Direct Gaze Triggers Higher Frequency of Gaze Change: An Automatic Analysis of Dyads in Unstructured Conversation

Georgiana Cristina Dobre
c.dobre@gold.ac.uk
Goldsmiths, University of London

Marco Gillies
m.gillies@gold.ac.uk
Goldsmiths, University of London

Patrick Falk
patrick.falk.17@ucl.ac.uk
University College London

Jamie A. Ward
J.Ward@gold.ac.uk
Goldsmiths, University of London

Antonia F. de C. Hamilton
a.hamilton@ucl.ac.uk
University College London

Xueni Pan
x.pan@gold.ac.uk
Goldsmiths, University of London

ABSTRACT

Nonverbal cues have multiple roles in social encounters, with gaze behaviour facilitating interactions and conversational flow. In this work, we explore the conversation dynamics in dyadic settings in a free-flow discussion. Using automatic analysis (rather than manual labelling), we investigate how the gaze behaviour of one person is related to how much the other person changes their gaze (frequency in gaze change) and what their gaze target is (direct or avert gaze). Our results show that when one person is looked at they change their gaze direction with a higher frequency compared to when they are not looked at. They also tend to maintain a direct gaze to the other person when they are not looked at.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → Intelligent agents.

KEYWORDS

multimodal interaction, eye gaze behaviour, social interaction

ACM Reference Format:
Georgiana Cristina Dobre, Marco Gillies, Patrick Falk, Jamie A. Ward, Antonia F. de C. Hamilton, and Xueni Pan. 2021. Direct Gaze Triggers Higher Frequency of Gaze Change: An Automatic Analysis of Dyads in Unstructured Conversation. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462244.3479962

1 INTRODUCTION

When we interact with other people we use both verbal and nonverbal signals, not only to make ourselves understood but also to check if the message is received as we intended. Gaze behaviour is one the nonverbal cues that facilitate interaction and the conversation flow, and in a live social interaction it can be very different from when watching a video [5]. Gaze is tightly coordinated with other nonverbal behaviours such as speech [15, 16]. Gaze cues have captured scientific interest since the 1960s. Researchers would either observe and label on a social interaction live (e.g. [17]) or they would video record the interaction for later analysis (e.g. [7]). Some of these studies look at a structured interaction between two participants [13] when they have predefined actions such as going through a set of predefined questions. A benefit of these types of tasks is the better control over the conversational roles (speaker/listener). Although they do have important contributions to the field, the results from structured tasks are not always applicable in free flow conversations, with clear limitations when used as a building block for nonverbal behaviour models used in autonomous virtual agents.

In unstructured tasks, participants are usually instructed to speak about a certain subject (free-flow conversations) or to speak with a confederate about a certain topic [14]. These tasks are closer to how people interact everyday and can capture different conversational dynamics between participants. Insights from studies with unstructured tasks could help create a nonverbal behaviour model for autonomous agents. Gaze behaviour, for example, is one of the social behaviours that has been well studied [3, 15, 17].

One major challenge in the analysis of unstructured conversation data is the annotation or labelling of the specific events within the recording, which are typically more time consuming than the structured ones. Although interesting results are emerging from these, it would be difficult to scale the manual annotations to large datasets. Also, it brings challenges when working with interactive autonomous agents, as the same manual data labelling needs to happen in real time, making it not truly autonomous.

We aim to explore conversational dynamics between two people into a free-flow discussion that could be later integrated in a nonverbal model for an autonomous agent for real-time social interactions. We use automatic data annotation methods and considered the gaze targets of either looking at the other person's face (direct gaze, DG) or not looking at the other person gaze (avert gaze, AG). We consider the following hypotheses:

H1: Listeners performs more DG than speakers. This is in line with the literature and acts as a validation of our method.

