Guidelines for Social XR Implementation of Social Cues

Alexander Toet[1], Maikel ter Riet[1], Tina Mioch[1], Tom F. Hueting[1], Jan B.F. van Erp[1,2], Omar Niamut[1]

[1]TNO, The Hague, The Netherlands

[2]University of Twente, Enschede, The Netherlands

ABSTRACT

In our digital age, human social communication is increasingly mediated. However, mediated social communication (MSC) systems will only become a viable alternative for in-person social interactions when they reliably and intuitively convey all relevant social and spatial cues needed by communication partners to establish effective communication, collaboration, mutual understanding, and trust. A lot of research has been done on mediated communication in general, and on the effect of social cues in particular. However, this research typically focuses on particular social cues, investigating effects of implementations in particular situations. In addition, the research often focuses on answering psychological or cognitive questions. The translation to design questions such as functional and technical requirements is often left to the developers of the technical systems. This leaves a gap between psychological research results and a translation towards practical guidelines for designers and developers. This paper aims to make a first step towards guidelines for the implementation of social cues for social XR implementations.

Keywords: Mediated Social Communication, Social Cues, Social Presence

INTRODUCTION

Social cues are implicit behaviors that convey social and contextual information by verbal or non-verbal signals, like facial expressions, gestures, posture, and eye gaze. Social cues are essential to understand and predict the intentions and behavior of others and to modify one's own behavior in response. For example, eye gaze signals where someone's attention is directed, while gestures can be used to enhance or clarify verbalizations. In natural (face to face or F2F) settings, at least 70% of interpersonal communication is non-verbal, while verbal communication accounts for only at most 30% (Mehrabian 1972).

Nowadays, human social communication is increasingly mediated. Technologies like videoconferencing software (e.g., Zoom, Microsoft Teams, Skype, etc.) afford a new form of virtual togetherness by facilitating shared and synchronous social activities, thereby substituting face-to-face (F2F) interactions. New immersive (VR, AR or MR-based) communication systems extend regular video- or audio-conferencing tools by affording social experiences that even more closely approximate the experience of F2F meetings.

However, it is widely recognized that current mediated social communication (MSC) systems do not provide the affective experience of in-person social interactions or social presence. Hence, physical distance is still experienced as a barrier to effective communication. The main reason is that MSC systems still do not provide all relevant social and spatial cues needed by communication partners to establish effective communication, collaboration, mutual understanding, and trust (Hacker et al. 2019). To become a valid alternative for in-person social interactions MSC systems need to reliably and intuitively convey these cues.

This paper is a first attempt to derive guidelines for the implementation of social cues in mediated social communication (MSC) systems, based on existing knowledge about the use and significance of social cues in F2F communication.

SOCIAL CUES IN COMMUNICATION

Social cues in F2F communication can broadly be classified into verbal, nonverbal, and other cues. The cues that contribute most to creating a sense of social presence in MSC are eye movements, touch, facial expressions, body language, proxemic cues, verbal cues, and paralinguistic cues (Sharan et al. 2021). In this section we will discuss the role of each of these social cues in establishing a sense of social presence in MSC, and their implications for the design of MSC systems.

Eye Gaze

In natural communication, gaze (where one looks, how long, and when) plays an essential part in human social behavior. The movements, orientation, pupil size and blink frequency of the eyes convey important nonverbal signal that serve both

interpersonal and practical functions. Eye movements include looking, staring, and blinking. People unconsciously use pupil size and blink frequency to evaluate others: people with large pupils and slow blink frequency are perceived as more sociable and more attractive (Weibel et al. 2010). They also use eye contact to build trust: a steady gaze is usually perceived as a sign of honesty whereas a shifty gaze and an inability to maintain eye contact is seen as an indicator of deception. Eye gaze is also used to signal both the end and the beginning of a speaking turn during a social interaction. Realistic gaze behavior induces a sense of engagement and social presence (Bailenson et al. 2005). Gaze orientation can both convey and direct attention (Frischen et al. 2007). In conversations, gaze can be used to signal a person's level of interest and attention (Bavelas et al. 2006) or degree of comprehension (Beebe 1977). People can in principle detect direct gaze rather accurately, even over a video link (Monk & Gale 2002). Pupil size is an important social cue that people implicitly consider in their social behavior and decision-making: individuals with large pupils are generally evaluated positively by observers, while those with small pupils are perceived negatively. Pupil size also influences approach-avoidance behavior (Brambilla et al. 2019).

MSC systems should provide a representation of eye movements in response to events or social signals that is sufficiently fast to create the impression of high alertness (in contrast to low alertness). The representation of eye pupils should be of sufficient resolution and contrast to enable users to clearly perceive pupil dilations and contractions. To effectively support eye contact, the disparity between the optical axis of the camera and the viewing direction of the communication partner should be minimized (Eijk et al. 2010).

