Depth from motion parallax: deictic consistency, eye contact and a serious problem with Zoom

Nikolaus F. Troje
Centre for Vision Research and Department of Biology
York University, Toronto, Canada

Abstract: 119 words
Main text: 5728 words
Two figures
62 references

Abstract
The dynamics of head and eye gaze between two or more individuals displayed during verbal and non-verbal face-to-face communication contains a wealth of information and is used for both volitionary and unconscious signalling. Current video communication systems convey visual signals about gaze behaviour and other directional cues, but the information they carry is often spurious and potentially misleading.

I discuss the consequences of this situation, identify the source of the problem as a more general lack of deictic consistency, and demonstrate that using display technologies that simulate motion parallax are both necessary and sufficient to alleviate it. I then devise an avatar-based remote communication solution that achieves deictic consistency and provides natural, dynamic eye contact for computer-mediated audio-visual communication.
The dynamics of mutual gaze

Zoom, Skype and similar remote video communication platforms play a significant role in modern society and have become an essential part of our communication behaviour. Streaming video adds a wealth of visual information to the auditory-only communication that we had mostly relied on for remote communication until about 20 years ago. Nevertheless, remote video communication does not achieve the same efficiency as normal face-to-face communication. It is experienced as more exhausting (Bailenson, 2021) and has been shown to be less efficient in the context of, for instance, teaching and learning (Serhan, 2020), employment interviews (Basch, Melchers, Kurz, Krieger, & Miller, 2021; Blacksmith, Willford, & Behrend, 2016), and therapeutic consultation (Hammersley et al., 2019). One central reason for this, as I discuss in this paper, is a lack of faithful directionality between the users which affects deictic (that is, oriented, directed) behaviour in general, and particularly impacts true, dynamic eye contact (Nguyen & Canny, 2007).

The dynamics of head and eye gaze between two or more individuals displayed during verbal and non-verbal face-to-face communication contains a wealth of information and is used for both volitionary and involuntary signalling. Current video communication systems convey visual signals about gaze behaviour and other directional cues, but the information they carry is often spurious and potentially misleading. What are the consequences of this situation and what can be done to mitigate the problem and improve remote video communication?

In real-world face-to-face communication, how my partner’s gestures, head orientation and gaze develop over time, and relative to my own location and gaze behaviour contains important information that helps me to monitor, on a fine time scale, how my communication partner divides her attention between me and other people or things, whether she understood what I just said, whether she trusts my words, and to which degree she is willing and able to engage in our conversation. It informs me whether she understands my facial expression and attends to my actions and words, but also tells me, on a fine time scale, when I am momentarily unobserved. A wealth of research confirms that the dynamics of head orientation and gaze contain critical information that determines in many ways a perceiver’s behaviour and the course a communication takes. Social gaze is a disambiguating cue in referential communication (Hanna & Brennan, 2007), informs social bonding, trust building and rapport (Wolf, Launay, & Dunbar, 2016), moderates the perception of facial expressions of emotion (Adams Jr & Kleck, 2005), provides signals to negotiate turn taking (Duncan, 1972), affects the efficiency of both verbal and non-verbal communication between social partners in many other ways (e.g. Goldman, 1980), and elicits direct and immediate, inescapable affective responses (Hietanen, 2018). Kleinke (1986) in a very comprehensive review provides many more examples that illustrate the richness of mutual gaze behaviour and the various functions it serves in communication and social interaction. Bohannon et al. (2013) contributes another review which emphasises on eye contact behaviour during video-mediated communication. Hessels
(2020), in a more recent review, focuses on how gaze to faces supports face-to-face interaction.

Not surprisingly, the anatomy of our eyes is well-adapted to send gaze-related signals. In humans, the sclera of the eye lost its pigmentation and appears white, thus contrasting with the darker iris. Humans also have wider eyes compared to other primates exposing more sclera on either side of the iris (Kobayashi & Kohshima, 1997, 2001). These traits are partially present in other great apes, too, but to a much lesser extent than in humans (Caspar, Biggemann, Geissmann, & Begall, 2021; Emery, 2000). No other animal is known to use gaze as a communicative cue – maybe with exception of dogs who most likely acquired it only during domestication by humans.

