperform the baseline method, and their performance is almost identical. We have visualized the mean absolute error of the best performing model in Figure 9.

5.2.3 Discrete and high-resolution data representation. The results of this experiment are summarized in Figure 11a using accuracy (the best performance is accuracy of 100%). The proposed temporal method (LSTM) reaches the highest accuracy when predicting the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input (HSa-H). Specifically, the accuracy is ∼70% for interlocutor A and ∼44% for interlocutor B. In the same configuration (HSa-H), the proposed non-temporal method (ANN) reaches its highest accuracy of ∼66% for interlocutor A and ∼42% for interlocutor B. The baseline method also reaches its



(a) Azimuthal angle error. (b) Polar angle error.

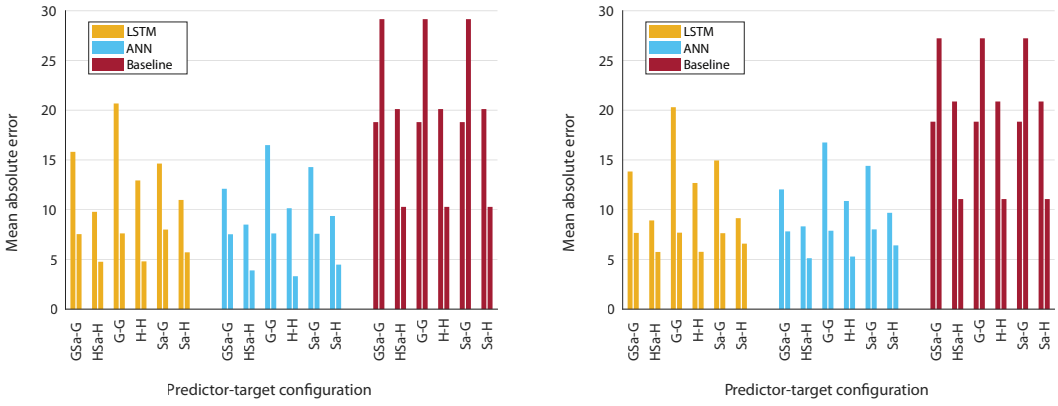
Fig. 8. Mean absolute error of the best performing continuous and passive model for the Kinect dataset. The mean location of all participants in all interactions is drawn using average head size (circles and rectangles). All physical proportions are kept.



(a) Azimuthal angle error. (b) Polar angle error.

Fig. 9. Mean absolute error of the best performing continuous and active model for the Kinect dataset. The mean location of all participants in all interactions is drawn using average head size (circles and rectangles). All physical proportions are kept.

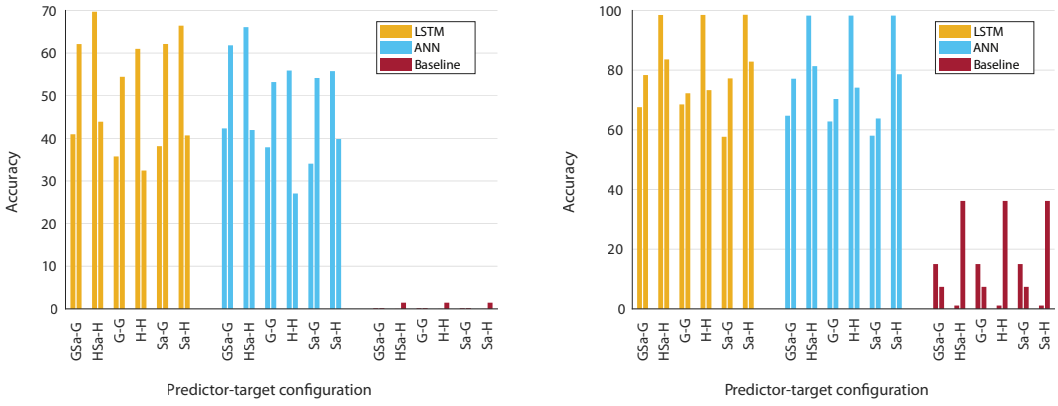
highest accuracy in this configuration and it is ~0.02% (interlocutor A) and ~1% (interlocutor B). In summary, the proposed methods outperform the baseline method, while the performance of the LSTM is better than the performance of the ANN.



