Capturing Obstructed Nonverbal Cues in Augmented Reality Interactions: A Short Survey

Anton Nijholt

Abstract We present a short survey on recovering nonverbal communication cues that are hidden by head-mounted devices while interacting in augmented reality. The focus is on recovering facial expressions and gaze behavior by using various kinds of sensors that are attached to or integrated with these devices. The nonverbal cues can be made visible for other co-located or remote interactants on devices or avatars.

Keywords Augmented reality · Head-mounted devices · Nonverbal interaction · Facial expressions

1 Introduction

While in 1968, Ivan Sutherland's virtual reality device, hanging from the ceiling of its laboratory, could indeed be seen as a threatening "Sword of Damocles", nowadays we have headsets for augmented and virtual reality with a multitude of sensors and functions, almost resembling ordinary glasses, and also various ways to control the augmented environments and manipulate its virtual objects. We can make use of cheap sensors, powerful processors, and high-bandwidth networking. Head tracking takes care of providing a correct user's perspective of the world. Augmented reality (AR) cameras and computer vision make it possible to align virtual objects with the real world. Gaze, gestures, and voice commands, built-in or custom, can be used to interact with the environment. Techniques have become available for modeling and updating a user's spatial environment.

Head-mounted devices (HMDs) for AR and virtual reality (VR) have become well-known. Smart glasses, after a rather unfortunate start, are gaining attention again and are once again on the road to enter the consumer market. In AR video see-through, both the real and virtual scenes are available in digital form which allows detailed manipulation and alignment. In optical see-through (OST) AR, the user sees the real world directly with virtual objects superimposed on it. The HMDs are

A. Nijholt (✉)
Faculty EEMCS, Human Media Interaction, University of Twente, Enschede, The Netherlands
e-mail: a.nijholt@utwente.nl

S. Bhaumik et al. (eds.), *Proceedings of International Conference on Industrial Instrumentation and Control*, Lecture Notes in Electrical Engineering 815,
https://doi.org/10.1007/978-981-16-7011-4_1

bulky, uncomfortable, and sometimes weigh more than half a kilo. Less advanced AR is possible with smart (AR) glasses, head-up displays, sometimes hooked up to a smartphone or other handheld that is used as an additional head-down display and control.

Smooth human–human interaction requires the use and recognition of social signals, verbal and nonverbal cues that are used by partners in face-to-face and multiple-user interactions. These communication cues cannot always be detected when one or more interaction partners wear a head-mounted augmented reality device. Face regions are covered, eye and eyebrow movements are hidden, and facial expressions cannot be recognized. Moreover, users of these HMDs may have to deal with information that pops up during a task, and this may disrupt their communication and their exchange of communication cues with local or remote collaborators.

In this paper, we survey the research attempts that aim at recovering social signals, in particular those that can be detected from facial expressions, from AR devices that nowadays hide regions of the face that in unobstructed human–human interaction would provide information about the affective and cognitive state of an interaction partner. We aim at recovering those signals, maybe even enhance them, to design a more natural—not obstructed by technology—interaction between co-located partners or between partners, colleagues, collaborators, or instructors in remote locations.

2 Nonverbal Communication Cues

Nonverbal cues play an important role in social and other human–human interactions. We make gestures while speaking, we move our head and body, our facial expressions change, and our speech has nonverbal components. Listening has also many active components. Mimicry occurs naturally. While interacting, we can point at objects, pick up, manipulate, or assemble them, leaf through a book, use a tool, et cetera. Our attitude toward the interaction, for example, consent, approval, contempt, discomfort, impatience, confidence, interest, anxiousness, and especially focus of attention, is visible in our body language and facial expressions. It shows in movements of the eyes, eyebrows, and mouth, in our head orientation, body pose, and gaze, our gestures, and the proxemic distance we keep.

In AR, we will have situations where humans wearing AR HMDs communicate with each other, with people that do not wear them, and with avatars, visible in their HMD, that represent remote interactants. Besides, there can be virtual humans with their own algorithmic identity and intelligence. These latter interactants will not be considered here. Wearing AR devices disrupts the display and the detection of nonverbal communication cues. Can that be repaired, can they be translated to artificial signals, or can we even think of using the technology to enhance these cues?

AR research can profit from past human–computer interaction research done in the context of virtual reality, intelligent virtual agents, and affective computing.