The accuracy in the perception of gaze direction has been measured to be slightly less than 1 degree at normal conversation distance (1.2 m), which means that the sensitivity to iris displacements (about 0.2 mm in this case) is only limited by the spatial acuity of the observer’s eye (Chen, 2002; Cline, 1967; Gamer & Hecht, 2007). Mareshal and colleagues (2013) have also demonstrated that observers are biased to register another person’s gaze as eye contact even if it is slightly off. Like other perceptual biases, this one is probably also reflecting the statistics of the real world. The prior probability that another person is looking just a few degrees away from our eyes (rather than either looking right into our face or looking definitely somewhere else) is so low that it seems to be a safe bet for our perceptual decision-making system to assume that the other person is in fact looking at us – even if the signal itself might be noisy and gaze direction potentially within Gamer & Hecht’s (2007) “cone of gaze”.

Perceiving eye contact triggers a fast physiological response that is controlled by the autonomic nervous system (Mazur et al., 1980) and can be measured in terms of pupil diameter (Honma, Tanaka, Osada, & Kuriyama, 2012), skin conductance (Hietanen, Leppänen, Peltola, Linna-Aho, & Ruuhiala, 2008), cardiac reactivity (Quigley & Stifter, 2006) and EEG activity (Hirsch, Zhang, Noah, & Ono, 2017). Senju and Johnson (2009) who review the literature on this “eye contact effect” suggest a model in which the initial response to detected eye contact is generated in a subcortical network that has long been known to play a role in early visual development of face recognition competence where it guides the preference for face-like visual stimuli already in neonate infants (Johnson & Morton, 1991). Senju and Johnson (2009) suggest that in the mature brain the same subcortical network still responds to eye contact events and uses its fast-track responses to modulate processing in several parts of a larger cortical network of the so-called social brain.

At this level, mutual eye contact might directly synchronize information processing in the brains of communicating partners as shown by Hirsch et al. (2017). It is thus a candidate for a physiological interpretation of the Interactive Brain Hypothesis brought forward by Di Paolo and colleagues (see e.g. Di Paolo & De Jaegher, 2012), by which mind-reading through interactive synchrony plays a key role for the understanding of interpersonal interaction and communication.
Responses to receiving gaze are gated by one’s own behaviour as have been shown elegantly by Jarick and Bencic (2019). The authors measured autonomous sympathetic arousal in terms of skin conductance responses in communicating dyads. Participants exhibited much higher degrees of arousal while making direct eye contact than when they were either only sending or only receiving social gaze. Other authors have also pointed out that eye contact behaviour can only be understood as emerging from the closed loop formed by the dyad (Gallagher, 2014; Laidlaw, Foulsham, Kuhn, & Kingstone, 2011; Myllyneva & Hietanen, 2016).

In summary, both the elaborate adaptations that humans have evolved to develop mutual gaze into a rich social information channel, and the variety of functions that it plays in communication emphasise the importance and complexity that eye contact plays during visual interactions as an emergent behaviour between two or more communication partners.

**Deictic consistency and the sense of place**

Interpreting gaze, head orientation, pointing gestures, and any other directed behaviour of another person in egocentric coordinates is based on the assumption that the communication partners are located in a common, normally behaving space. The ancient Greek mathematician Euclid elegantly operationalized the properties of “normal” space as experienced by the human agents in terms of five axioms. The first and central one states that for any two different points in space, there exists a unique line containing these two points. Here, I will refer to that characteristic as “deictic consistency” or sometimes just as “directionality” (Troje, 2023) (see also Nguyen and Canny (2005) who used the term “spatial faithfulness” for a very similar concept). Deictic consistency means that two individuals X and Y generally agree on their orientations and locations relative to each other (Fig. 1a). That agreement, in turn, is a crucial prerequisite to also achieve consensus about egocentric locations of other objects and it is needed to transform the two individuals’ egocentric locations into a coherent allocentric space in which both exhibit co-presence. The line of sight, that is, the unique Euclidean line shared by both individuals, can now serve as an unambiguous reference to represent mutual location and orientation in space.