Motivated by the approach/avoidance conflict in social interactions (see Section 2), we are interested in how receiving DG could influence someone's gaze behaviour - **H2a: When someone is being looked at (receiving DG), they would switch back and forth between performing DG and AG with a higher frequency compared to when they are not being looked at.** In line with this hypothesis, when someone is not being looked at, they change less frequently between DG and AG. We hypothesise that in this situation, this person will direct their gaze to the

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10.
https://doi.org/10.1145/3462244.3479962

other conversation partner - **H2b: When someone is not being looked at (receiving AG), they would look more at the other person's face (DG) than somewhere else (AG).**

2 LITERATURE REVIEW

Gaze is an important part of social interaction with functions such as regulating the conversational turns [16], providing extra information in ambiguous situations [18], giving insights about how people think or feel [4], getting the other person's attention or approval [8], signalling attractiveness, dominance and threat[2, 9, 10].

During a conversation, the amount of time one person looks at the other varies considerably [11, 16]. On average, listeners tend to give more DG to speakers, with these DGs being broken only by very short AG periods [16]. However, the DG and AG of the speaker tend to be, on average, more equal in length. In other words, the speaker's *AG* is considerably longer than that of the listener's. This led to **H1**, which also serves as a validation of our data and method.

Another factor in gaze is an approach/avoidance conflict. Argyle and Dean [2] proposed that eye contact or mutual gaze (when both participants are gazing at the other) is, on one hand, actively sought in conversations for increasingly closeness and self validation. On the other hand, there is a tendency to avoid excessive mutual gaze, as it can be overly intimate and arousing. This leads to a conflict that is normally resolved by reaching an equilibrium level of mutual gaze. This was the motivation for our **H2**, when a participant is being looked at by their conversational partner (i.e. when their own gaze towards the partner would result in mutual gaze), they will be more actively managing the level of mutual gaze through their own gaze behaviour and will therefore switch between directed gaze (towards the partner) and averted (away) more often (**H2a**). They will also look less overall (lower overall mutual gaze), **H2b**.

Hence, the interaction dynamics between people are greatly influenced by the participants in the interaction. Looking at only one's behaviour gives only partial insights into the nonverbal behaviour in interactions. Works such as [1, 6, 12] take into account data from all participants in that interaction to detect or generate different aspects of a social interaction. Although it helps advance the field, a nonverbal behaviour model based solely on one's data leads to behaviour that is neither flexible nor contextual.

Understanding the gaze behaviour dynamics between two people can inform constructions of gaze and non-verbal behaviour models for conversational agents. This is particularly relevant now as some Virtual Reality (VR) headsets come with gaze tracking capacity (e.g. VIVE Pro Eye), enabling a whole range of applications in gaming and social skills training, where gaze behaviours appear to be similar to when taking place in real world interactions [20]. Importantly, the agent's nonverbal behaviour has implications in maintaining the user's plausibility illusion [21]. For instance, poor coordination between gestures and speech can make the agent be seen as nervous or not eloquent, while poorly timed gaze behaviour can disturb the conversation's smooth flow [19].

In this work, we look at the gaze and speech of participants during unstructured conversational tasks. We aim to validate some of the previously reported gaze behaviours, but also to look at how people's gaze changes when they are looked at or not by the other person. This initial work aims to strengthen our understating of gaze dynamics. We plan to further include these findings in building nonverbal behaviour models for autonomous agents in virtual environments however, this is out of the scope of this paper.

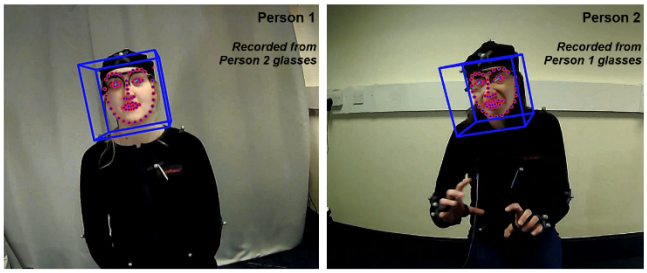

Figure 1: Dyad recording setup. Person 1 and 2 are seated in front of each other. The footage is recording from a camera on each person's PupilLab Glasses. The blue box and the facial landmarks are exported from OpenFace software and added to the original video.