(a) Passive methods. (b) Active methods.

Fig. 10. Mean absolute error for the continuous methods. The first bar in each predictor-target configuration is the mean absolute error (in degrees) for the azimuthal angle and the second bar is the mean absolute error (in degrees) for the polar angle.

5.2.4 Discrete and low-resolution data representation. The results of this experiment are summarized in Figure 11b using accuracy (the best performance is accuracy of 100%). The proposed temporal method (LSTM) reaches the highest accuracy when predicting the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input (HSa-H): ~98% for interlocutor A and ~84% for interlocutor B. Using the same configuration, the proposed non-temporal method (ANN) reaches its highest accuracy of ~98% for interlocutor A and ~81% for interlocutor B. The baseline method also yields its best performance in this configuration and it is ~1% (interlocutor A) and ~36% (interlocutor B). In summary, the proposed methods outperform the baseline method, while the performance of the LSTM is better than the performance of the ANN.



(a) High-resolution methods. (b) Low-resolution methods.

Fig. 11. Accuracy for the discrete methods. The first bar in each predictor-target configuration is the accuracy (in percentage) for interlocutor A and the second bar is the accuracy (in percentage) for interlocutor B.

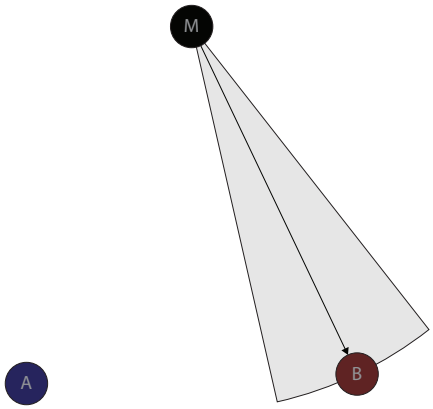
5.3 OptiTrack Dataset

The results in this section are based on the OptiTrack dataset. As mentioned previously, we present results only from the listening state of the moderator. We have summarized the data distribution in different sets in Table 3.

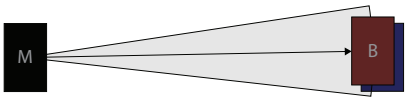
Table 3. Data distribution in different sets in the OptiTrack dataset.

	Number of Frames		
Set	**Train**	**Test**	**Validation**
Eye-gaze direction	60645	11257	3755
Head orientation	138082	25756	8529

5.3.1 Continuous and passive data representation. The results of this experiment are summarized in Figure 14a using mean absolute error (the best performance is error of 0°). Here, the $[-1, 1]$ range corresponds to 220° in azimuthal angles and 135° in polar angles. The proposed temporal method (LSTM) reaches the lowest error of ∼14° for the azimuthal angle and ∼7° for the polar angle when estimating the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input (HSa-H). The proposed non-temporal method (ANN) also reaches the lowest error in this configuration and it is ∼13° for the azimuthal angle and ∼7° for the polar angle. The baseline method, however, reaches its best performance when estimating the moderator's eye-gaze direction using the eye-gaze direction of the interlocutors and the speech activity as input (GSa-G). The error in this case is ∼21° and ∼10° for the azimuthal and the polar angle respectively. In summary, the proposed methods outperform the baseline method, and their performance is almost identical. We have visualized the mean absolute error of the best performing model in Figure 12.



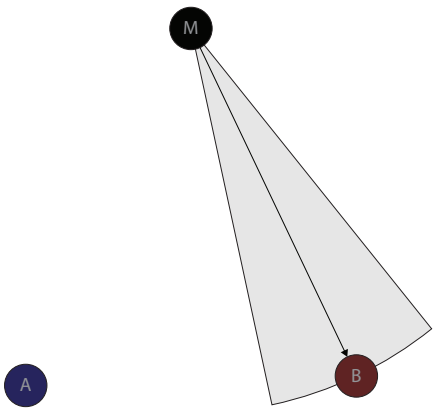
(a) Azimuthal angle error.



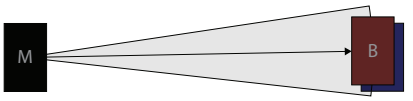
(b) Polar angle error.