Nonverbal interaction behavior has been studied in various large-scale European projects and networks of excellence on nonverbal augmented and multiparty (human–human, human-virtual agent) interaction such as augmented multi-part interaction (AMI) and social signal processing (SSPnet) [1]. Face-to-face interaction research is usually based on Sachs' turn-taking model of conversations [2]. This is a model with a speaker speaking, a listener providing nonverbal feedback, and continuous role switching. Within turns, there is also simultaneous expressive behavior that requires synchronization. Mimicking [3], for example, is a naturally occurring phenomenon. Gaze and facial expressions are sources of important nonverbal information.

Simultaneity and synchronization require timing and anticipation. In [4], more observations and examples of what we call "anticipatory synchronization" can be found. Performing joint tasks, for example, building a sandcastle, walking together, shaking hands, moving furniture, and cooperative and competitive game and sports activities require synchronization as well. In (future) AR where we deal with real and virtual objects, such activities require coordination and mutual adjustments as well.

In the next section, we will survey the many ways AR technology obstructs the communication of nonverbal cues during interactions and the attempts to reduce this obstruction. Many parameters need to be taken into account. Devices can obstruct the emitting and detection of social cues, causing problems with turn-taking, delay in responses, and keeping interpersonal distance. Lack of or reduced eye contact is another cause of poor and unbalanced communication. Interesting virtual content, not available to the non-user, can distract the AR user, and this can also lead to not emitting or paying attention to social cues in a conversation. Moreover, among the parameters are the different AR technologies. We have OST and video see-through devices. We can also look at hybrid AR devices, spatial AR, and handheld, eyes-down, AR. Moreover, we should take into account that AR technology is progressing, HMDs are becoming smaller, come to resemble ordinary glasses, and nevertheless have many sensors, that not only provide information about the environment but also about the users of AR devices and their interaction partner, whether or not this partner is using an AR or VR device. Our focus is on the use of HMDs that use OST AR.

3 Recovering Occluded Nonverbal Cues

Obstruction in perceiving nonverbal signals does not only occur while using AR or VR HMDs. For example, naturally occurring interaction supporting gestures are hindered by devices that must be operated simultaneously. But also darkened or mirrored glasses, ordinary glasses, scalp hair, beards, and mustaches, can make it difficult to recognize faces and facial expressions. Ear, nose, or lip piercings and make-up may have an impact too. A niqab only leaves the eyes uncovered. Face masks to prevent the spread of viruses cover the lower region of the face and confuse facial expression recognition software [5, 6]. Face masks that are decorated with facial expressions using LEDs and partially transparent face masks that let people see your facial expressions are available on the consumer market. Face accessories that are meant to evade facial recognition have been designed as well, including IR LEDs embedded in glasses that are intended to confuse surveillance cameras.

In contrast to situations where a user is isolated in his task, revealing obstructed signals from face, body posture, and hand gestures while using AR or VR devices is a necessity if we want to convey the users' eye gaze and emotions during social and collaborative interactions. The focus here is on the face because facial regions are the most noticeable occlusions caused by AR and VR technology, that is, by wearable and head-attached devices. Not seeing the other person's face degrades the quality and the efficiency of the interaction. Moreover, the information from eye and facial movements can be integrated with other information, such as head and body pose, nonverbal speech, and physiological information. Also, for this latter information, we should ask how the technology that is used may prevent otherwise naturally occurring nonverbal behavior. Being able to disclose social signals leads to the question of how to communicate that now digitally available information and for whom. This depends on whether we have co-located AR-supported face-to-face communication or mediated communication between multiple users in distant locations and a mixed-reality collaboration space.

Below we will review the most important approaches to recover eye contact, gaze awareness, and facial expressions from occluded facial regions. We should take into account that head-attached VR and AR devices are getting smaller and come to resemble ordinary glasses. Nevertheless, eye and eyebrow movements will be difficult or not at all to detect. Capturing of social and attention signals should also be done when we need to represent remote interaction participants as avatars in collaborative AR and VR.

