Gaze and face-to-face interaction

Gérard Bailly, Alaeddine Mihoub, Christian Wolf, Frédéric Elisei

Chapter contribution to

Eye-tracking in interaction

Geert Brône & Bert Oben

Submitted for publication to

Advances in Interaction Studies (John Benjamins)

Target: 10000-12000 words

Count: 11199 words

Title: Gaze and face-to-face interaction: from multimodal data to behavioral models

Short title: Gaze and face-to-face interaction

Gaze and face-to-face interaction: from multimodal data to behavioral models

Gérard Bailly[1], Alaeddine Mihoub[1,2], Christian Wolf[2,3] & Frédéric Elisei[1]

[1] GIPSA-Lab, Univ. Grenoble-Alpes & CNRS, St Martin d'Hères – France

[2] INSA-Lyon, Villeurbanne – France

[3] Université de Lyon & CNRS, Villeurbanne – France

1 Abstract

This chapter describes experimental and modeling work aiming at describing gaze patterns that are mutually exchanged by interlocutors during situated and task-directed face-to-face two-ways interactions. We will show that these gaze patterns (incl. blinking rate) are significantly influenced by the cognitive states of the interlocutors (speaking, listening, thinking, etc.), their respective roles in the conversation (e.g. instruction giver, respondent) as well as their social relationship (e.g. colleague, supervisor).

This chapter provides insights into the (micro-)coordination of gaze with other components of attention management as well as methodologies for capturing and modeling behavioral regularities observed in experimental data. A particular emphasis is put on statistical models, which are able to learn behaviors in a data-driven way.

We will introduce several statistical models of multimodal behaviors that can be trained on such multimodal signals and generate behaviors given perceptual cues. We will notably

compare performances and properties of models which explicitly model the temporal structure of studied signals, and which relate them to internal cognitive states. In particular we study Semi-Hidden Markov Models and Dynamic Bayesian Networks and compare them to classifiers without sequential models (Support Vector Machines and Decision Trees).

We will further show that the gaze of conversational agents (virtual talking heads, speaking robots) may have a strong impact on communication efficiency. One of the conclusions we draw from these experiments is that multimodal behavioral models able to generate co-verbal gaze patterns should be designed with great care in order not to increase cognitive load. Experiments involving an impoverished or irrelevant control of the gaze of artificial agents (virtual talking heads and humanoid robots) have demonstrated its negative impact on communication (Garau, Slater, Bee, & Sasse, 2001).

2 Introduction

The social relevance of the eyes has been largely investigated. If visually salient objects attract attention, cognitive demands of the visual search easily override contrastive properties – i.e. spatiotemporal multimodal salience – of the objects (Henderson, Malcolm, & Schandl, 2009). This is particularly the case for faces (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005) and notably of faces having direct eye contact – see Senju et al (Senju & Hasegawa, 2005) for a review. Võ et al (Võ, Smith, Mital, & Henderson, 2012) argue for a functional, information-seeking use of gaze allocation during dynamic face viewing.

The proper replication of the movement and appearance of the human eye is a challenging issue when building virtual agents or social robots able to engage into believable and smooth communication with human partners (Marschner, Pannasch, Schulz, & Graupner, 2015; Ruhland et al., 2014). We here review some key issues that pave the way towards context-

aware gaze models. The chapter is organized as follows. We first argue for the importance of getting multimodal interactive motion capture data that will enable us to study multi-party interactions as dynamically coupled systems. We then review statistical models that can capture regularities and generate context-aware behaviors. We finally draw the reader's attention to the impact of the appearance of the avatar's eye on gaze perception by human viewers and the need for taking care of every processing stage of the perception-action loop, namely the active multimodal scene analysis and comprehension, the behavior planning and execution, as well as the final rendering of movements.

3 Interactive gaze

3.1 Eyes in the visual scene

Since the seminal works of Yarbus (Yarbus, 1967), Langton (S. R. H. Langton, 2000) and Itti et al (Itti, Dhavale, & Pighin, 2003), numerous studies have questioned visual attention and proposed models to capture the lawful control parameters of scan paths of static images and videos. Visual saliency – the set of perceptual quality which makes some regions of our visual field stand out from their neighborhood – and its interplay with other senses, such as audition (Coutrot, Guyader, Ionescu, & Caplier, 2012) or touch (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2009) – has drawn much of attention from disciplines such as experimental psychology, image and signal processing or machine vision (Duffner & Garcia, 2015). Modeling bottom-up visual saliency has been the subject of numerous research efforts during the past 20 years, with many successful applications in computer vision and robotics. Recently, Borji et al (Borji, Sihite, & Itti, 2013) performed an exhaustive comparison of 35 state-of-the-art saliency models over challenging synthetic and natural image vs. video datasets. Evaluation scores typically consist in comparing human heat maps – computed by

pooling gaze data from several subjects watching the data – with saliency maps computed by the competing models. Top-down factors driven by the cognitive demand (Goferman, Zelnik-Manor, & Tal, 2012) – and notably the task – as well as the presence of agents (Schauerte & Stiefelhagen, 2014) do also strongly influence the scan paths. Borji et al notably evidence that eye fixations in video clips with many actors and moving objects get lowest scores and suggest that gaze patterns are often driven in this context by complex cognitive processes that necessitate a minimum understanding of what is going on in the (audio)visual scene.

3.2 Conversational gaze

Speaking faces are effectively salient and relevant regions of interest in a visual scene – in particular when the audio channel is available (Coutrot & Guyader, 2014; Li, Tian, & Huang, 2014). The scan path to speaking faces mainly goes through the mouth, the eyes, the nose ridge and the forehead (Buchan, Paré, & Munhall, 2007; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). The proportion of eye- vs. mouth-directed fixations has been shown to depend on cognitive demand: as an example, Lansing et al (Lansing & McConkie, 1999) have evidenced that observers spend more time looking at and direct more gazes toward the upper part of the talker's face when asked to make decisions about intonation patterns than about the words actually being spoken.

While most of the work about gaze and attention has been performed using non interactive stimuli – individual minds and brains observing representations of other people through essentially pre-recorded natural or synthetic videos – several studies have been performed on interactive gaze, i.e. in situations of sensorimotor reciprocity, i.e. situated face-to-face conversations where speakers can see and hear each other. The fact that the observer's actions cannot influence the individuals when watching static images or movies has in fact a strong impact on joint behaviors. Gaze patterns are known to differ between in situ two-ways

interaction settings vs. video replay or video simulation. Foulsham et al. (2011) have shown that people were more likely to be gazed at in a video condition than in a live condition when they were close to the observer in the scene (e.g., were approaching in order to pass by). Using more intimate settings, Laidlaw et al. (2011; 2012) further demonstrated that participants sitting in a waiting room looked at a videotaped confederate more often and for a longer duration than at a live confederate: videotaping elicits unlimited screening while live interactions respect the elementary social ground rules. Risko et al (2016) give numerous examples showing that the presence of another person can substantially alter patterns of gaze in social contexts. Studies of social attention *in the wild* have been favored by the recent availability of light-weight – and more and more discrete – mobile eye trackers.

When engaging in overt attention such as the one required during face-to-face interaction, the cognitive activity matters. Lee et al (2002) collected gaze data from one female speaker during informal face-to-face conversation. They showed that the distribution of the magnitudes of gaze shifts in listening mode is much narrower than that of talking mode, indicating that, when the subject is speaking, eye movements are more dynamic and active. Conversely, gaze of listeners is much more likely to be focused on the source of information, i.e. the speaker. Vertegaal et al (2001) measured subjects' gaze at the faces of their conversational partners during four-person conversations. They show that speakers gazed at their interlocutors about 1.6 times less than listeners. More recently, Otsuka et al (2014; 2011) have shown that conversational regimes – namely convergence, dyad-link, and divergence among multiparty conversations – as well as participants status – addressed/unaddressed participants, overhearing/eavesdropping bystanders – strongly influence gaze patterns and head directions between participants.