Fig. 12. Mean absolute error of the best performing continuous and passive model for the OptiTrack dataset. The mean location of all participants in all interactions is drawn using average head size (circles and rectangles). All physical proportions are kept.

5.3.2 Continuous and active data representation. The results of this experiment are summarized in Figure 14b using mean absolute error (the best performance is error of 0°). Here, the $[-1, 1]$ range corresponds to mean of ∼100° in azimuthal angles and mean of ∼57° in polar angles. The proposed temporal method (LSTM) reaches the lowest error of ∼15° for the azimuthal angle and ∼6° for the polar angle when predicting the moderator's eye-gaze direction using the eye-gaze direction of the interlocutors and the speech activity as input (GSa-G). The proposed non-temporal method (ANN) reaches also reaches the lowest error in this configuration. Specifically, the error is ∼13° for the azimuthal angle and ∼7° for the polar angle. It is worth mentioning that the performance of both models (LSTM and ANN) in the HSa-H configuration is almost identical to that in the GSa-G configuration. The baseline method also reaches the lowest error in this configuration (GSa-G) and it is ∼21° (azimuthal angle) and ∼14° (polar angle). In summary, the proposed methods outperform the baseline method, while the performance of the ANN is better than the performance of the LSTM. We have visualized the mean absolute error of the best performing model in Figure 13.



(a) Azimuthal angle error.

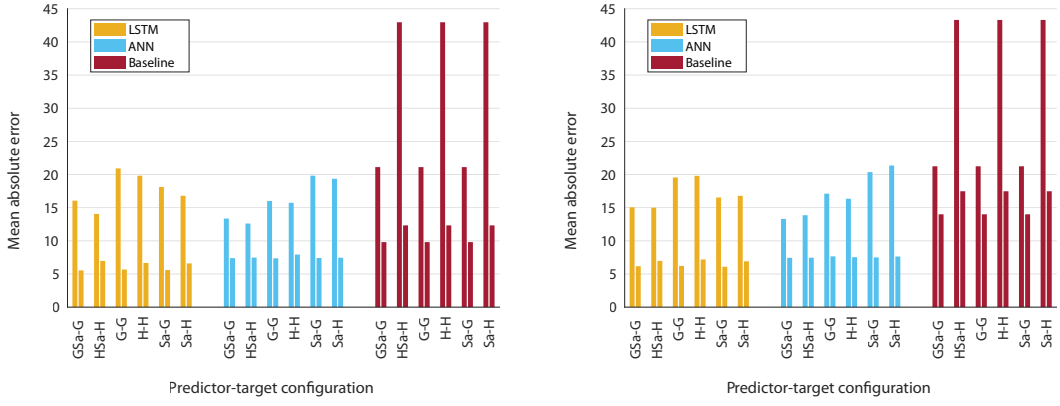


(b) Polar angle error.

Fig. 13. Mean absolute error of the best performing continuous and active model for the OptiTrack dataset. The mean location of all participants in all interactions is drawn using average head size (circles and rectangles). All physical proportions are kept.

5.3.3 Discrete and high-resolution data representation. The results of this experiment are summarized in Figure 15a using accuracy (the best performance is accuracy of 100%). The proposed temporal method (LSTM) reaches the its highest accuracy when predicting the moderator's head orientation using the head orientation of the interlocutors and the speech activity as input (HSa-H). Specifically, the accuracy is ∼75% for interlocutor A and ∼70% for interlocutor B. Using the same inputs, the proposed non-temporal method (ANN) reaches its highest accuracy of ∼70% for interlocutor A and ∼61% for interlocutor B. The baseline method also reaches its highest accuracy in this configuration and it is ∼12% for interlocutor A and ∼0.4% for interlocutor B. In summary, the proposed methods outperform the baseline method, while the performance of the LSTM is better than the performance of the ANN.