3.1 *Private and Co-Located Use of Captured Behavior*

An early example of capturing facial expressions of a user wearing special glasses is "expression glasses" [7]. Piezoelectric contact sensors (strain gauges) are used that sense facial muscle movement around the eyes. Applications in which the measurements are communicated to other devices and to have feedback on one's own emotions are mentioned in this paper. Optical (infrared photo-reflective) sensors do not require physical contact with the user's face. For example, [8] reports about smart eyewear that uses 17 integrated sensors that measure the distance between the sensors and the skin surface of the user's face. Movements of facial muscles cause three-dimensional deformations of the skin. Mainly based on information around the eyes, eight different facial expressions are distinguished. This information is meant to be used, among other things, to give the spectacle wearer information about his affective states during his social interactions in daily life. Hence, although this research provides approaches to capturing facial expressions, it has not been done in the context of AR or VR research. That would preferably also require information about the head pose and gaze direction, for example, to have this information transmitted to other interactants, co-located or remote.

In a co-located situation, we can have interaction between a participant with HMD and collaborators or bystanders without HMD. VR HMDs have a longer history than AR HMDs, and several research papers have addressed this situation. That is, revealing an HMD user's face but also what she is looking at is useful knowledge for co-located partners that do not use an HMD. In the case of VR, the front of the HMD can be used to make information visible to bystanders. This is not possible with OST AR. If we can capture the face, it can be displayed elsewhere, visible to others. If the bystanders were to wear an HMD, it could appear more naturally in that HMD. We mention a few methods of face capturing in VR HMDs that are aimed at making the face visible to bystanders.

The face capturing methods are interesting for AR, but the display of the face should be done differently. For example, in [9], a 3D-face model of the user is rendered on the front of the VR HMD. Only the obscured area of the human face is displayed, giving the illusion of a transparent HMD. No eye-tracking of the HMD user is involved, just random movements of eyeballs, eyebrows, and eyelids. It is projected according to the bystander's head position which is tracked using a front-facing camera on the HMD. This method would conflict with the OST property of an AR device, and a different solution for the display should be designed. A similar observation can be made for the approach in [10]. Here a VR HMD has been given a front-facing screen displaying the scene the user is looking at, overlaid with a cartoon representation of the user's eyes and their movements. The latter, obtained with an eye-tracker, reveals the user's visual attention to nearby bystanders or collaborators. The HMD user is in control, but since a collaborator has the same view, they can discuss the scene and the next actions. Again, such a display cannot be realized on an OST HMD, and the display should be done somewhere else in the location or on a bystander's HMD.

In [11], the users wear OST HMDs with head pose detection and line-of-sight sensing. 3D face shape construction of a user's face with various eye expressions representing line-of-sights and eyelid motions takes place offline. The textures are overlaid in real time on the position and pose of the real faces wearing the HMDs. The system is used in a shared mixed-reality space where people share virtual objects and gaze awareness.

AR displays have become part of cross-device systems, distributed display environments, either for use by the AR user or by providing external observers or collaborators with a similar experience as the AR user. Smartphones and smartwatches can be combined with head-worn AR devices. The AAR system [12] uses a HoloLens AR HMD with an actuated head-mounted projector that shares the AR user's view with external users and that also allows view-dependent rendering of virtual objects on a nearby wall for an external user. In [13], the interaction between an AR HMD user and non-HMD users is also realized with an HMD combined with a dynamic projector that displays the augmented content onto planar surfaces.

Finally, in [14], a "headset removal" process is applied to a VR headset that incorporates eye-tracking. The HMD user performs in a virtual world and is observed in a live stream by the audience. A personalized face model together with a set of textures indexed by eye gaze has been captured offline. It is aligned and blended with the visual portion of the face in the live stream. The eye-tracking information is used to search the gaze database for a corresponding gaze image which is then rendered as appearing behind a translucent part of the headset. A smooth merging is obtained by interpolating between successive gaze images. The translucency helps to hide small imperfections.