Most of these studies consider faces in the visual field as dynamic stimuli and potential regions of interest that can attract fixations of a target speaker according to his/her cognitive demand and his/her role in the conversation. Few studies have nevertheless considered gaze patterns has the by-product of a coordinated action, i.e. sensible consequences of an underlying coupled system in which the interlocutors play an active role and coordinate behaviors for sharing common grounds and goals (see notably the importance of the speech channel in Richardson, Dale, & Kirkham, 2007).

Figure 1. First experimental setting used by Bailly et al (G. Bailly, Raidt, & Elisei, 2010) to study mutual gaze patterns using computer-mediated face-to-face interactions.

3.3 Mutual gaze patterns

Settings where gaze patterns of all parties involved in the conversation are monitored in parallel with other modalities (e.g. voice, body, head, face and hand gestures) are rare. Several studies have of course examined mutual multimodal behaviors using synchronous videos (Cummins, 2012) and manual annotations, but the accuracy of gaze estimation by human viewers can hardly go beyond the basic contrast between eye contact vs. gaze aversion. Bailly

et al (G. Bailly et al., 2010; Raidt, Bailly, & Elisei, 2007) designed a computer-mediated face-to-face interaction with two pinhole cameras and two Tobii® eye-trackers both embedded onto two computer screens that displayed live videos of the interlocutors (see Figure 1). Using a similar setup, Barisic et al (2013) used dual eye-tracking to investigate real-time social interactions: they eliminated the problem of live video capture by tele-representing the interlocutors by virtual avatars (see also the experiment performed by Boker et al., 2009 where they manipulate the control parameters of Active Appearance Models). The Barisic et al system was inspired by Carletta et al (2010) who demonstrated a dual-tracking system using an experimental paradigm for cooperative on-screen assembly of two-dimensional models. More recently, Brône and Oben (2015) recorded several face-to-face interactions with two head-mounted eye-trackers and associated scene cameras.

In our experiments with dual videos and eye-trackers (G. Bailly et al., 2010), we analyzed the typical distributions of the fixations and blinking rates of one target female participant over the facial elements (eyes, mouth, nose ridge, other parts of the face) of the face of her 10 different interlocutors and the mirror distributions of the interlocutors' gaze on her own facial parts. We showed that these distributions depend on their joint cognitive states, e.g. speaking turns are almost always associated with an eye contact and more specifically with a saccade of the speaker's gaze towards the right eye of the interlocutor, speakers mainly monitors the eyes of their interlocutor while listeners monitors their lips, etc. The interaction scenario was a speech game where interlocutors have to read, utter and repeat so called Semantically Unpredictable Sentences (SUS) (Benoît, Grice, & Hazan, 1996) such as "the hammer fires the cake that spikes". These utterances are quite difficult to recognize: when the speaker reads aloud a sentence for the first time, the speaker and the listener have respectively to speak clearly, lip-read and monitor the others' gaze to ensure that the message is correctly passed

over. When the listener repeats back what he/she has understood, the respective speaking and listening conditions differ: the text giver knows the textual content and "only" listens to check if the text is correctly spelled out by the receiver. The role of the speaker or listener in the conversation has an impact on the a priori knowledge of what will be exchanged and we expect the role to influence the multimodal behaviors of the speakers. Each member of the dyad was thus text giver and receiver in alternation. We showed that their respective roles and a priori knowledge on the exchanged linguistic content both impact the gaze patterns and the blinking rate, e.g. blinking frequency is much higher while speaking (here an average of 0.6 blinks per seconds) than while listening (0.1 blinks per seconds) and blinking frequency is almost null when receivers listen to the giver's first reading. Note that large head shifts (such as occurring when the text giver finishes reading the target sentence and gets ready to speak to the receiver) are also systematically accompanied by a blink.

4 Learning & generating gaze patterns

4.1 Grounding gaze patterns

Several generative models of gaze patterns have been proposed. The most long-winded line of research has been initiated by Itti et al (Itti et al., 2003; Itti, Dhavale, & Pighin, 2006) who proposed a photorealistic attention-based gaze animation that is grounded on a model of saliency and a biological model of the eye/head saccade subsystem. Itti et al propose a winner-takes-all strategy for allocating the current fixation to the most salient region of a map that combines bottom-up saliency with top-down task-relevance and attention guidance. Sun (Sun, 2003) proposed a hierarchical saliency model that first decomposes the scene into a pyramid of regions of interest and further constraints the fixations to first exhaust salient sub-regions of an image before switching to another region. Raidt et al (Picot, Bailly, Elisei, &

Raidt, 2007) augmented the Inhibition of Return (IoR) mechanism proposed by Itti et al for attention guidance with a stack of attention – in which are stored the position and local appearance of the most recent *N* (=4) regions of interest that have been fixated – and added an object recognizer that triggers object-specific scrutinization mechanism, in particular when detecting a face. *A priori* knowledge about important regions of interest (objects, faces, etc.) is then easily recruited and ad hoc gaze patterning can be triggered.

Note that generation of gaze patterns can also be grounded on speech signals. A number of systems use speech as an input from which to generate facial expressions involving the mouth, head, eyes, and eyebrows (Albrecht, Haber, & Seidel, 2002) – for a review of data-driven mapping techniques addressing this problem see (Ruhland et al., 2014).

We mentioned in section 3.2 that the structure of conversation has a strong impact on gaze patterns. The generation of gaze paths should thus benefit from an incremental estimation of the cognitive, psychological and physiological state of the interlocutor(s) as well as the locutionary and illocutionary contents of the speech acts. Several authors have proposed statistical models that generate gaze patterns in context. Lee et al (Lee et al., 2002) proposed a statistical eye movement synthesis model for gazing at faces that exploits empirical distributions of durations of fixations and amplitudes of saccades depending on the talking/listening mode of the speaker. Vinayagamoorthy et al (Vinayagamoorthy, Garau, Steed, & Slater, 2004) and Gu et al (Gu & Badler, 2006) further refined the model for virtual characters. There is a rich set of models that exploit empirical distributions for various mechanisms related to the conversational structure (e.g. topic-signaling, turn-taking, etc.) or participant characteristics (e.g. roles, social status, etc.) Gaze patterns are typically described as automata (Mutlu, Kanda, Forlizzi, Hodgins, & Ishiguro, 2012) or belief networks

(Pelachaud & Bilvi, 2003) and trigger saccades according to empirical means and standard deviations of spatial and temporal gaze parameters.

Gaze is thus both conditioned by both bottom-up information (i.e. multimodal input) and top-down cognitive demands, especially for establishing and monitoring socio-communicative relations.

4.2 Learning joint behaviors

If several researchers have studied the conversational dynamics (Cummins, 2012; Dale, Fusaroli, Duran, & Richardson, 2013; Fusaroli & Tylén, 2016; Schmidt, Morr, Fitzpatrick, & Richardson, 2012), few works have tried to actually model the links between multimodal behaviors of interlocutors mediated by the structure of their conversation and use these models to predict multimodal joint behaviors. The seminal work of Pentland and colleagues on social signal processing (Pentland, 2004, 2007) has opened the route to both the inference of paralinguistic information from raw signals exchanged during social interaction but also to the generation of such social signals for recommendation systems or autonomous agents. They built a computational model based on Coupled Hidden Markov Models (CHMMs) to characterize the dynamics of dyadic interactions. The degree of coupling was shown to correlate with the success of the intended goals. The work of Otsuka et al (2011; 2005) is also a very inspiring landmark: they proposed to use a Dynamic Bayesian Network (DBN) to estimate addressing and turn taking ("who responds to whom and when?") and predict gaze shifts between participants of a multi-party conversation. Speech activity, head and gaze shifts across participants were here mediated by the conversational regime (see section 3.2).

The progress of machine learning techniques offers very powerful tools to mine multimodal scores. They offer elegant and efficient ways to perform decision or regression

tasks. It is quite tempting to use regression tools to perform a direct mapping between input features (observed behaviors, a priori contextual knowledge) and desired output behaviors. Thus, Ishii et al. (2014) proposed a support vector machine (SVM) to establish a direct mapping between gaze transition patterns and the timing of speech turns in multi-party meetings. In this context, interaction sequences are represented as temporal sequences where one temporal unit is often the "frame", i.e. one instant of time in an (audio-)visual sequence whose duration depends on the acquisition frequency, typically around 40ms. Frame-based classifiers are very sensitive to the placement of the analysis window and often exhibit noisy temporal output sequences when the window is sliding over the input. These drawbacks are also exhibited by non-deterministic mapping techniques such as Gaussian Mixture Regressors (GMR).