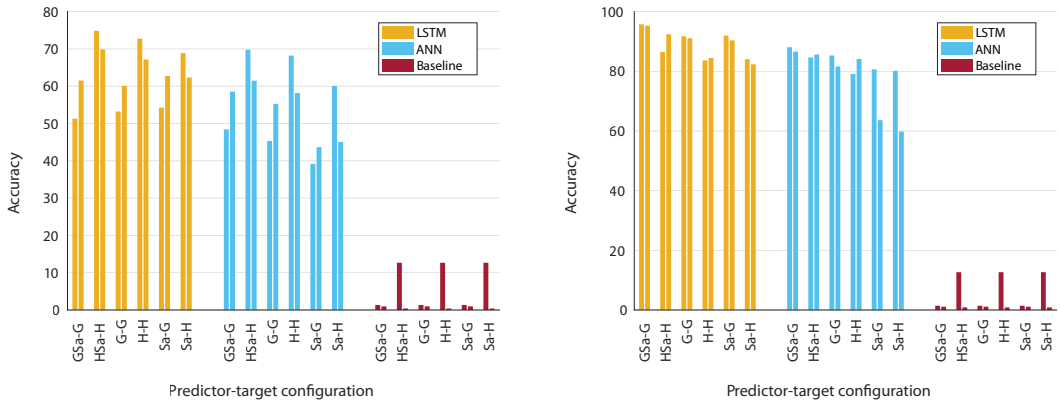
5.3.4 Discrete and low-resolution data representation. The results of this experiment are summarized in Figure 15b using accuracy (the best performance is accuracy of 100%). The proposed temporal method (LSTM) reaches the highest accuracy when predicting the moderator's eye-gaze direction



(a) Passive methods. (b) Active methods.

Fig. 14. Mean absolute error for the continuous methods. The first bar in each predictor-target configuration is the mean absolute error (in degrees) for the azimuthal angle and the second bar is the mean absolute error (in degrees) for the polar angle.

using the eye-gaze direction of the interlocutors and the speech activity as input (GSa-G). Specifically, the accuracy is ~96% for interlocutor A and ~95% for interlocutor B. Using the same inputs, the proposed non-temporal method (ANN) also reaches its highest accuracy of ~88% for interlocutor A and ~87% for interlocutor B. The baseline method, however, reaches it highest accuracy in the HSa-H configuration and it is ~13% and ~1%. In summary, the proposed methods outperform the baseline method, while the performance of the LSTM is better than the performance of the ANN.



(a) High-resolution methods. (b) Low-resolution methods.

Fig. 15. Accuracy for the discrete methods. The first bar in each predictor-target configuration is the accuracy (in percentage) for interlocutor A and the second bar is the accuracy (in percentage) for interlocutor B.

6 DISCUSSION

The purpose of the conducted experiments is to investigate to what extent in the given context, complex visual attention behavior can be modeled solely from low-level multimodal signals, without taking semantic information into account. Comparing GSa-G/HSa-H and G-G/H-H shows that speech activity coupled with interlocutors' eye-gaze direction or head orientation is always a better predictor for the moderator's eye-gaze direction or head orientation than using interlocutors' eye-gaze direction or head orientation alone. Further, comparing Sa-G/Sa-H and G-G/H-H shows that speech activity alone is a very competitive predictor for the moderator's eye-gaze direction or head orientation. Finally, comparing GSa-G/HSa-H and Sa-G/Sa-H yields better performance for GSa-G/HSa-H. In summary, the results of the experiments clearly show that in this context (open-world dialogues), models that take into account the speech activity significantly outperform models which do not use this information.

6.1 Limitations

The proposed methods are capable of generating plausible candidate gaze targets when the moderators are in listening state. However, the methods do not consider the current goals/intentions of the moderators which yields significantly lower performance when the moderators are in speaking state. This observation motivates the development of different methods for the speaker and listener roles in this context, as it was suggested by [Peters et al. 2005]. A higher level process which specifies the moderators' goals should be present and augmented with the proposed methods to generate candidate gaze targets when the moderators are in speaking state.

6.2 Contributions

The proposed methods were built and evaluated with multiple days of recordings with multiple pairs of interlocutors. Generally, every new pair of interlocutors will present different stimuli, both in terms of dynamics and behavior, to the moderators. This suggests that the methods performance is promising (given the variability of the stimuli) and they could be used for generation of candidate gaze targets that are natural in the context.