3 MULTIMODAL DATASET WITH DYADS

The multimodal data was recorded as part of this research project. The data was recorded in a room with two stools, so that each participants pair was facing each other at a distance of appoximately 1.5 m (Fig. 1). A projector screen to the participants' side showed instructions, with pre-recorded audio cues played from a speaker. The researcher was separated from the participants with a curtain. They remained in the same room, but could not be seen nor did they interact with participants during the experiment. A video camera recorded the whole session. Each participant wore a lapel microphone. Their voices were registered on an audio file (left and right channels). Each of them also wore the PupilLab glasses (https://pupil-labs.com) that recorded their eye, gaze data, and a video stream of that person's view. Upper body motion capture was also recorded but excluded from this work.

There were 62 participants recruited from a local mailing list. They were paired up as 31 dyads given their availability. Participants acclimatised to the experimental set up through a PupilLab glasses calibration session and a task of watching a short cartoon. We did not include these parts in the analysis. Next, they were engaged in three types of tasks: discussion, picture description and meal planning (recipe). There were five sessions in the following order: discussion 1, picture description 1, recipe, picture description 2 and discussion 2. The activity took on average one hour to complete. Here only discussion 1 & 2 and recipe were included as they both are unstructured tasks where participants were not told when to speak or listen. They were left to talk freely. During the discussion task, the participants talked about a short cartoon video that they previously watched. This task lasted two minutes, and took place on two occasions for each pair, resulting in a total of four minutes of dialogue for each dyad. In the recipe task, the participants spoke freely in order to plan a meal that uses ingredients both dislike. This task took approximately five minutes for each dyad.

The gaze target data was exported from the PupilLab software, and it can have low confidence when the eyes are closed (blinks) or when the target gaze can not be detected due to eye shape, the

participant's makeup, or if the PupilLab glasses were not well fitted. Out of the 31 dyads and 93 task datasets (three tasks per dyad), we removed 37 datasets as the overall gaze target confidence was less than 65%. We consider 56 tasks from 23 dyads. There were 18 same-sex dyads and 5 mixed-sex dyads (41 female and 5 male).

Out of the total 56 datasets there were: 21 for discussion 1 ($D1$), 18 for discussion 2 ($D2$), and 17 for the recipe (R) task. A total of 164 minutes were recorded, with 78 from $D1$ and $D2$ combined, and 86 from R. Out of these 23 dyads, 11 had the speech recorded only from one lapel microphone due to a technical error. They were excluded from the speech-related results (i.e., H1). From those with full audio available, we considered all three tasks from 8 of the dyads, D1 from one dyad, R from one dyad, $D1$ and R from one dyad and finally both discussions ($D1$, $D2$) for one dyad. There are in total 10 recordings for $D1$ and R, and 9 for $D2$. This brings a total of 29 tasks and 88 minutes of data (50 from R and 38 from $D1$&2).

3.1 Data post-processing

We post-processed the data from the PupilLabs glasses and the audio files. Here, the term *audio* describes the sound that comes from a participant - it includes the speech but also laughter or backchannels.

From the PupilLabs software we exported the gaze targets and the person's view in video format. We used the video for getting the face location of the other person (the person they were looking at). To generate the face position data, we used OpenFace software [22]. From OpenFace we calculated a square to fit the participant's face. However, as the returned values represented the face contour (excluding the forehead), we enlarged it with 10%, to capture the edge of the face. With the gaze target data for each participant and the face coordinates of the other person, we were able to detect the behaviour of looking or not looking at the other person's face (DG/AG). The data was recorded at a 30 frames per second frequency. However, to limited the noise in the data, we scaled it down to 6 frames per second, combining each 5 frames.

Each channel from the audio files was post-process by applying the Google's WebRTC Voice Activity Detector (https://webrtc.org/) via the python interface *pyvad* version 0.1.3 (https://pypi.org/project/pyvad/). The detector output was binary voiced or unvoiced data (value 1 or 0) per sample for each audio channel with a sample rate of 22050 Hz. Each channel represents one person from a given dyad. As the gaze data is represented with a 6 frames per second (166ms frequency), we used the same frequency for the audio data. We summed the values for each 166ms window: if the window was unvoiced, the resulted value was 0, whereas if the window was fully voiced, the resulted value for that window was 3675 (dividing the sample rate by six: 22050/6). Hence, the outcome voice detection file had a frequency of 166ms, and each of these datapoints had a value between 0 and 3675.