Sequential models such as Hidden Markov Models (HMM) or Dynamic Bayesian Networks (DBN) partially resolve these issues by mediating the correspondence between input/output observations via hidden states or latent variables. These elementary temporal units segment the interaction into homogenous spatio-temporal patterns that can be then combined into larger interaction units. These interaction units can then combine these elementary patterns according to a task-specific syntax and model complex joint sensorimotor behaviors by splitting the regression problem into task-dependent subspaces.

For an introduction to these regression techniques that link input to output observations, see the I/O HMM proposed by Bengio & Frasconi two decades ago (Bengio & Frasconi, 1996). Semi-Hidden Markov Models (Mihoub, Bailly, & Wolf, 2014) have interesting properties for modeling and controlling the durations of these joint sensorimotor states – i.e. the hidden states of the Markov chains that link input observations with desired output

features. Moreover, Dynamic Bayesian Networks combine time dependency and structural constraints (with latent variables) with direct causal relations between multimodal features.

4.3 A sample interactive game

We illustrate these concepts through results of recently performed experiments on multimodal face-to-face interaction (Mihoub, Bailly, Wolf, & Elisei, under revision), where gaze was studied together with other signals, such as gesture and speech. Gestural deixis usually involves the combination of "what" information – using deictic words (this, that,…) or the name of the object/agent – and "where" information – using deictic gestures such as head, gaze, body or finger pointing**.** For that purpose, we designed a game inspired by the famous "put that there" paradigm (Bolt, 1980). This interactive scenario – simple as it can appear at first sight – is a very interesting benchmark for studying and learning human strategies used to maintain mutual attention and coordinate multimodal deixis of objects and places – similar to the visual worlds used by Tanenhaus and colleagues (Allopenna, Magnuson, & Tanenhaus, 1998) and Clark (2003).

The interaction consists in a cube game involving an instructor and a manipulator (Fig. 2), the latter following orders of the former, which are typically formulated like "*Put the red dotted cube at the left of the one with the green cross"*. The task is collaborative: the instructor is secretly informed (via a sketch displayed on a tablet) about the pattern to be reproduced by asking the manipulator to move cubes from a source manipulator space to a target chessboard (see Figure 2). The objective of the statistical model is to learn and reproduce the instructor's coverbal behaviors in terms of gaze and gesture given his/her speech and behavior of the interlocutor. This statistical model may be then transferred to a conversational agent (virtual avatar, lamp avatar or humanoid robot) capable of instructing a human manipulator.

Figure 2: ego-centric view as seen from the instructor and extracted from the cube game experiment. The current fixation point of the right eye – monitored by a Pertech® head-mounted eyetracker – is cued by a back circle (here surrounding the red cube pointed by the hand of the instructor). Time stamps are used to synchronize multimodal streams.

The signal flow between the two participants is modeled by capturing several social signals: the manipulator gestures (MP), instructor speech (SP), instructor gesture (GT) and instructor gaze (FX). In order to learn generic behaviors, all signals are discrete and refer to a limited set of possible references:

- *manipulator gestures* are distinguished as: rest, grasp, manipulate, end, none
- *instructor gestures* are discretized through a dictionary of 5 regions of interest: rest, cube to be displaced, position of target tile, position of the reference cube, none
- *gaze* also refers to one of 8 possible regions of interest: manipulator's face, source manipulator space, chessboard, cube to be displaced, position of target tile, position of the reference cube, tablet, none

- *verbal instructions* are discretized into 5 elements corresponding to the key lexical elements: cube to be displaced, position of target tile, position of the reference cube, else, none

These discrete variables were annotated semi-automatically by only one expert: FX and GT were segmented automatically but labelled by-hand in order to avoid the unnecessary development of target identification algorithms; SP was first aligned with speaker-independent phonetic models and further checked by hand; MP strokes are completely segmented and annotated by hand. Consistency of this multi-stream labelling is essentially post-hoc checked using so-called coordination histograms (Mihoub, Bailly, Wolf, & Elisei, 2016). Modality-specific micro-controllers are then supposed to generate continuous segment-specific (i.e; arm, eye, head) movements as well as speech from these discrete instructions (see our recent evaluation of such a framework in Nguyen, Duc-Anh, Bailly, Gérard, & Elisei, Frédéric, 2016).

We also suppose that the underlying cognitive task follows a specific syntax, which is related to the structure of the interactive task. This syntax is modeled through an intermediate layer, mediating between low-level observations, called *interaction unit (IU)* in line with Ford et al (Ford, 2004), which takes 6 values in our experiments: getting instructions, seeking cube, pointing the cube, pointing the destination position, verification, validation.

In the following, we analyzed a set of 30 game plays in which the instructor interacted with 3 different partners (10 game plays with each one). Each game play consists in placing 10 cubes given an entirely filled manipulator space (16 cubes) and an empty task space. The first cube should be placed in the center tile of the chessboard. The mean duration of a single game is around 1 minute and 20 seconds (~2000 frames, 40ms per frame).

4.4 Learning joint behaviors with dynamic Bayesian networks

Given this experimental setting, the question arises how these different – observed or latent – variables interact and whether causal relations exist between them. The motivation for this analysis is twofold. First, the derivation of a relational graph is an interesting scientific result in itself, which can provide valuable insights into the underlying cognitive process. Secondly, with respect to the goal of this study, causality graphs can be used as a modeling tool for the design of efficient inference algorithms capable of predicting desired variables (here coverbal actions) given observed quantities (here verbal actions).

Of course, the contingent causality relations between the underlying cognitive processes are hidden and as such cannot be retrieved with absolute certainty. Statistical models provide estimations for these relations, which are derived using different mathematical concepts such as correlation and mutual information. The resulting so-called causality graphs provide information on conditional independence properties of the variables of the system. In particular, for each variable A of the model, the graph provides the so-called Markov blanket ∂A, defined as the set of variables which, when conditioned on it, make all other variables independent of A.

In the case of time series, where each considered variable is present for each time instant of a sequence, dynamic Bayesian networks (DBN) have been established as an important tool for modeling structured problems, for learning and inference. They are particularly attractive and useful for modeling the dynamics of multimodal behaviors in face-to-face interactions (Huang & Mutlu, 2014). DBNs are directed acyclic graphs in which nodes represent random variables and edges represent conditional dependencies. Semantically and intuitively an edge from a parent node X to a child node Y means that node X has influence over node Y. An exact description of how independence statements can be derived from the graph is beyond the

scope of this chapter. The interested reader is referred to (Koller & Friedman, 2009) and (K. Murphy, 2002).

In some situations and depending on the application, this dependency structure may be manually provided by an expert of the target domain. Alternatively, several statistical methods have been introduced to learn the graphical structure of a DBN automatically from data (Trabelsi, Leray, Ben Ayed, & Alimi, 2013). In our application, our DBN structure (see Figure 3) has been entirely learned from training data. The intra-slice structure is learnt using the K2 algorithm (Cooper & Herskovits, 1992). The inter-slice structure is learned using the REVEAL algorithm (Liang, Fuhrman, Somogyi, & others, 1998). We employed the Bayes Net Toolbox (K. P. Murphy, 2001) for training and inference. The resulting causality network (see Figure 3) presents very interesting intra-slice properties such as:

- The interaction units influence both perception and action streams (black arrows), and thus paces the joint behaviors
- The instructor reacts to the manipulator actions (MP impacts SP, GT and FX) (blue arrows)
- The speech activity (SP) of the instructor influences his co-verbal behavior (GT and FX) (green arrows). This is consistent with co-verbal contingency (McNeill, 1992)
- Each random variable (slice t+1) is influenced by its history (slice t) (gray arrows)

… as well as inter-slice properties that cue the causal relations within the perception-action loop, notably:

- The deictic chain that chains gaze, pointing gesture and verbal indexing (FX → GT→ SP) leads to an effective manipulation (→ MP)

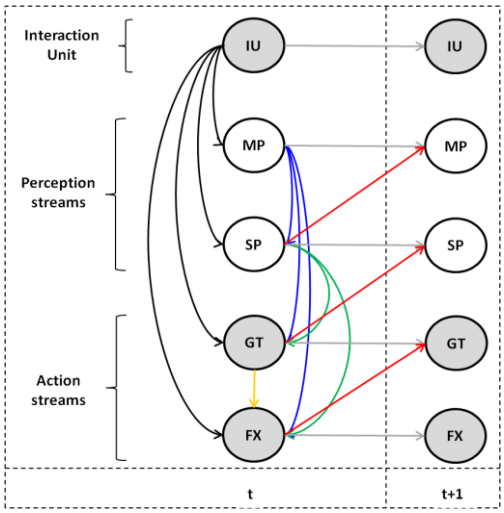

Figure 3. The DBN model learned from training data. Variables in gray circles are to be predicted in the inference stage.