Touch

In natural communication, touch implies physical presence and proximity. The meaning of a social touch is highly dependent on the accompanying verbal and nonverbal signals of the sender and the context in which the touch is applied. In mediated social communication, the ability to touch a communication partner can significantly enhance the feeling of social presence. Mediated social touch, like normal interpersonal touch, can modulate physiological responses, increase trust and affection, help to establish bonds between humans and avatars or robots, and initiate pro-social behavior (van Erp & Toet 2015).

An MSC system providing social touch interaction should afford a reliable bidirectional transmission of the relevant parameters. To achieve realistic social touch through a multisensory MSC system, congruency of the multisensory signals in space, time, and meaning is of eminent importance. For instance, touch should be congruent with other (mediated) display modalities (visual, auditory, olfactory) to communicate the intended meaning. Touch signals should be highly synchronized (within 10 ms) with corresponding audio and visual signals. Especially in closed-loop interaction (e.g., when holding or shaking hands), signals that are out of sync may severely degrade the interaction, thus requiring (near) real-time processing of touch and other social signals and generation of adequate social touches in reaction.

Facial Expressions

Facial expressions are quite universal and form a large part of nonverbal communication, conveying emotions (like happiness, sadness, anger, and fear), and supporting speech and turn-taking behavior. The recognition of different emotions requires access to different facial features: the mouth is an important feature for the recognition of happiness, the eyes and eyebrows are used to assess sadness, while the recognition of fear involves a holistic processing of all these facial features (Beaudry et al. 2014). Eyebrow movements are probably the most relevant of all facial gestures in conversations. They play an important role in maintaining attention, thereby facilitating a sense of social presence. Facial expressions and bodily cues are processed as an integrated (Gestalt-like) unit rather than independently (Aviezer et al. 2012). Posture plays a role in eliminating ambiguity of facial expressions (Karaaslan et al. 2020). Hence, systems that transmit facial expressions should also reliably convey bodily movements. Social information communicated via both facial expression and body posture is based on subtle movements (Vesper & Sevdalis 2020). Therefore, for mediated communication to be experienced as a natural or face-to-face interaction, communication systems need to represent the subtle facial expression and body language cues in a naturalistic dynamic fashion. Viewing the mouth also supports the intelligibility of speech: in conditions where there is noise in the audio channel, people fixate more frequently and longer at the mouth of their conversation partner (Yi et al. 2013). There is also a significant interaction between facial expressions and gaze direction: direct gaze enhances the perception of approach-orientated facial expressions such as anger and joy, whereas an averted gaze enhances the perception of avoidance-orientated facial expressions such as fear and sadness (Liang et al. 2021).

Given that humans process facial expressions, gaze direction and bodily cues in an integrated and dynamic way, MSC systems should convey these features in a highly synchronized dynamic fashion.

Body Language

Body language consists of body and head posture and movements and hand gestures. Examples of body posture are arm- or leg-crossing, while head and hand movements include nodding, pointing, waving, etc. Posture can communicate attitude and emotional state (Dael et al. 2012). It can also be a measure of rapport that manifests itself through the psychological effect of mirroring, where a person mimics the body poses, gestures, or even general attentiveness of another, often subconsciously (Lafrance & Broadbent 1976). Gesture has many purposes ranging from illustrating words and ideas to directing attention and referring to objects, and is primarily used to accompany speech (McNeill 1992).

In MSC, behavioral realism determines the degree of social presence, especially when the user's representation signals mutual awareness. Users are more involved, mutually aware, and socially present when they can use hand tracking to create

meaningful gestures (Yassien et al. 2020). Accurate hand and finger movements support more subtle communication to express in-depth feelings or share more complicated information (Maloney et al. 2020).

Proxemics

The human brain represents a person's peripersonal space (PPS: the region where all physical interactions between the individual and the environment take place) through the integration of different (visual, auditory, olfactory, haptic) sensory inputs, which are all coded relative to specific body parts, or even to the body as whole, through somatosensory processing. Personal space is a nonverbal social cue, which depends on factors like sociocultural norms, situational factors, individual traits, familiarity, and task (Kinoe 2018). Proxemics refers to how people perceive their position in space relative to others (Hall 1966). It allows individuals to utilize space to communicate comfort, anger, friendliness, and standoffishness through four distance zones: intimate (<0.45 m), personal (0.45–1.2 m), social (1.2–3.6 m), and public (>3.6 m). Each distance zone has a specific range of proximity affording certain types of communication. For example, the intimate zone is common for communicating through physical contact activities such as expressing affection, comfort, physical stress, protection. People show similar proxemic spacing behavior in virtual worlds as in the physical world (Williamson et al. 2021). Perceived interpersonal distance (proximity) is a significant determinant of social presence and quality of communication in immersive VR (Ennis & O'Sullivan 2012). Additionally, in VR many other variables may influence perceived proximity, such as the appearance and behavioral characteristics of the users.