---

1 Deixis is Greek for ‘pointing’. The adjective ‘deictic’ is frequently used in linguistics, but also in semiotics, for words or actions that reference from the speaker’s (or actor’s) position to another location in space or time.
Deictic consistency holds wherever there is space, that is, everywhere (Fig. 1a). It is such a ubiquitous property that we take it for granted to a point where there seems to be no reason to even provide it with a name. However, the concept becomes useful and important when considering video conferencing situations, where deictic consistency is generally violated (Fig. 1b). The reason is the offset (both direction and distance) between the location of the camera that records person X and the dynamically changing location of person Y as appearing on (or maybe better: somewhere behind) X’s screen. There are two different aspects to that offset that are important to discriminate. The first is the fixed offset between the location of the camera and the centre of the screen where the picture of the other person typically appears. For a typical 13 inch laptop screen with built-in camera used
from a distance of 70 cm that is about 9 deg vertically, but it can be much larger and can vary in different directions if a separate webcam is used. The second aspect refers to the generally smaller dynamic changes that occur due to movements of the observer’s head in front of their screen. Even for a sitting observer, involuntary lateral sway movements relative to a conversation partner at a distance of 1m will easily reach amplitudes of 5 deg and more.

The fixed offset between camera and screen can be taken care of by teleprompter designs where the camera is hidden behind a beam splitting mirror such that it appears to be behind the centre of the screen that displays the talking head (e.g. Doherty-Sneddon et al., 1997). The same can be achieved by employing machine-learning-based image processing to synthesise new views from a monocular video stream (T.-C. Wang, Mallya, & Liu, 2021). Other approaches exist that do not even attempt to generate new views of the whole scene but only manipulate the orientation of either eyes or the orientation of the face alone to account for the disparity between the location of the video camera and centre of the screen (e.g. Giger, Bazin, Kuster, Popa, & Gross, 2014; Kuster, Popa, Bazin, Gotsman, & Gross, 2012; Wood, Baltrušaitis, Morency, Robinson, & Bulling, 2018).

Note, however, that these approaches do not account for the dynamic part that is due to changing position of the user. Suppose X and Y are facing each other. If a real user X changes her location in front of a real interlocutor Y even only slightly, she either moves out of his line of sight, or Y has to respond to her move by making active eye or head movements to adjust to her changing location. The dynamic aspect of the Zoom problem is harder to deal with as it requires tracking the user locations. On the other side, fixing the dynamic, interactive aspect of the offset is probably more significant as it conveys information to control the closed-loop dynamics of a functional dyadic encounter. The violation of deictic consistency does not only affect directional behaviour between two communicating individuals, but it also disrupts a more general “sense of place”. Slater (2009) makes that point very clear, when he talks about the “place illusion” that users of virtual reality experience, but, of course, the experience of having a location in real space has the same causes and constraints, too. Here, we would not label it “illusory”. Being able to change location in space and receiving visual confirmation that the change has occurred is critical for the experience of a more general “sense of place”. Being in control of visual location is also the central ingredient to validate other deictic signals such as head orientation, directed gestures and eye orientation of another person, for establishing joint attention (Tomasello, 1995) and for perspective taking, i.e. the ability to form a theory of what another person can see from their particular location (Cole & Millett, 2019).

Being in control of place is also the basis for true dynamic eye contact behaviour. The lack of directionality in video communication means, that our visual system can no longer rely on the cues that trigger the eye contact effect and underlie the significance of eye contact behaviour as an important backchannel for human communication.

In real spaces where deictic consistency holds, person X perceives to be looked at by person Y if (and only if)
person Y is really foveating person X. There are exceptions, for instance in crowded environments where multiple people may be aligned along the line of sight of a looker, but those are rare. Our visual system therefore evolved to rely on and make use of the assumption of deictic consistency.

We should emphasize here, that communication without visual information is possible and can be very efficient and satisfying. We can enjoy a conversation in the dark or on the phone without video. Visual signals are missing in these cases, but we can take that into account.