The second goal of our work is to be able to infer desired hidden quantities from observed quantities, in particular in a real-time setting with low latency (i.e. limited look-ahead observations) so that these behavioral models can be used to control artificial agents engaged in effective interactions with humans. To this end, the proposed models should be able (1) to estimate the interaction units from perceptual observations (speech activity/ manipulation of the partner); when the two partners cooperate, the sequential organization of the interaction units should ideally reflect the shared mental states of the conversation partners at that particular moment; (2) to generate suitable actions (hand gestures of the instructor and his own gaze fixations) that reflect his current awareness of the evolution of the shared plan. We used the junction tree algorithm (Jensen, Lauritzen, & Olesen, 1990) to perform offline estimation by computing the MPE (most probable explanation) of IU, GT and FX given the whole sequence of MP and SP. The junction tree algorithm gives an exact solution of the estimation problem, i.e. the inferred variable is the most probable one according the probabilistic formulation.

Table 1 gives the results of the proposed method compared to a more classical setting with Hidden Markov Models (HMMs). For all models, 30-fold cross validation was applied. The Levenshtein distance (Levenshtein, 1966) is adopted for the evaluation because it computes a warped distance between predicted and original signals, which is tolerant to minor misalignments. In particular, this avoids getting extremely low scores (near zero) in presence of small latencies.

While the graphical structure was entirely learned in the DBN setting, HMMs are characterized by an imposed structure consisting of (i) hidden variables satisfying the Markov property and (ii) observed variables, which are conditionally independent from each other given the hidden variables. Relaxing these restrictions translates into higher classification rates, as can be seen in table 1. While the improvement of the estimation of instructor deixis (direction of finger) is minor, large gains are obtained with respect to the estimation of the interaction unit, and the gaze direction. In particular, this shows that gaze is a complex phenomenon whose estimation can be significantly improved if the conditional dependencies on other variables are taken into account correctly. The fixed dependency structures of classical HMMs seem to be too restrictive in this context.

Table 1. Estimation performance of the DBN model compared to classical hidden Markov models.

	IU	**GT**	**FX**
DBN	85%	87%	71%
HMM	72%	85%	60%

Because of the Levenshtein distance, these results neglect minor misalignments between the reference and the generated scores. We also compared inter-modal synchronization patterns using so-called coordination histograms (Mihoub et al., 2016). We showed that DBN also better reproduce the natural micro-coordination between multimodal streams. In fact, HMM impose to model all transitions between discrete observations at onsets of hidden states.

HMM and DBN may be improved to cope with interactive and highly rhythmical patterns. We have shown (Mihoub, Bailly, & Wolf, 2015) that semi-HMM that explicitly model the duration of hidden units – so called state occupancy – better capture sensorimotor loops. Similar proposals have been recently done for DBN (Donat, Bouillaut, Aknin, & Leray, 2008). Note finally that Deep Neural Networks (DNN) able to cope with highly structured sequences such as Long Short-Term Memory (LSTM) or Clockwork Recurrent Neural networks (CW-RNN). DNN (Hochreiter & Schmidhuber, 1997) (Sak et al., 2014) offer performative alternatives to Graphical models when large training data is available.

4.5 Adapting joint behaviors

One challenge of the original proposal of Pentland et al. (Pentland, 2004) was to observe and characterize the dynamics of the social glue – i.e. activities or interactions that strengthen the relational ties in a group of people (Lakin, Jefferis, Cheng, & Chartrand, 2003) – via dynamical models of multimodal joint behaviors. We have shown (Mihoub, Bailly, Wolf, & Elisei, 2015) that models of mutual gaze patterns can in fact implicitly capture social features that are encoded via very shallow signals that may escape to human expertise. For instance, for the speech game described in section 3.3 where a female speaker interacted with 10 different interlocutors, we computed distances between datasets and models of gaze behaviors of different dyads (i.e. applying model trained on interlocutor A to the dataset of interlocutor

B). We then performed multidimensional scaling (MDS) on the distance matrix. Analysis evidenced the significant impact of pre-existing social relationships (colleagues vs. students) between interlocutors. Such data mining techniques can be used to detect meaningful dimensions that structure the interactive human behaviors. By using repertoires of behavioral models – or more comprehensive statistical models – trained on multiple dyads or social groups, one may expect to faithfully select and adapt autonomous systems to their audience, notably the role it has to play in the conversation and the social relationship it wants to establish. As an example, De Kok et al (2013) used so-called "speaker descriptors" (mean and standard deviation of pitch and energy, speech rate and average gaze shift per minute) to select an appropriate model of back channeling – trained as Conditional Random Fields (CRF) models – amongst a collection of pre-analyzed dyadic interactions.

4.6 Effective gaze tracking and generation

Note that the behavioral models proposed above rely on (1) an active visual scene analysis that should deliver estimations of the gaze direction of the conversational partners as well as track the positions of potential objects of interest and (2) a faithful gaze generator that effectively direct the agent's gaze towards the intended targets. Section 5 sketches the current state of the art concerning non-invasive gaze estimation. Section 6 further underlines the importance of accurate gaze control and rendering.

5 Active gaze estimation from images and videos: gaze patterns and interaction models

If behavioral models can be trained using data collected on interlocutors with invasive motion capture systems – such as head-mounted eyetrackers or other wearable sensors – autonomous agents should rely on egocentric sensors such as embedded cameras.

The direct estimation of gaze direction from images or videos can be a hard challenge according to the chosen experimental setup. The best standard solutions use special hardware with multiple head-mounted cameras often operating in the infrared spectrum. Although complex, these solutions can now be miniaturized enough to be integrated into mobile devices, and latest technologies allow eye-tracking to be integrated into head-mounted gears like the Google® or SMI® glasses.

Gaze estimation *in the wild* tries to solve this problem from RGB images or images taken from RGB-D (consumer depth) cameras. A major challenge here is to be able to generalize to different head poses and to different individuals. Calibration to the subject at hand is a preponderant methodology, although auto-calibration and calibration free methods are on the rise. Estimation in the wild requires preliminary face detection or head tracking. This estimation is greatly enhanced by depth information such as provided by RGBD sensors such as the Kinect®. As an example, Funes-Mora and Odobez (2014) learn a user-specific 3D head model in an off-line stage. During on-line estimation, the 3D head pose is tracked by aligning new 3D data with the model using iterated closest points (ICP) initialized with a Viola-Jones face detector.

Methods on gaze estimation itself can roughly be classified into two families of approaches. Geometric methods fit a 2D or 3D model of the eye to data, as for instance in

(Valenti & Gevers, 2012). These methods are often chosen when specific hardware is available in a multi-camera setup and/or when data quality is high. Appearance based methods, on the other hand, use direct regression of gaze from appearance features learned from training samples. Issues are the quality of training data in terms of the resolution of the input eye images, and in terms of number of subjects; the ability to generalize; and the problem of obtaining reliable training labels in the case of supervised learning. As an example, Sugano et al (Sugano et al. 2014) proposed a regression plus synthesis approach, where random forests are trained in an offline stage on a mixture of real and synthetic data, which has been created by 3D reconstruction from multiple cameras. In Funes-Mora and Odobez (2014), gaze direction is estimated in the head coordinate system (after head tracking) using regression from histogram of oriented gradient (HoG) features and then mapped back to the global coordinate system. In Duffner and Garcia (2015), visual focus of attention (VFOA, i.e. discrete gaze information restricted to a set of chosen focal points) is inferred with an HMM. In a sequential setting, particle filtering tracks faces and VFOA jointly. Again, the observation model resorts to image primitives such as HoG features or color histograms.