Participants were in close proximity, hence the microphone from one person was recording some of the activity from the other person. We considered this when post-processing the voice detection files. Given person A with their microphone mA and person B with their microphone mB, in the ideal scenario, mA would record only A's voice and mB only B's voice. In reality, as A starts speaking (while B remains quiet), mA captures the A's speech, however, mB also captures some of this speech. In this situation, in the data from the voice detection file, the values from mA are higher than the values from mB (the voice detection files contains values between 0 and 3675, see above). Because of this, we compared the values from mA and mB by each dataframe and marked as 'speaking' the person whose voice detection value is higher. If the value is equal, then both of them are marked as speaking. After this second data post-processing, the voice detection file has binary values: 0 for listening and 1 for speaking.

This post-processing might also introduce very short speaking duration sections (less than one second) that are not from the person wearing the microphone but rather captured from the other person's speech. To tackle this issue, we filtered any sections of speech shorter than one second. This also removed some of the backchannels or laughter that appear in the audio as the voice activity detector does not account for them.

4 DATA ANALYSIS AND RESULTS

4.1 Gaze behaviour during conversational roles

Firstly, we analysed the data to validate the most common gaze behaviour recorded in previous literature [16]. In line with our hypothesis, the speaker has a higher amount of AG behaviour (looking away their partner's face) while the listener has a higher DG (looks at their partner's face). We split the data into two parts based on the conversation role label (speaking or listening). Then we calculated the percent of which participant is looking at their partner or is averting their gaze, for both parts. On average, the listener looked more at their partner (69%) while the speaker had a DG of (61%). The percentages differed based on the task. In $D1$ and $D2$, the listener had 71% DG while the speaker only 60%. The difference is smaller in the Meal Planning (Recipe) task with 65% of direct gaze while listening and (61%) while speaking.

A repeated measure two-way ANOVA with the conversational *role* (speaker and listener) and the *task* (D1, D2 and R) as factors was performed with SPSS software $v27$. No *interaction* effect was found ($F(2,7) = 4.148, p = 0.065, \eta^2 = 0.542$), and no effect was found for *Task* ($F(2,7) = 0.238, p = .794, \eta^2 = 0.64$). However, there was an effect on ***role*** ($F(1,8) = 71.024, p < .001, \eta^2 = 0.899$). As expected, speaker performed significantly less DG, confirming **H1**. Figure 2a shows the values for each task and by role.

4.2 The effect of being looked at on own gaze

We were interested in the hypothesis that when someone is being looked at, they change their gaze differently compared to when they were not ($H2a$). Here we analyse how much they were changing their gaze behaviour per second. The gaze behaviour can be either *DG* (looking at the other person's face), or *AG* (looking away from the other person's face). Here we used all 56 tasks. We first separated the data in two datasets: when the participant is looked at (dataset L) and where they are not (dataset nL). We did this for each participant in the dyad. Next, we computed the sum of all the changes in gaze behaviour of the person being looked at (from dataset L) or not being looked at (from dataset nL). We then calculated how many seconds are in L and in nL. With these values, we calculated the frequency of gaze change per second by dividing the total seconds from the gaze change value (see Equations 1 & 2).