In MSC, proxemic cues like personal space and spatialized audio should be available to induce a feeling of social presence (Yang et al. 2020).

Verbal Cues

Linguistic cues refer to the choice of words and structure of sentences, i.e. dialect and syntax, that users employ during communication. Compared to non-spatialized audio, the availability of the spatialized voice and auditory beacons in a mixed-reality remote collaboration system significantly enhances the users' sense of social presence and the spatial perception of the environment (Yang et al. 2020).

MSC systems should provide spatial audio, preferably in combination with information on the users' gaze direction and hand gestures, to enhance social presence and mutual understanding. This will benefit tasks or discussions involving details or characteristics of their shared environment.

Paralinguistic Cues

Paralinguistic cues are social cues that are embedded in vocal communication but are separate from actual language or semantic content. This includes prosodic cues like vocal pitch, tone, loudness, and inflection. Prosodic synchrony in mediated

communication improves collective intelligence (problem solving ability; Tomprou et al. 2021).

The requirements for the use of audio and video information in MSC systems depends on their intended application. For social conversation, the system should provide both video and audio, and possibly haptic interaction, to simulate natural F2F communication. For decision making applications, audio-only may be preferred, since this eliminates the risk that certain communicating partners dominate the process by showing visually salient social cues turns (Tomprou et al. 2021). Vocoders that make voices sound more extrovert may be applied to enhance the sense of social presence (Lee & Nass 2003).

GUIDELINES FOR IMPLEMENTING SOCIAL CUES

MSC systems should provide a representation of eye movements in response to events or social signals that is sufficiently fast to create the impression of high alertness. The representation of eye pupils should be of sufficient resolution and contrast to enable users to clearly perceive pupil dilations and contractions. To support eye contact, the disparity between the optical axis of the camera and the viewing direction of the communication partner should be minimized. Touch signals should be synchronized and congruent with corresponding audio and visual signals. Given that humans process facial expressions, gaze direction and bodily cues in an integrated and dynamic way, MSC systems should convey these features in a synchronized dynamic fashion. Body language should be represented in a realistic manner since it determines the degree of social presence to a large extent by signaling mutual awareness. Proxemic cues like personal space and spatialized audio should be available to enhance the feeling of social presence. Spatial audio, in combination with information on the users' gaze direction and hand gestures, can enhance social presence and mutual understanding. Prosodic cues like vocal pitch, tone, loudness, and inflection should be provided when collective intelligence plays a role in mediated communication.

CONCLUSION

To become a viable alternative for in-person social interactions, MSC systems should reliably and intuitively convey all relevant social and spatial cues needed by the communication partners to establish effective communication, collaboration, mutual understanding, and trust, and to induce a true sense of social presence. Social cues that contribute most to creating a sense of social presence are eye movements, touch, facial expressions, body language, proxemic cues, verbal cues, and paralinguistic cues. This paper is a first attempt to derive guidelines for the implementation of social cues in mediated social communication systems, based on existing knowledge about the use and significance of social cues in F2F communication.

REFERENCES

Aviezer, H., Trope, Y., and Todorov, A. (2012), "Holistic person processing: Faces with bodies tell the whole story." *Journal of Personality and Social Psychology, 103* (1), 20-37.

Bailenson, J.N., Beall, A.C., Loomis, J., Blascovich, J., and Turk, M. (2005), "Transformed social interaction, augmented gaze, and social influence in immersive virtual environments." *Human Communication Research, 31* (4), 511-537.

Bavelas, J.B., Coates, L., and Johnson, T. (2006), "Listener Responses as a Collaborative Process: The Role of Gaze." *Journal of Communication, 52* (3), 566-580.

Beaudry, O., Roy-Charland, A., Perron, M., Cormier, I., and Tapp, R. (2014), "Featural processing in recognition of emotional facial expressions." *Cognition and Emotion, 28* (3), 416-432.

Beebe, S.A. (1977), "Effects of eye contact, posture, and vocal inflection upon comprehension and credibility." 37, 5436-5437. US: ProQuest Information & Learning.

Brambilla, M., Biella, M., and Kret, M.E. (2019), "Looking into your eyes: observed pupil size influences approach-avoidance responses." *Cognition and Emotion, 33* (3), 616-622.

Dael, N., Mortillaro, M., and Scherer, K.R. (2012), "Emotion expression in body action and posture." *Emotion, 12* (5), 1085-1101.

Eijk, R.v., Kuijsters, A., Dijkstra, K., and IJsselsteijn, W.A. (2010), "Human sensitivity to eye contact in 2D and 3D videoconferencing." In, *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, 76-81.