The problem with Zoom and similar computer-mediated audio-visual conferencing systems is not so much the absence of directional visual cues, but rather that both communication partners receive a flow of directional cues, critical parts of which are spurious. The cues may trigger responses in our social brain that are not warranted. My visual system registers meeting someone’s gaze, but she may not look at me at all. While my cognitive systems might have learned that I should not rely on these mechanisms to function properly, my visual system may still trigger autonomic responses to spurious information. I never know whether she is currently looking at me or not. Directional cues are present, but they can no longer be relied upon. I have to actively suppress my natural, reflexive responses to them. That generates cognitive load and, in my opinion, is the main problem with current video conferencing platforms.

Of course, there are many other problems that make videoconferencing more exhausting and less efficient than face-to-face communication, but at least some of them relate directly to the lack of deictic consistency discussed here. For instance, Bailenson (2021) points out that in multiuser video conferences each single user feels more or less constantly being looked at, not just by one other person, but by all of them – which is rare in an in-person conference. What makes it worse is again, that there is no way to escape their gaze by changing location. Bailenson also discusses the additional cognitive load that results from the fact that discussants in video communication can no longer rely on the learned and evolved automaticity in their gestures and non-verbal body language. Rather they tend to vet, control, and moderate their own behaviour, being aware that it can potentially send spurious, unintended signals. As I am arguing here, these signals and the misunderstandings they cause are primarily due to misleading directional cues, including eye gaze.

Dissatisfaction with one’s own appearance is another factor that applies to computer-mediated audio-visual communication more than to face-to-face communication (Ratan, Miller, & Bailenson, 2022; Shockley et al., 2021). Again, the reason why face appearance dissatisfaction poses a larger problem in video communication may connect to the fact that the user has no reliable information about when exactly she is being looked at and when that is not the case.

**Virtual Reality**

The differences between real things in real space and pictures or movies taken of the same scene and projected
on a computer screen are not captured by simply stating that the former is 3D and the latter is 2D. Looking at pictures means doing two things at the same time. We see and recognize the medium as a three-dimensional, typically planar object in physical 3D space, but we also process the contents of the picture, the depiction, and understand it as representing a space that may exist elsewhere or that does not exist at all but in the imagination of the artist and the spectator. Art theoreticians talk about the “twofoldness” of pictures in this context (e.g. Wollheim, 1998).

While the canvas might be flat, the depicted space is not necessarily 2D. The picture might very well convey depth cues such as occlusion, linear perspective, texture gradients, shadows and shading that together provide detailed distance and 3D shape information and elicit a rich perception of depth. On the other hand, a picture typically lacks some other depth cues, most importantly binocular disparity, and motion parallax.

Speaking more precisely, pictures do not really lack these cues. Rather, binocular disparity informs the binocular observer about the planar shape of the medium. The same is true for motion parallax: A moving observer experiences a parallax pattern that speaks to the planar geometry of the canvas. Unlike available pictorial depth cues, binocular disparity and motion parallax do not provide information about the depth of the depiction but about the 3D shape of the medium.

Display systems have changed substantially over the course of human cultural and technological evolution. Figural paintings first appeared more than 40,000 years ago and were the only pictorial artifact for a long time. Remaining symbolic for a long time, it wasn’t until the 15th century that renaissance artists started to consider pictures as trompe l’oeil, that is, as a way to produce visual stimulations that mimic, at least to some degree, the ones our eyes sample when looking into the real world. Many of the more recent developments in display technology still follow that goal. Photography provided a means to freeze the optic array sampled at a fixed location in space with much more fidelity than paintings could. The invention of film added motion to pictorial representations. Soon after, attempts occurred to feed different stimuli to the two different eyes, conveying stereoscopic depth first to static, then to moving pictures. But it was not until technology was able to track an observer’s head and then use location information instantly to update the renderings on a stereoscopic head-mounted display (HMD) without perceivable lags that we can now also create artificial spaces with true deictic consistency and consequently the illusion of having a place in such spaces (Slater, 2009). We invented virtual reality. Technically, we just added another depth cue, self-induced motion parallax, to pre-existing display technology, but phenomenologically it triggered the leap from looking at imagery as a passive observer to becoming an active agent with control over location and orientation in the depicted space.

Note, that concurrent VR is still struggling with simulating some depth cues (e.g. accommodation, Cooper, Piazza, & Banks, 2012) but these do not affect deictic consistency.