Note that recent methods augment image-based information with contextual cues, such as multimodal contingency, visual saliency, learned gaze patterns and other interaction models. The objective is to leverage the strong linkage between gaze and other verbal and non-verbal signals in human interactions. These contextual features are exploited to improve the quality of gaze estimation or, alternatively, to contextualize training labels.

In Sugano et al (2013), gaze is estimation is combined with visual saliency, i.e. *a priori* information on the attractiveness of certain locations in the image. Saliency is extracted through face detection, as faces are more likely to be looked at, and additional low-level information calculated from texture. The saliency information is accumulated into Gaze

probability maps, which are used as input (soft targets) to train Gaussian process regressors. Alnajar et al (2013) exploit human gaze patterns, which are learned in an offline stage. The goal is to perform gaze estimation in an uncalibrated setting: initial gaze sequences are first estimated through classical regression and then aligned with known template gaze patterns.

Such signal-dependent bottom-up information are often complemented with a priori top-down information such as the possible regions of interest (RoI) in the scene, the dialog structure, the respective roles of the speakers, the possible conversational regimes, etc. As an example, Sheiki and Odobez (Sheiki and Odobez, 2014) modeled gaze patterns during multi-party meetings and exploit the very properties of these interaction scenarios, notably speaking activity – i.e. people tend to look at speakers –, verbal content – i.e. people tend to look at verbally referenced objects, with a mean delay of 2s (Richardson, Dale, & Shockley, 2008) – and topic – gaze mirrors mental state (Teufel, Alexis, Clayton, & Davis, 2010).

Furthermore, gaze estimation may benefit from other modalities. The kinematic chain of attention involves the whole body, from the orientation of feet and body to the orientation of head and eyes. Conversely, the orientation of all these segments contribute to gaze estimation (Hietanen, 1999) (S. R. Langton, Honeyman, & Tessler, 2004). Thus, the so-called Midline effect (Fuller, 1992) (Hanes & McCollum, 2006) rules the relation between gaze and head orientation.

Finally, gaze estimation should be linked to action and scene comprehension. Active perception refers to the ability of agents to act to better perceive (see Bajcsy, 1988 for a general theoretical framework for active perception). Such actions comprise self-motion (e.g. moving away from interfering sources or near to sources of interest), verbal (e.g. asking interlocutors to play again) as well as nonverbal communication (e.g. displaying facial expressions expressing doubt or surprise) so that to renew percepts. Ferreira et al (2013) have

notably proposed a Bayesian framework for multimodal perception through an active attentional and behavioral exploration of the environment. Optimal exploration can be then expressed in terms of various criteria such as entropy minimization or maximization of reward such as *a posteriori* probability of given events.

Active visual perception also refers to the fact that humans should move eyes to perform a fine-grained analysis of a given region of interest (RoI). In fact, the 6 to 7 million cones that provide the eye's color sensitivity are much more concentrated in the central spot known as the macula. In the center of that region is the "fovea centralis", a 0.3 mm diameter rod-free area with very thin, densely packed cones. Saccades are performed to bring the RoI into the fovea. There are thus major consequences: (1) the input vision stream that delivers information on the actual state of possible ROI is full of *missing data* since this information is only renewed by overt attention shifts: estimation of the gaze of others is thus performed on demand of behavioral models (see early work conducted by Yarbus, 1967), i.e. when such information is necessary to keep track of another's intentions; (2) this action-for-perception is one of the basic mechanism for shared attention mechanisms and, more generally, theory of mind models (ToM) (Baron-Cohen, Jollife, Mortimore, & Robertson, 1997) that conversely enable others to infer our own intentions. Mutual gaze reading is thus a perquisite for effective intelligent interaction. Easing gaze reading for the conversational partners is a key issue for the development of social avatars and androids.

6 Easing gaze reading

The generation of task-relevant and interlocutor-adaptive gaze patterns is of course a crucial step towards building credible human-agent interactions. These high-level control strategies should not conceal the low-level control and embodiment issues. The embodiment

of the artificial agent matters and may bias the perception of the intended movements by human viewers.

As outlined previously, the estimation of the gaze direction of others is a complex by-product of multiple cues that involve eye-related features – notably the position of the iris in the eyelid opening – as well as features related to other deictic features including low-level cues such as body and head posture and high-level cues such as a priori information on potential targets. We demonstrate below that several eye-related features impact the gaze reading.

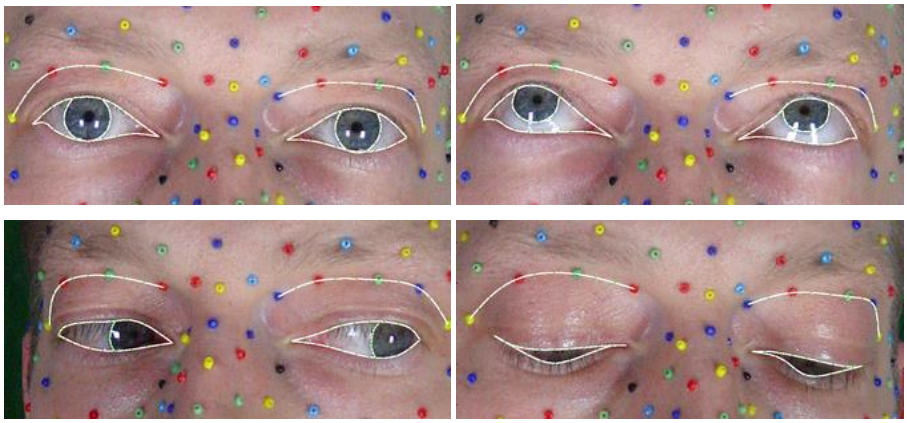

Figure 4. Photogrammetric data showing the deformation of eyelids according to gaze direction (from Gérard Bailly, Elisei, Raidt, Casari, & Picot, 2006)

6.1 Eye appearance

Trutoiu et al (Trutoiu, Carter, Matthews, & Hodgins, 2011) studied temporal and spatial deformations of eyelids when blinking and showed that viewers are quite sensitive to the dynamics of eye blinking. Elisei et al (Elisei, Bailly, & Casari, 2007) showed empirically that gaze shifts are accompanied with eyelids movements (see Figure 4). Oyekoya et al (Oyekoya,

Steed, & Steptoe, 2010) confirmed that eyelid movements play an important part both in conveying accurate gaze direction and improving the visual appearance of virtual characters.

In a series of books (M. Tomasello, 2009; Michael Tomasello, 2008), Tomasello and his colleagues notably propose that the phylogenetic specificity of humankind rests in its species-specific adaptation for sociability. The account offered by Tomasello contrasts human cooperation and altruism with nonhuman primate competition, and proposes that human altruism leads to shared intentionality (the ability to share attention to a third party – object or agent – and, more generally, to share beliefs and intentions). Tomasello (Michael Tomasello, Hare, Lehmann, & Call, 2007) further proposed the cooperative eye hypothesis (CEH). The CEH suggests that the eye's visible characteristics evolved to ease gaze following. Kobayashi & Kohshima (Kobayashi & Kohshima, 2001) have notably shown that humans have the smallest iris proportion in the eye opening and the largest contrast between iris and sclera colors among the primate and non-primate species with eyes.

Conversely in HAI, humans expect social agents to offer back cooperation, altruism and share goals and plans. Such cooperative behavior will also be favored by the agents' gaze readability. This readability is both a control and a design issue: the eyes should be controlled and move in an appropriate and predictive way but should also be designed so that the eye's visible characteristics are similar to those that humans have developed for the sake of social interaction.