Ennis, C., and O'Sullivan, C. (2012), "Perceptually plausible formations for virtual conversers." *Computer Animation and Virtual Worlds, 23* (3-4), 321-329.

Frischen, A., Bayliss, A.P., and Tipper, S.P. (2007), "Gaze cueing of attention: visual attention, social cognition, and individual differences." *Psychological bulletin, 133* (4), 694-724.

Hacker, J., Johnson, M., Saunders, C., and Thayer, A.L. (2019), "Trust in Virtual Teams: A Multidisciplinary Review and Integration." *Australasian Journal of Information Systems, 23*

Hall, E.T. (1966), "The hidden dimension." Garden City, NY: Doubleday.

Karaaslan, A., Durmuş, B., and Amado, S. (2020), "Does body context affect facial emotion perception and eliminate emotional ambiguity without visual awareness?" *Visual Cognition, 28* (10), 605-620.

Kinoe, Y. (2018), "Interpersonal dstancing in cooperation: Effect of confederate's interpersonal distance preferences." In: J. Zhou, & G. Salvendy (Eds.), *Human Aspects of IT for the Aged Population. Applications in Health, Assistance, and Entertainment. ITAP 2018.* ,Vol Lecture Notes in Computer Science, vol 1092, 334-347:Springer International Publishing.

Lafrance, M., and Broadbent, M. (1976), "Group Rapport: Posture Sharing as a Nonverbal Indicator." *Group & Organization Studies, 1* (3), 328-333.

Lee, K.M., and Nass, C. (2003), "Designing social presence of social actors in human computer interaction." In, *SIGCHI Conference on Human Factors in Computing Systems*, 289–296:Association for Computing Machinery.

Liang, J., Zou, Y.-Q., Liang, S.-Y., Wu, Y.-W., and Yan, W.-J. (2021), "Emotional Gaze: The Effects of Gaze Direction on the Perception of Facial Emotions." *Frontiers in Psychology, 12* (2796)

Maloney, D., Freeman, G., and Wohn, D.Y. (2020), ""Talking without a Voice": Understanding Non-verbal Communication in Social Virtual Reality." *Proc. ACM Hum.-Comput. Interact., 4* (CSCW2), Article 175.

McNeill, D. (1992), "Hand and mind: What gestures reveal about thought." University of Chicago Press.

Mehrabian, A. (1972), "Nonverbal communication." Chicago, Ill, USA: Aldine-Atherton.

Monk, A.F., and Gale, C. (2002), "A Look Is Worth a Thousand Words: Full Gaze Awareness in Video-Mediated Conversation." *Discourse Processes, 33* (3), 257-278.

Sharan, N.N., Toet, A., Mioch, T., Niamut, O., and van Erp, J.B.F. (2021), "The relative importance of social cues in immersive mediated communication." In: T. Ahram, & R. Taiar (Eds.), *IHIET Future Systems 2021 Virtual Conferences*,Vol LNNS 319, 1-8, Cham:Springer Nature Switzerland.

Tomprou, M., Kim, Y.J., Chikersal, P., Woolley, A.W., and Dabbish, L.A. (2021), "Speaking out of turn: How video conferencing reduces vocal synchrony and collective intelligence." *PLOS ONE, 16* (3), e0247655.

van Erp, J.B.F., and Toet, A. (2015), "Social touch in human-computer interaction." *Frontiers in Digital Humanities, 2* (Article 2), 1-13.

Vesper, C., and Sevdalis, V. (2020), "Informing, Coordinating, and Performing: A Perspective on Functions of Sensorimotor Communication." *Frontiers in Human Neuroscience, 14* (168)

Weibel, D., Stricker, D., Wissmath, B., and Mast, F.W. (2010), "How Socially Relevant Visual Characteristics of Avatars Influence Impression Formation." *Journal of Media Psychology, 22* (1), 37-43.

Williamson, J., Li, J., Vinayagamoorthy, V., Shamma, D.A., and Cesar, P. (2021), "Proxemics and Social Interactions in an Instrumented Virtual Reality Workshop." *The 2021 CHI Conference on Human Factors in Computing Systems*, Article 253: Association for Computing Machinery.

Yang, J., Sasikumar, P., Bai, H., Barde, A., Sörös, G., and Billinghurst, M. (2020), "The effects of spatial auditory and visual cues on mixed reality remote collaboration." *Journal on Multimodal User Interfaces, 14* (4), 337-352.

Yassien, A., ElAgroudy, P., Makled, E., and Abdennadher, S. (2020), "A design space for social presence in VR." In, *11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, Article 8:Association for Computing Machinery.

Yi, A., Wong, W., and Eizenman, M. (2013), "Gaze Patterns and Audiovisual Speech Enhancement." *Journal of Speech, Language, and Hearing Research, 56* (2), 471-480.