With the introduction of motion parallax, VR has the potential to reintroduce true eye contact into computer-
mediated audio-visual communication. Neither binocular disparity nor any other depth cue is able to convey deictic consistency. A stereo camera streaming to a stereoscopic headset can convey a very powerful experience of depth, but unless the location of the two cameras within the virtual world is controlled directly by the user’s movements in space, deictic consistency cannot be achieved. Motion parallax is critically required to elicit the sense of place (Troje, 2019).

**Depth from motion parallax: MPDepth**

Virtual reality (VR) reintroduces deictic consistency and co-presence, and it has thus the potential to improve virtual meetings (Michael-Grigoriou & Kyrlitsias, 2022). However, VR comes with serious limitations. First, in order to implement binocular disparity, two images need to be delivered separately to the two eyes. To achieve this, users must wear awkward headsets that shield them from their real environment, limit their mobility therein, and cover face and facial expressions. Second, while it is straightforward to interact with computer-generated graphical contents (including avatars) there is no easy way to render a photorealistic copy of another person in virtual reality. Wearing an HMD means that important parts of the face are occluded, which makes video recording pointless, and hinders other ways to record facial motion and facial expression.

In the real world, binocular disparity and motion parallax are always coupled. However, putting the two cues into conflict can be easily done in virtual reality. Wang and Troje (2020) were experimenting in virtual reality with a tool they later called the “Alberti Frame” frame (X. M. Wang & Troje, 2023) that could either behave like an empty window frame or it could be set to behave like a screen onto which the same scene was projected that was visible through the window. The frame could also be toggled into two other modes. In “stereo-only” mode, the imagery within the frame behaved like a window with respect to binocular disparity, but still like a screen with respect to binocular disparity. In “parallax-only” mode the opposite was the case. The frame responded to changes in observer location with the same motion parallax that was perceived through the window, however binocular disparity was indicating the flatness of a screen. If both stereopsis and motion parallax were enabled, the frame became a window again that was indistinguishable from the original one.

Putting motion parallax and stereopsis into conflict generates sensations that reflect expectations our visual system derives from the usual coupling between the two depth cues. For instance, an observer moving laterally in front of the stereo-only display perceives illusionary parallactic motion in the opposite direction of the one that would be expected to be experienced in the real world (Allison, Rogers, & Bradshaw, 2003; X. M. Wang & Troje, 2023). Likewise, if binocular observers are asked to adjust the gain between their own lateral movements and the parallactic motion generated in the parallax-only mode of the Alberti Frame, they set it to a value that is reduced to 60% of the parallax they would experience in the real world (X. M. Wang et al., 2020). Monocular observers, presented with the same stimuli do not show that behaviour. They adjust the gain to values very close to what they would be in the real world (Fig. 2). It seems that stereoscopic depth generates predictions about
expected motion parallax which are then used to discount for the experienced image motion. As a result, the sensation of the parallactic motion becomes less salient. If the same motion occurs in the absence of binocular disparity, it is not predicted and appears to be more salient. There seems to be too much motion.

Fig. 2: In VR, observers were standing upright in front of two large openings representing two adjacent walls of a hexagonal gazebo located in a forest. The left opening was a normal open window. The right opening behaved like a window with respect to motion parallax, but like a screen with respect to binocular disparity (parallax-only condition, left and middle box), or it behaved like a screen with respect to motion parallax, but like a window with respect to binocular disparity (stereo-only, right box). Binocular, but not monocular observers perceived veridical parallax to be too large and scaled it down. No such scaling was observed for binocular disparity. Box and Whisker plots, adapted from Wang et al. (2020).

Interestingly, if observers are asked to adjust stereoscopic depth in the stereo-only mode of the Alberti Frame such that it fits the window stimulus, we do not see similar effects (Fig. 2, for more details, see X. M. Wang et al., 2020). In the absence of motion parallax, observers still adjust the simulated inter-pupillary distance to a veridical value. Binocular disparity seems to predict expected motion parallax, but motion-parallax itself does not predict binocular disparity.