There is a specific motivation behind each of the proposed data representation methods. The continuous and passive case attempts to represent the interaction from the perspective of the moderators, while the continuous and active case attempts to represent the interaction from the perspective of the interlocutors. The discrete and high-resolution case partitions the space in detailed regions for high resolution candidate gaze targets generation, while the discrete and low-resolution case partitions the space in broader regions in order to capture the general gaze behavior (*look at* or *look away*).

All of the proposed methods can be extended to interactions with more than three participants because they do not assume a specific number of people (all methods however, require at least three participants). In such cases, each unique pair of interlocutors produces signals. By considering all possible unique pairs of interlocutors, the methods can generate several candidate gaze targets sets (one set per unique pair of interlocutors) and a final decision for the gaze target can be made, for example, with a simple voting procedure.

The predictions of the methods when using discrete data representations can be seen as spatial saliency maps (one per interlocutor). In these cases, each region has a value in the interval $[0, 1]$ which corresponds to how likely it is for this region to be a gaze target. This representation can be further augmented with biologically-inspired visual (even auditory) saliency maps. Borrowing the terms from the biologically-inspired approaches, the methods here can be seen as bottom-up

processes, while the previously mentioned "higher level process which specifies the moderators' goals", as a top-down one.

We also presented a simple heuristic baseline method and compared its performance with the proposed data-driven methods. Given the context of dynamic multiparty situated interactions without visual/auditory distractions, in the current study we did not compare the proposed methods with biologically-inspired one; the latter (i.e, visual saliency maps) usually do not incorporate information related to face-to-face interactions, but rather are used as mechanisms for selecting visual stimuli in cluttered visual scenes. However, the presented heuristic baseline method can be seen as producing auditory saliency maps because the most salient location there, is always the speaking interlocutor.

In summary, the contributions of this study are the introduction of novel data representations suitable for modeling spatial relations in the context of multiparty open-world dialogues and the investigation of the utility of different real-time measurable predictors for the accurate generation of gaze targets. The study also compares non-temporal and temporal models for eye-gaze direction or head orientation generation and evaluates the proposed methods using different input devices.

6.3 Application

The methods proposed in this study have direct application in robotic agents that need to engage in multiparty dialogues with humans and generate human-like gaze behaviors. While the proposed methods still lack confidence scores in the speaking state, they can accurately generate gaze for listening agents similar to how humans do when listening. The benefits of appropriately generated active listening gaze behavior in robotic agents include, for example, smoother regulation of turn-taking and participation in joint actions with humans.

7 CONCLUSIONS

This study investigates the utility of eye-gaze direction or head orientation and speech activity of humans as predictors for generating real-time candidate gaze targets for a robot in multiparty open-world dialogues. The study proposes, develops and evaluates several methods to address this challenge. The methods achieve good performance in the given context, but this performance is limited to the cases when the modeled person takes a listening role.

The study presents experiments that investigate to what extent visual attention behavior can be modeled solely from low-level multimodal signals. The results show that speech activity coupled with the eye-gaze direction or head orientation is the best predictor (from the ones investigated) for eye-gaze direction or head orientation. Furthermore, the results clearly show that in this context, gaze targets generation methods that takes into account the speech activity significantly outperform ones which do not use this information.

Directions for future work include the use of hand motions as additional predictors and comparing the proposed methods with one of the more advanced multiparty heuristic methods proposed in the literature. Another direction is the development of similar methods that address the generation of candidate gaze targets when the person being modeled is speaking. Final goal is to deploy these methods on a robot and generate an actual gaze behavior. This in turn can be studied and used in multiparty human-robot interactions.

ACKNOWLEDGMENTS

This research was supported by the CHIST-ERA project IGLU and KTH SRA ICT The Next Generation.

We would like to acknowledge the NVIDIA Corporation for donating the GeForce GTX TITAN cards used for this research, and the Swedish National Infrastructure for Computing (SNIC) at the Parallel Data Center (PDC) at KTH for computational time allocation.

We would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

H. Admoni and B. Scassellati. 2014. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *Proceedings of the ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 196–199.

H. Admoni and B. Scassellati. 2017. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

J. Anderson, M. Matessa, and C. Lebiere. 1997. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction* 12, 4 (1997), 439–462.