6.2 Estimating gaze direction of avatars

Several studies have shown that multiple cues influence the estimation of gaze direction. Gaze direction is a complex by-product of body, head and eye orientation. Contextual features combine with these bottom-up cues to direct attention. There is surprisingly few works

assessing the perception of the gaze of virtual or robotic agents by human observers (see however Cuijpers & van der Pol, 2013 experiments with the Nao). Remarkable experiments have been conducted by Al Moubayed et al (S. Al Moubayed, Edlund, & Beskow, 2012; Samer Al Moubayed, Skantze, & Beskow, 2012) with a lamp avatar called Furhat. They notably compared the estimation by human observers of the gaze of Furhat, its virtual model and the video of its performance both displayed on screen. They showed that Furhat is the only display escaping from the Mona Lisa effect[1] (Gregory, 1997) and delivering very accurate gaze direction, independently of the observer's viewing angle. Similar experiments have been conducted by Delaunay et al. (Delaunay, Greeff, & Belpaeme, 2010). Onuki et al. (Onuki, Ishinoda, Kobayashi, & Kuno, 2013) compared the impression given by mechanical eyes and a lamp avatar: they concluded that eyes with a round outline shape and a large iris were most suitable for precision and subjective agreement.

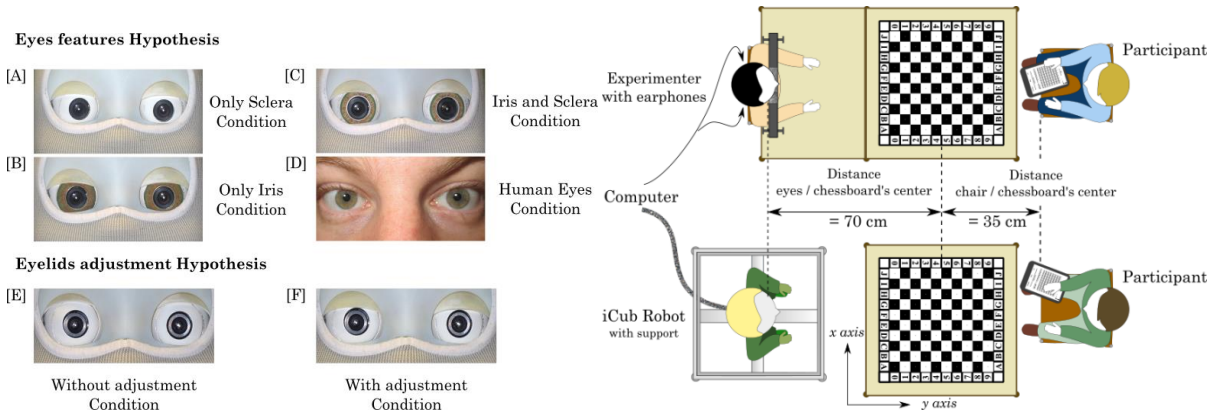

Figure 5. The robot's and human's eyes and robot's eyelids used for the eye direction experiment, accordingly to our hypotheses. [A]: eyes with no iris; [B] eyes with large colored iris caps; [C] eyes with human-sized colored iris caps; [D] human eyes; [E] robot's eye gaze without adjustment of eyelids position; [F]

[1] A person depicted in portrait paintings does not appear slanted even when observers move around. Moreover its gaze seems to follow you when its gaze is facing the original view.

robot's eye gaze with eyelids position adjusted. Right: the interaction set-ups where the participants are either faced to a robot with 3 different iris sizes or a human informer gazing at given tile of a chessboard that they were asked to guess.

We also performed (Foerster, Bailly, & Elisei, 2015) a comparative evaluation of the impact of the iris size and the coordination between eye direction and eyelid aperture for the estimation of gaze direction of our iCub humanoid robot Nina by human viewers (see above). We show that the coordination between eye direction and eyelid aperture significantly contribute in reducing estimation errors. We confirmed the findings of Onuki et al. for the benefit of endowing avatars with large irises. We also compared the performance of the robot with that of a human informer: surprisingly the robot outperformed the human challenger!

7 Future trends

We addressed the challenge of endowing avatars with social gaze. We have demonstrated that these avatars should pay attention to the analysis of audiovisual scene they step in but also to the overt behaviors and estimated intentions of the other agents sharing the environment and conversing with them. We argue for the benefits of building statistical models of multimodal behaviors from human demonstrations, i.e. by collecting traces of exemplary interactions comprising gaze tracks and behavioral signals of all interlocutors together with the estimation of underlying organization of conversational structure and goals.

Figure 6. Beaming the GIPSA-Lab iCub: the human tutor (left) monitors the head and eyes of the iCub robot (right) while perceiving (viewing and listening) the remote scene via a head mounted display that plays back the audiovisual streams captured by the cameras and microphones embedded into its eyes and ears. This cognitive gift artificially provides the robot with situated social skills.

Such a supervised training faces numerous issues. When performing off-line on collected traces, the performance of current statistical models – even on simple and controlled scenarios involving a reduced set of conversational units – is still far from perfect. Big data is certainly required to inspect the multiple factors that may influence behaviors, along the ever-changing linguistic, paralinguistic and nonlinguistic dimensions of social interactions. But the use of such off-line models for monitoring one-line interactions is still an issue. Moreover the retargeting of human behavior on artificial embodiments faces two main challenges: (a) the source and target degrees of freedom have different properties in terms of dynamics, kinematics and appearance; (b) the expected behaviors of human interlocutors – that are heavily conditioning the input features of predictive behaviors – will be impacted in an unpredictable way by the retargeted behavior and appearance of the avatar. We are presently exploring an original way of coping with this double challenge by immersive teleoperation (Gérard Bailly, Elisei, & Sauze, 2015): the human tutor provides a robot with social behaviors

by perceiving and acting in the scene though its robotic effectors (see figure above). The human tutor provides the cognitive abilities and the robot the sensorimotor affordances. The robot stores these passively-experienced behaviors into a behavioral memory it will then mine to build socially-effective models. More autonomous strategies – such as developmental learning or learning by curiosity – should take over such a human bootstrapping procedure and replace direct supervision with indirect reward.

We believe that off-line learning of behavior models from massive amounts of data (big-data) will further boost the recognition and predictive performance of the discussed data-driven methodologies. Recently, deep neural networks have been rediscovered in computer vision and machine learning and proven to be extremely efficient, in particular for sequential data (Karpathy et al., 2014) . Handling multi-modality is increasingly shown to be important in these models, where combining and modeling audio and video channels can provide significant gains in applications like audio-visual speech recognition (Ngiam et al., 2011) and audio-visual gesture recognition (Neverova, Wolf, Taylor, & Nebout, 2016).

While the availability of massive amounts of training data has been beneficial to various fields of research, it can be argued that supervised learning using annotated data had most impact in a majority of cases, e.g. in visual object recognition trained on >1 million annotated images (Krizhevsky, Sutskever, & Hinton, 2012). These amounts of data are currently unavailable in face-to-face interaction, and it might be argued that a large effort by the community is necessary in order to create a corpus of sufficient size.

In scientific terms, we conjecture that research in data-driven learning of behavior models from massive amounts of data will require tackling the task of learning hierarchical models capable of learning interactions in several layers of abstraction: high level components addressing the important cognitive aspects we also dealt with in this chapter (turn taking,

back-channeling etc.), low level components modeling information related to the scene, which is often of geometric nature (spatial arrangements between the actors in the scene and their body parts, positions of various objects of interest in the scene etc.), as well as intermediate levels of representations between these two extremes. We believe that semi-supervised learning and weakly-supervised learning of DNN will bring advances to this field.

8 Acknowledgments

This research supported by ANR (SOMBRERO ANR-14-CE27-0014, Robotex ANR-10-EQPX-44-0 and Persyval ANR-11-LABX-0025), the Rhone-Alpes region (ARC6) and the UJF (EMSOC 2012084RECF991). We also want to thank François Foerster, Carole Plasson and Miquel Sauzé for their valuable contributions. A special thanks to Ghatfan Hasan for keeping Nina alive. We thank the two anonymous peer reviewers whose comments and suggestions greatly improved the initial draft of this paper.