This last observation probably contributes to the fact that, when presenting, again in virtual reality, visual stimuli in parallax-only mode there is no measurable carry-over that distorts perceived stereoscopic depth. Users perceive such displays as flat, on the one hand, but respond to directional signals in the same way as if they were presented in 3D with consistent cues from stereopsis and motion parallax (X. M. Wang & Troje, 2023). A non-stereoscopic screen which nevertheless responds to changing user locations by updating the projection such that it mimics the view that a monocular viewer would have received when looking at a 3D scene through a window is sufficient to re-establish deictic consistency.

We recently designed a system that implements the parallax-only mode of the Alberti Frame outside of HMD-
mediated VR on regular computer screens. We track the user’s head location in 3D space with a device mounted sensor (which may just be the built-in web camera) and then apply it to the rendering pipeline to simulate self-induced motion parallax. Binocular disparity, which, as we have seen, does not contribute to achieving deictic consistency anyway, is not provided, and consequently, no head gear is needed. The system can be used to display any object that can be modelled by 3D computer graphics on a flat screen in a space that provides deictic consistency. We refer to this approach as “MPDepth” (Depth from Motion Parallax).

The main application of MPDepth is not looking at inanimate objects in new ways, but it is avatar-based remote communication: A screen mounted sensor tracks the user’s location and uses it to steer a virtual camera that renders the avatar of the communication partner such that head movements in front of the user’s screen immediately result in the parallax and viewpoint changes expected if the two interlocutors were sitting on opposite sides of a window framed by the edges of the computer screen. Even the small movements of a stationary user sitting in front of a computer while engaging in a video conference, are enough to convey the perception of depth and reliable, faithful directionality.

The same sensor that controls the virtual camera is also used to track head pose and facial motion required to animate the user’s avatar on the screen of his communication partner.

Our current prototype allows us to systematically study the role of dynamic eye contact in dyadic communication, with several studies currently underway. The main manipulation in these experiments is the switch between two viewing conditions where the first renders the other person’s avatar from a fixed camera location, thus simulating the Zoom situation, while the other condition provides deictic consistency such that normal dynamic eye contact is enabled.

The implementation of that approach requires solutions to several different problems. The solutions that we employ in our current setups are based on the RGB-D sensor that Apple uses in iPads and iPhones and functionality that is available within Apple’s ARKit framework, on the one hand, and the game engine Unity, on the other hand. ARKit provides the tools to track a user’s head relative to the device using the RGB-D camera of newer iOS devices. That information is used locally to determine the location of the virtual camera that renders the remote person’s avatar on the local screen. Note, that this needs to be done quasi lag-less to convincingly convey the sensorimotor contingency between user movements and the resulting changes in the projection, that characterizes active motion parallax and thus provide a sense of place.

ARKit also provides a realtime stream of information about non-rigid facial deformations including movements of mouth and eyes. That information is required at the remote end of the communication line where it is used to animate the user’s avatar. Here, small transmission lags (generally <150 ms) are inevitable, but tolerable and no different from those experienced in conventional video conferencing.
Being a versatile game engine widely used for multiplayer games, Unity’s development framework also provides solutions to deal with such transmission lags and we take advantage of them. Finally, we employ Unity for avatar animation and rendering. The virtual camera used to project the avatar onto the screen is an off-axis camera, a functionality which is also conveniently provided within Unity. Off-axis projection discounts for the foreshortening that the projection on the screen experienced from a user location that is not anywhere on the central normal axis of the screen. It thus ensures, that the user, even when in an off-axis location, obtains the same optic array as if she were looking through a window framed by the edges of the screen.

The screens of both users are now behaving like a common window through which they can interact. The space between them behaves Euclidian. The line of sight is the same for both interlocutors and can serve as a reference that ensures deictic consistency. Deictic gestures, head orientation and gaze can be faithfully interpreted similarly to how the visual system interprets them in real space. Consequently, eye contact behaves veridically, too. If user X crosses user Y’s gaze she reliably knows that user Y also experiences eye contact. Also, if user X sees user Y rendered in the centre of her screen, she knows that user Y also sees her in the centre of his screen. Likewise, if user Y appears to be cut off at the edge of the screen, user X knows that she appears similarly to him. To center him again on her virtual window, she only has to move a bit away from the occluding edge of the screen. Users simply behave as if they are interacting through a real window. This also means that there is no need for our system to mirror back to the user their own rendering. Zoom and Skype need that as an artificial feedback that enables users to monitor how and where their communication partner sees them.