S. Andrist, B. Mutlu, and M. Gleicher. 2013. Conversational gaze aversion for virtual agents. In *Proceedings of the Intelligent Virtual Agents*. Springer Berlin Heidelberg, Berlin, Heidelberg, 249–262.

S. Andrist, X. Tan, M. Gleicher, and B. Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, 25–32.

M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. 2005. Towards a humanoid museum guide robot that interacts with multiple persons. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. IEEE, 418–423.

D. Bohus and E. Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of the International Conference on Multimodal Interfaces*. ACM, New York, NY, USA, 1–8.

A. Borji and L. Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207.

C. Breazeal and B. Scassellati. 1999. A context-dependent attention system for a social robot. In *Proceedings of the Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1146–1153.

F. Chollet et al. 2015. Keras.

A. Colburn, M. Cohen, and S. Drucker. 2000. *The role of eye gaze in avatar mediated conversational interfaces*. Technical Report. Microsoft.

S. Frintrop, E. Rome, and H. Christensen. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception* 7, 1 (2010), 1–39.

E. Gu and N. Badler. 2006. Visual attention and eye gaze during multiparty conversations with distractions. In *Proceedings of the Intelligent Virtual Agents*. Springer Berlin Heidelberg, Berlin, Heidelberg, 193–204.

E. Hall. 1990. *The Hidden Dimension*. Anchor, Garden City, NY, USA.

S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

M. Hoffman, D. Grimes, A. Shon, and R. Rao. 2006. A probabilistic model of gaze imitation and shared attention. *Neural Networks* 19, 3 (2006), 299–310.

A. Holroyd, C. Rich, C. Sidner, and B. Ponsler. 2011. Generating connection events for human-robot collaboration. In *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 241–246.

C. Huang and B. Mutlu. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. ACM, New York, NY, USA, 57–64.

C. Ishi, C. Liu, H. Ishiguro, and N. Hagita. 2010. Head motion during dialogue speech and nod timing control in humanoid robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, Piscataway, NJ, USA, 293–300.

L. Itti and C. Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40, 10 (2000), 1489–1506.

L. Itti and C. Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3 (2001).

S. Khullar and N. Badler. 2001. Where to look? Automating attending behaviors of virtual human characters. *Autonomous Agents and Multi-Agent Systems* 4, 1 (2001), 9–23.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *Computing Research Repository* abs/1412.6980 (2014).

C. Koch and S. Ullman. 1987. *Shifts in selective visual attention: Towards the underlying neural circuitry*. Springer Netherlands, Dordrecht, 115–141.

D. Kontogiorgos, V. Avramova, S. Alexandersson, P. Jonell, C. Oertel, J. Beskow, G. Skantze, and J. Gustafson. 2018. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Paris, France.

C. Liu, C. Ishi, H. Ishiguro, and N. Hagita. 2012. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, 285–292.

B. Mutlu, J. Forlizzi, and J. Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. IEEE, 518–523.

C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. 2005. A model of attention and interest using gaze behavior. In *Proceedings of the Intelligent Virtual Agents*. Springer-Verlag, London, UK, 229–240.

C. Rich, B. Ponsler, A. Holroyd, and C. Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, Piscataway, NJ, USA, 375–382.

J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer. 2008. Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 962–967.

T. Spexard, M. Hanheide, and G. Sagerer. 2007. Human-oriented interaction with an anthropomorphic robot. *IEEE Transactions on Robotics* 23, 5 (2007), 852–862.

K. Stefanov and J. Beskow. 2016. A multi-party multi-modal dataset for focus of visual attention in human-human and human-robot interaction. In *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Paris, France.

J. Trafton, M. Bugajska, B. Fransen, and R. Ratwani. 2008. Integrating vision and audition within a cognitive architecture to track conversations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, 201–208.

H. Walker, W. Hall, and J. Hurst. 1990. *Clinical Methods: The History, Physical, and Laboratory Examinations*. Butterworth Publishers, Boston, MA, USA.

Y. Zhang, J. Beskow, and H. Kjellström. 2017. Look but don't stare: Mutual gaze interaction in social robots. In *Proceedings of the International Conference on Social Robotics*. Springer International Publishing, Cham, 556–566.