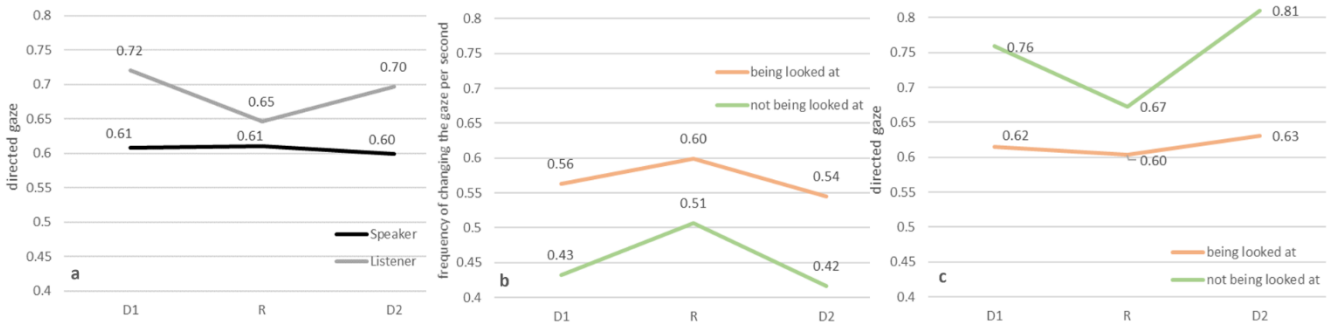

Figure 2: *a*: DG percent while being a speaker or a listener during each task. *b*: Gaze change frequency while being or while not being looked at during each task. *c*: DG percent while being or while not being looked at by task. The tasks are in chronological order (D stands for discussion (1 and 2) and R for recipe)

$$L_{gaze\_change\_fq} = \frac{\sum L_{gaze\_change}}{\sum L_{duration}} \quad (1)$$

$$nL_{gaze\_change\_fq} = \frac{\sum nL_{gaze\_change}}{\sum nL_{duration}} \quad (2)$$

Figure 2b shows the results by task, with the $nL_{gaze\_change\_fq}$ having lower value compared to $L_{gaze\_change\_fq}$. A repeated measure 3×2 two-way ANOVA (task and L: looked at/nL: not looked at). No *interaction* effect was found ($F(2, 15) = 1.001, p = 0.391, \eta^2 = 0.118$), so as for *Task* ($F(2, 15) = 1.590, p = .236, \eta^2 = 0.175$). However, there was an effect on ***Looked at*** ($F(1, 16) = 21.681, p < .001, \eta^2 = 0.575$). This confirms **H2a** that when being looked at, participants change their gaze pattern significantly more frequently compares to when they are not.

Given that participants were making fewer changes in gaze while they were not being looked at, we analysed what was their behaviour during those periods, which led us to **H2b**. Hence we calculated the percentage of DG of each person while being or not being looked at. We considered all 56 task datasets for this analysis.

As before, we first separated the data into two datasets: the part when the participant is looked at (dataset L) and the part where they are not looked at (dataset nL). We considered each participant in the dyad separately. Next, we summed the DG by the person being looked at (from dataset L) or not being looked at (from dataset nL). The percent is calculated by dividing the amount of direct gaze by the dataset size, either dataset L or dataset nL (see Equation 3 & 4).

$$L_{direct\_gaze\_percent} = \frac{\sum L_{direct\_gaze}}{L_{size}} \quad (3)$$

$$nL_{direct\_gaze\_percent} = \frac{\sum nL_{direct\_gaze}}{nL_{size}} \quad (4)$$

A repeated measure 3×2 two-way ANOVA was performed with two factors (task and gaze behaviour of their partner - L: looked at/nL: not looked at). No *interaction* effect was found ($F(2, 15) = 2.995, p = 0.080, \eta^2 = 0.285$), and no effect was found for *Task* ($F(2, 15) = 2.026, p = .166, \eta^2 = 0.213$). However, there was an effect on ***Looked at*** ($F(1, 16) = 28.091, p < .001, \eta^2 = 637$). This confirms **H2b** that when not being looked at, participants tended to perform more DG than when they were (see figure 2c).

5 DISCUSSION

A limitation of this study is our dataset containing only dyads between strangers. In future work we plan to look at the effect of familiarity of the conversation partners on the gaze behaviours.

As with manual annotations, it is possible that our automatic data analysis process brings some degree of error. Hence our H1 serves as a validation and was supported. We manually annotate a limited part of our data and the results were largely in agreement with our automatic analysis. We also used statistical tests to compensate for noise. Further, with our automatic annotation method, small errors can be compensated by the use of a large number of frames of data (higher frequency and longer time), also making it possible to scale to much larger datasets which would lead to better generalisation.

For all three hypotheses, the Recipe task has a smaller effect. These differences in results could be explained by the task's nature. In R, the participants were asked to come up with a meal plan with foods they both dislike. This led to silent periods where they were thinking about the food they do not like, but also to more speech overlap and sections of laughter.