Of course, MPdepth implemented as outlined here, only works for setups where each user has their own screen. It does not work if multiple users are using the same screen. The rendering on that screen can only respond to the head location of one user. For that user, directional constistency is given, but for someone else who looks at the screen, the display won’t look right.

It is, however, possible to arrange setups where more than two people can meet – as long as they are all at different locations and all have their own setup. For a three-way meeting, each participant would need a round desk with one chair for the user and two screens that serve as windows behind which the two interlocutors appear. A similar setup might be feasible for four people but certainly not for larger groups. For one-on-one meetings, however, MPdepth can be used to create solutions that run on off-the-shelf hardware.

**Concluding remarks and future perspectives**

Computer-mediated audio-visual communication provides many advantages over audio-only telephony, but it also introduces problems that relate to the lack of deictic consistency which affects directional visual cues such as pointing, head orientation, eye orientation, and, as a consequence, the dynamics of natural eye contact. Virtual Reality, in principle, has the potential to alleviate that problem, but it introduces many others. In contrast,
MPdepth is a screen-based, head-free, very practical solution that only requires to simulate veridical motion parallax. Our current system is avatar-based, but given the fast development of viewpoint synthesis technology (e.g. Khakhulin, Sklyarova, Lempitsky, & Zakharov, 2022; Ma et al., 2021), we expect that we can soon manipulate a monocular video stream in realtime to achieve photorealistic MPdepth.

For some specific applications, representing the users as avatars might be advantageous. For instance, when supplying online technical support, both customer and support staff may want to retain anonymity. On the other hand, a technical conversation that involves instruction and trouble-shooting would very much benefit from functional eye contact as a backchannel to communicate attention and comprehension. If used to render visual chatbots, the use of clearly artificial avatars, whose appearance matches the artificial nature of the communication, may also be perceived as a welcome feature.

The lack of directionality and sense of place is not just a property of screen-based visual communication but is inherent to the perception of pictorial visual representations in general. A picture (in the wider sense, which also includes movies), by its nature, is not a virtual reality. It is not able to simulate the optic array the observer would receive if he were looking directly into the depicted scene. Pictures can be contemplated from viewpoints that differ substantially from the centre of projection of the picture (Vishwanath, Girshick, & Banks, 2005). If the human visual system would interpret a picture as a snapshot into a real 3D world, the deviations between the centre of projection and the current vantage point would translate into massive geometric distortion of the perceived geometry of the scene (Cooper et al., 2012; Farber & Rosinski, 1978). However, that is not the case. Humans perceive pictures as what they are: representations of the real world which behave very differently from the depicted scene itself. We adopt a different perceptual mode, it seems. We process pictures in “picture mode” not in the default “presence mode” (Troje, 2019). The distinction between perception in picture mode and perception in presence mode might map onto similar distinctions between vision-for-action and vision-for-perception and even on the distinction between the dorsal and ventral visual pathways identified by Goodale and Milner (1992; see also Nanay, 2011). Recent electroencephalographic data seem to corroborate the view that the brain treats images and solid objects categorically differently (Fairchild, Marini, & Snow, 2021; Snow & Culham, 2021).

We don’t know yet how deep this distinction goes and how it is represented in the human brain. But it is an interesting thought and if confirmed, it may challenge a basic assumption at the very root of vision research – namely that findings in vision research that are obtained with pictorial stimulus material generalize to human visual processing in the real world. It may also mean that the difference between visual communication on screens, on the one hand, and real or simulated face-to-face communication, on the other hand, is engaging different perceptual modules in our visual brain with potential consequences for the quality and efficiency of communication that go beyond the role of deictic consistency and eye contact.
References


Mareschal, I., Calder, A. J., & Clifford, C. W. (2013). Humans have an expectation that gaze is directed toward them. *Current Biology, 23*(8), 717-721.