9 References

Al Moubayed, S., Edlund, J., & Beskow, J. (2012). Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. *ACM Transactions on Interactive Intelligent Systems*, *1*(2), article 11 (25 pages).

Al Moubayed, S., Skantze, G., & Beskow, J. (2012). Lip-reading: Furhat audiovisual intelligibility of a back-projected animated face. *Intelligent Virtual Agents - Lecture Notes in Computer Science*, *7502*, 196–203.

Albrecht, I., Haber, J., & Seidel, H.-P. (2002). Automatic Generation of Non-Verbal Facial Expressions from Speech. In J. Vince & R. Earnshaw (Eds.), *Advances in Modelling, Animation and Rendering* (pp. 283–293). Springer London. Retrieved from http://dx.doi.org/10.1007/978-1-4471-0103-1_18

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Alnajar, F., Gevers, T., Valenti, R., & Ghebreab, S. (2013). Calibration-free gaze estimation using human gaze patterns (pp. 137–144). Presented at the Computer Vision (ICCV), 2013 IEEE International Conference on, Sydney, Australia: IEEE.

Bailly, G., Elisei, F., Raidt, S., Casari, A., & Picot, A. (2006). Embodied conversational agents : computing and rendering realistic gaze patterns. In *Pacific Rim Conference on Multimedia Processing* (Vol. LNCS 4261, pp. 9–18). Hangzhou - China.

Bailly, G., Elisei, F., & Sauze, M. (2015). Beaming the gaze of a humanoid robot. In *Human-Robot Interaction (HRI)* (pp. 47–48). Portland, OR.

Bailly, G., Raidt, S., & Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication - Special Issue on Speech and Face-to-Face Communication*, *52*(3), 598–612.

Bajcsy, R. (1988). Active Perception. *IEEE, Special Issue on Computer Vision*, *76*(8), 996–1005.

Barisic, I., Timmermans, B., Pfeiffer, U., Bente, G., Vogeley, K., & Schilbach, L. (2013). In it together: using dual eyetracking to investigate real-time social interactions. Presented at the Proceedings from SIGCHI Conference on Human Factors in Computing Systems, Paris.

Baron-Cohen, S., Jollife, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813–822.

Bengio, Y., & Frasconi, P. (1996). Input-output HMMs for sequence processing. *IEEE Transactions on Neural Networks*, *7*(5), 1231–1249. https://doi.org/10.1109/72.536317

Benoît, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, *18*, 381–392.

Bindemann, M., Burton, A. M., Hooge, I. C., Jenkins, R., & de Haan, E. F. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, *12*(6), 1048–1053. https://doi.org/10.3758/BF03206442

Boker, S. M., Cohn, J. F., Theobald, B.-J., Matthews, I., Brick, T. R., & Spies, J. R. (2009). Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars. *Philosophical Transactions of the Royal Society - Biological Sciences*, *364*(1535), 3485–3495.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study. *Image Processing, IEEE Transactions on*, *22*(1), 55–69. https://doi.org/10.1109/TIP.2012.2210727

Brône, G., & Oben, B. (2015). InSight Interaction: a multimodal and multifocal dialogue corpus. *Language Resources and Evaluation*, *49*(1), 195–214. https://doi.org/10.1007/s10579-014-9283-2

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1–13.

Carletta, J., Hill, R. L., Nicol, C., Taylor, T., de Ruiter, J. P., & Bard, E. G. (2010). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods*, *42*(1), 254–265. https://doi.org/10.3758/BRM.42.1.254

Clark, H. H. (2003). Pointing and placing. *Pointing: Where Language, Culture, and Cognition Meet*, 243–268.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*(4), 309–347. https://doi.org/10.1007/BF00994110

Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, *14*(8), 5.

Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, *5*(4), 2.

Cuijpers, R. H., & van der Pol, D. (2013). Region of eye contact of humanoid Nao robot is similar to that of a human. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social Robotics* (Vol. 8239, pp. 280–289). Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-02675-6_28

Cummins, F. (2012). Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes*, *27*(10), 1525–1549.

Dale, R., Fusaroli, R., Duran, N., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation*, *59*, 43–95.

de Kok, I. (2013). *Listening heads* (PhD Thesis). University of Twente, Enschede, The Netherlands.

Delaunay, F., Greeff, J., & Belpaeme, T. (2010). A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 39–44). Osaka, Japan.

Donat, R., Bouillaut, L., Aknin, P., & Leray, P. (2008). Reliability analysis using graphical duration models (pp. 795–800). Presented at the Availability, Reliability and Security, 2008. ARES 08. Third International Conference on, IEEE.

Duffner, S., & Garcia, C. (2015). Visual Focus of Attention estimation with unsupervised incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, to appear.

Elisei, F., Bailly, G., & Casari, A. (2007). Towards eyegaze-aware analysis and synthesis of audiovisual speech. In *Auditory-visual Speech Processing* (pp. 120–125). Hilvarenbeek, The Netherlands.

Ferreira, J. F., Lobo, J., Bessiere, P., Castelo-Branco, M., & Dias, J. (2013). A Bayesian framework for active artificial perception. *IEEE Transactions on Cybernetics*, *43*(2), 699–711. https://doi.org/10.1109/tsmcb.2012.2214477

Foerster, F., Bailly, G., & Elisei, F. (2015). Impact of iris size and eyelids coupling on the estimation of the gaze direction of a robotic talking head by human viewers. In *Humanoids*. Seoul, Korea.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Fuller, J. H. (1992). Head movement propensity. *Experimental Brain Research*, *92*(1), 152–164.

Funes Mora, K. A., & Odobez, J.-M. (2014). Geometric generative gaze estimation (G3E) for remote RGB-D cameras (pp. 1773–1780). Presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH: IEEE.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145–171.

Garau, M., Slater, M., Bee, S., & Sasse, M. A. (2001). The impact of eye gaze on communication using humanoid avatars. In *SIGCHI conference on Human factors in computing systems* (pp. 309–316). Seattle, WA.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(10), 1915–1926.

Gregory, R. (1997). *Eye and Brain: The Psychology of Seeing*. Princeton, NJ: Princeton University Press.

Gu, E., & Badler, N. I. (2006). Visual attention and eye gaze during multiparty conversations with distractions (pp. 193–204). Presented at the Intelligent Virtual Agents, Springer.

Hanes, D. A., & McCollum, G. (2006). Variables contributing to the coordination of rapid eye/head gaze shifts. *Biological Cybernetics*, *94*, 300–324.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856. https://doi.org/10.3758/PBR.16.5.850

Hietanen, J. K. (1999). Does your gaze direction and head orientation shift my visual attention? *Neuroreport*, *10*(16), 3443–3447.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, C.-M., & Mutlu, B. (2014). Learning-based Modeling of Multimodal Behaviors for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction* (pp. 57–64). New York, NY, USA: ACM. https://doi.org/10.1145/2559636.2559668

Ishii, R., Otsuka, K., Kumano, S., & Yamato, J. (2014). Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 694–698). Florence, Italy.

Itti, L., Dhavale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *SPIE 48th Annual International Symposium on Optical Science and Technology* (Vol. 5200, pp. 64–78). Bellingham, WA.

Itti, L., Dhavale, N., & Pighin, F. (2006). Photorealistic attention-based gaze animation. In *IEEE International Conference on Multimedia and Expo* (pp. 521–524). Toronto, Canada.

Jensen, F., Lauritzen, S., & Olesen, K. (1990). Bayesian updating in recursive graphical models by local computations. *Computational Statistics Quaterly*, *4*(1), 269–282.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks (pp. 1725–1732). Presented at the Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE.

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. *Journal of Human Evolution*, *40*(5), 419–435.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. Boston, MA: MIT Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing (NIPS)*. Lake Tahoe, NV.

Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Social attention to a live person is critically different than looking at a videotaped person. *Proc. Natl. Acad. Sci. (PNAS)*, *108*, 5548–5553. https://doi.org/10.1073/pnas.1017022108

Lakin, J., Jefferis, V., Cheng, C., & Chartrand, T. (2003). The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. *Nonverbal Behavior*, *27*(3), 145–162.

Langton, S. R. H. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *Quarterly Journal of Experimental Psychology*, *53A*(3), 825–845.