6 CONCLUSION

In this work, we analyse gaze and speech behaviour to gain insights into conversational dynamics between dyads, during an unstrctured conversation. We used an automated method to annotate speech and gaze data from 56 unstructured tasks, from 46 participants. We found that people tended to have a higher frequency of gaze change (from averting to directing and vice versa) when they were being looked at compared to when they were not. During the times where the participants were being looked at, they were also directing their gaze to their partners more compared to when they were not. Alongside proportions of gaze, we also looked at how it changes when being looked at (hence the use of gaze change frequency as a dependent variable). It is a direct contribution to understanding human interaction towards developing a diagnostic tool for neurological disorders such as autism and depression. Also, the work contributes to a more realistic gaze model for VR applications such as soft skill training, language learning, and entertainment, by modelling more complex dynamics for NPCs.

REFERENCES

[1] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations. In *2019 International Conference on Multimodal Interaction*. 74–84.

[2] Michael Argyle and Janet Dean. 1965. Eye-contact, distance and affiliation. *Sociometry* (1965), 289–304.

[3] Michael Argyle and Roger Ingham. 1972. Gaze, mutual gaze, and proximity. *Semiotica* 6, 1 (1972), 32–49.

[4] Simon Baron-Cohen, Sally Wheelwright, Jolliffe, and Therese. 1997. Is there a" language of the eyes"? Evidence from normal adults, and adults with autism or Asperger syndrome. *Visual cognition* 4, 3 (1997), 311–331.

[5] Roser Cañigueral, Jamie A Ward, and Antonia F de C Hamilton. 2021. Effects of being watched on eye gaze and facial displays of typical and autistic individuals during conversation. *Autism* 25, 1 (2021), 210–226.

[6] Soumia Dermouche and Catherine Pelachaud. 2019. Generative Model of Agent's Behaviors in Human-Agent Interaction. In *2019 International Conference on Multimodal Interaction*. 375–384.

[7] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* (1972). https://doi.org/10.1037/h0033031

[8] Jay S Efran and Andrew Broughton. 1966. Effect of expectancies for social approval on visual behavior. *Journal of Personality and Social Psychology* 4, 1 (1966), 103.

[9] Steve L Ellyson, John F Dovidio, and BJ Fehr. 1981. Visual behavior and dominance in women and men. In *Gender and nonverbal behavior*. Springer, 63–79.

[10] Nathan J Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews* 24, 6 (2000), 581–604.

[11] Ralph V Exline. 1963. Explorations in the process of person perception: visual interaction in relation to competition, sex, and need for affiliation. *Journal of personality* (1963).

[12] W. Feng, A. Kannan, G. Gkioxari, and C. L. Zitnick. 2017. Learn2Smile: Learning non-verbal interaction through observation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4131–4138.

[13] Megan Freeth, Tom Foulsham, and Alan Kingstone. 2013. What affects social attention? Social presence, eye contact and autistic traits. *PloS one* 8, 1 (2013), e53286.

[14] Roy S Hessels, Gijs A Holleman, Alan Kingstone, Ignace TC Hooge, and Chantal Kemner. 2019. Gaze allocation in face-to-face communication is affected primarily by task structure and social context, not stimulus-driven factors. *Cognition* 184 (2019), 28–43.

[15] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905.

[16] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63.

[17] A. Kendon and M. Cook. 1969. The consistency of gaze patterns in social interaction. *British journal of psychology (London, England : 1953)* (1969). https://doi.org/10.1111/j.2044-8295.1969.tb01222.x

[18] Ross G Macdonald and Benjamin W Tatler. 2013. Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of vision* 13, 4 (2013), 6–6.

[19] Michael Neff and Catherine Pelachaud. 2017. Animation of Natural Virtual Characters. *IEEE Computer Graphics and Applications* 37, 4 (2017), 14–16.

[20] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 1 (2019), 1–40.

[21] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557.

[22] Amir Zadeh, Yao Chong Lim, Tadas Baltrušaitis, and Louis Philippe Morency. 2018. Convolutional experts constrained local model for 3D facial landmark detection. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*. https://doi.org/10.1109/ICCVW.2017.296