Langton, S. R., Honeyman, H., & Tessler, E. (2004). The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & Psychophysics*, *66*(5), 752–771.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, *42*(3), 526–539.

Lee, S. P., Badler, J. B., & Badler, N. (2002). Eyes alive. *ACM Transaction on Graphics*, *21*(3), 637–644.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Li, J., Tian, Y., & Huang, T. (2014). Visual saliency with statistical priors. *International Journal of Computer Vision*, *107*(3), 239–253.

Liang, S., Fuhrman, S., Somogyi, R., & others. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing* (Vol. 3, pp. 18–29). Retrieved from http://politespider.com/papers/grn/REVEAL%20-%20%20a%20general%20reverse%20engineering%20algorithm%20for%20inference%20of%20genetic%20network%20architectures.pdf

Marschner, L., Pannasch, S., Schulz, J., & Graupner, S.-T. (2015). Social communication with virtual agents: The effects of body and gaze direction on attention and emotional responding in human observers. *International Journal of Psychophysiology*, *97*(2), 85–92. https://doi.org/10.1016/j.ijpsycho.2015.05.007

McNeill, D. (1992). *Hand and Mind. What Gestures Reveal about Thought*. Chicago: Chicago University Press.

Mihoub, A., Bailly, G., & Wolf, C. (2014). Modelling perception-action loops: comparing sequential models with frame-based classifiers. In *Human-Agent Interaction (HAI)* (pp. 309–314). Tsukuba, Japan.

Mihoub, A., Bailly, G., & Wolf, C. (2015). Learning multimodal behavioral models for face-to-face social interaction. *Journal on Multimodal User Interfaces (JMUI)*, *9*(3), 195–210. https://doi.org/10.1007/s12193-015-0190-7

Mihoub, A., Bailly, G., Wolf, C., & Elisei, F. (2015). Learning multimodal behavioral models for face-to-face social interaction. *Journal on Multimodal User Interfaces*, 1–16. https://doi.org/10.1007/s12193-015-0190-7

Mihoub, A., Bailly, G., Wolf, C., & Elisei, F. (2016). Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognition Letters*, *74*, 82–89. https://doi.org/Graphical models for social behavior modeling in face-to face interaction

Murphy, K. (2002). *Dynamic bayesian networks: representation, inference and learning* (PhD Thesis). UC Berkeley, Computer Science Division, Berkeley, CA.

Murphy, K. P. (2001). The Bayes Net Toolbox for MATLAB. *Computing Science and Statistics*, *33*, 2001.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *1*(2), 12.

Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2016). ModDrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *38*(8), 1692–1706.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning (pp. 689–696). Presented at the International conference on machine learning (ICML), Bellevue, WA.

Nguyen, Duc-Anh, Bailly, Gérard, & Elisei, Frédéric. (2016). Conducting neuropsychological tests with a humanoid robot: design and evaluation. In *IEEE International Conference on Cognitive Infocommunications – CogInfoCom*. Wroclaw, Poland.

Onuki, T., Ishinoda, T., Kobayashi, Y., & Kuno, Y. (2013). Designing robot eyes for gaze communication. In *IEEE Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)* (pp. 97–102). Fukuoka, Japan.

Otsuka, K. (2011). Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings. In *International Conference on Human-Computer Interaction (HCI)* (Vol. 12, pp. 171–179). Orlando, FL.

Otsuka, K., Takemae, Y., & Yamato, J. (2005). A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *International Conference on Multimodal Interfaces (ICMI)* (pp. 191–198). Seattle, WA.

Oyekoya, O., Steed, A., & Steptoe, W. (2010). Eyelid kinematics for virtual characters. *Computer Animation and Virtual Worlds*, *21*(3–4), 161–171.

Pelachaud, C., & Bilvi, M. (2003). Modelling gaze behavior for conversational agents. In *International Working Conference on Intelligent Virtual Agents* (Vol. LNAI 2792). Kloster Irsee, Germany.

Pentland, A. S. (2004). Social dynamics: Signals and behavior. Presented at the International Conference on Developmental Learning, La Jolla, CA.

Pentland, A. S. (2007). Social Signal Processing. *IEEE Signal Processing Magazine*, *24*(4), 108–111.

Picot, A., Bailly, G., Elisei, F., & Raidt, S. (2007). Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. In *International Conference on Intelligent Virtual Agents (IVA)* (pp. 272–282). Paris, France.

Raidt, S., Bailly, G., & Elisei, F. (2007). Mutual gaze during face-to-face interaction. In *Auditory-visual Speech Processing*. Hilvarenbeek, The Netherlands.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination common ground and the coupling of eye movements during dialogue. *Psychological Science*, *18*(5), 407–413.

Richardson, D. C., Dale, R., & Shockley, K. (2008). Synchrony and swing in conversation: coordination, temporal dynamics, and communication. In I. Wachsmuth, M. Lenzen, & G. Knoblich (Eds.), *Embodied Communication* (pp. 75–93). Oxford, UK: Oxford University Press.

Risko, E. F., Laidlaw, K. E. W., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience*, *6*, 143. https://doi.org/10.3389/fnhum.2012.00143

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the Fourth Wall of Cognitive Science Real-World Social Attention and the Dual Function of Gaze. *Current Directions in Psychological Science*, *25*(1), 70–74.

Ruhland, K., Andrist, S., Badler, J., Peters, C., Badler, N., Gleicher, M., … Mcdonnell, R. (2014). Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems (pp. 69–91). Presented at the Eurographics State-of-the-Art Report.

Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., & Mao, M. (2014). Sequence discriminative distributed training of long short-term memory recurrent neural networks. *Entropy*, *15*(16), 17–18.

Schauerte, B., & Stiefelhagen, R. (2014). "Look at this!" learning to guide visual saliency in human-robot interaction (pp. 995–1002). Presented at the Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on, IEEE.

Schmidt, R., Morr, S., Fitzpatrick, P., & Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior*, *36*(4), 263–279.

Senju, A., & Hasegawa, T. (2005). Direct gaze captures visuospatial attention. *Vision Cognition*, *12*, 127– 144.

Sugano, Y., Matsushita, Y., & Sato, Y. (2013). Appearance-based gaze estimation using visual saliency. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(2), 329–341.

Sun, Y. (2003). *Hierarchical object-based visual attention for machine vision* (Thesis). Institute of Perception, Action and Behaviour, University of Edinburgh, Edinburgh, UK.

Teufel, C., Alexis, D. M., Clayton, N. S., & Davis, G. (2010). Mental-state attribution drives rapid, reflexive gaze following. *Attention, Perception, & Psychophysics*, *72*(3), 695–705.

Tomasello, M. (2008). *Origins of Human Communication*. Boston, MA: MIT Press.

Tomasello, M. (2009). *Why We Cooperate*. Cambridge, MA: MIT Press.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, *52*, 314–320.

Trabelsi, G., Leray, P., Ben Ayed, M., & Alimi, A. M. (2013). Benchmarking dynamic Bayesian network structure learning algorithms (pp. 1–6). Presented at the Modeling, Simulation and Applied Optimization (ICMSAO), 2013 5th International Conference on, IEEE.

Trutoiu, L. C., Carter, E. J., Matthews, I., & Hodgins, J. K. (2011). Modeling and animating eye blinks. *ACM Transactions on Applied Perception (TAP)*, *8*(3), 1–17. https://doi.org/10.1145/2010325.2010327

Valenti, R., & Gevers, T. (2012). Accurate eye center location through invariant isocentric patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(9), 1785–1798.

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2009). Poke and pop: Tactile–visual synchrony increases visual saliency. *Neuroscience Letters*, *450*(1), 60–64.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*, 926–940.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Conference on Human Factors in Computing Systems* (pp. 301–308). Seattle, WA: ACM Press New York, NY, USA.

Vinayagamoorthy, V., Garau, M., Steed, A., & Slater, M. (2004). An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. *The Computer Graphics Forum*, *23*(1), 1–11.

Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *13*(3), 1–14.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In L. A. Riggs (Ed.), *Eye Movements and Vision* (Vol. VII, pp. 171–196). New York: Plenum Press.