3.2 Transmitting and Displaying Captured Nonverbal Information

Head pose, gaze, and facial expressions need to be captured to have them displayed on an avatar in a co-located or remote mixed-reality interaction situation. There is a variety of research addressing these issues. In [15], the aim is to capture head motions and facial expressions of HMD users and to implement them on their avatars that can perform in face-to-face conversations. The inside of the HMD is augmented with ultra-thin strain gauges that make contact with the face to measure the movements of the occluded upper face regions. Also, there is a separate head-mounted camera to capture the movements in the visible lower face region. In a subsequent paper [16], to improve the quality of face-to-face communication, the focus is on tracking mouth movements with the camera attached to the HMD and regressing the images to the parameters that control the animations of the lip movements of an avatar. In [17], it is shown that also optical (photo-reflective) sensors attached inside an Oculus Rift HMD can be used to distinguish the five basic facial expressions of an HMD user. Capturing of deformations caused by mouth movement is, however, limited. In their case, the expressions are reproduced in real time on a cartoon avatar. In [18], the IR gaze-tracking camera within the Oculus Rift headset is used to infer facial expressions of captured images of the user's eyes.

Projecting a 3D face model into a video mimicking the head pose and facial expressions of an HMD user is the aim of [19]. The visual parts of the face are kept while replacing those hidden by the HMD. The system is trained to learn the facial expressions of the user from the lower part of the face only. In real time, the upper part of the face and its expression changes are recovered in synchrony with the movements of the lower (visible) part. The reconstructed face can be transmitted to remote conversational partners or collaborators. In Microsoft's Holoportation project [20], persons are captured in full 3D and virtually teleported into a remote participant's physical space. Also here, the lack of eye contact when users wear their headsets was experienced as a limitation of the mixed-reality interaction. In this case, tiny inward-looking cameras on the HoloLens have been used to capture the eye regions, and mesh blending and texture mapping are performed on the 3D face geometry and shown as appearing being behind the HoloLens visor in a translucent view.

Rekimoto's "face-through HMD" [21] uses an infrared cut filter (an IR mirror) that reflects infrared light and is transparent for visible light. This mirror is placed between the user's face and the lenses of the HMD. Side-attached IR cameras capture the reflection from the eye and eyebrow region on the filter. A normal camera mounted on the bottom of the HMD in front of the face captures the other face regions, and with colorization of the IR image, the three camera images are merged on a prescanned 3D face of the user in real time. Although presented for a VR device, it will be interesting to see the method applied to an AR device that requires optical see-through. In [22], an HMD with face capturing is presented where the HMD enclosure is made from transparent material covered by IR pass filters. The filters block visible light but transmit infrared light. Therefore, an IR camera, mounted on the HMD and in front of the user's eyes, can be used to capture the user's face behind the IR pass filters, while the user sees unobstructed VR images. The IR images have to be colorized. Infrared cameras in the HMD, that is, behind the lenses, are used in [23] for capturing the eye regions, while an RGB camera records the visible face regions. Offline, a personalized 3D head model is constructed, and various head poses are recorded in a reference image data set. Head poses are tracked, and queried images are retrieved and synthesized with the camera images with accurate tracking of head motion, facial expression, and eye movement.

For completeness, we mention a recent paper [24], where filling in the HMD-occluded face region is done by using the information from a subject's face represented in RGB-D images. However, only a subject's head pose is taken into account. Changes in facial expressions and gaze behavior are not modeled. This makes the approach not yet suitable for conveying social cues during communication.

4 Conclusions

We discussed the importance of nonverbal cues in human–human interaction, how their generation, display, and perception are hindered by AR technology. Many approaches to recover nonverbal communication cues can be distinguished. This is an ongoing research topic that is not only important in the context of AR technology but also in the more general context of wearables, whether they are handheld, head-worn, on the wrist, or in clothes. In the research that was discussed, we see the influence of advancing technology on how the problem can be tackled, from the maker and craft approaches to the use of advanced and miniature sensors integrated into HMDs. We also see the lack of attention to other, non-HMD-oriented technology to measure affective and cognitive states of the interactants and use that information in the repair of occluded nonverbal communication cues. This other technology includes the use of head-worn EEG scalp or ear sensors, headbands with pressure sensors, or physiological sensors that help to indicate stress or attention. Moreover, AR interactions take place in smart environments, and the users and their devices become part of the Internet of Things (IoT) so that more information about users and their activities is known and can be used to support them.

References

1. Brunet, P.M., Cowie, R., Heylen, D., Nijholt, A., Schröder, M. (eds.): Special issue on conceptual frameworks for multimodal social signal processing. J. Multimodal User Interfaces **6**(3–4), 95–99 (2012)
2. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language **50**(4), 696–735 (1974)
3. Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., Pantic, M.: A multimodal database for mimicry analysis. In: D'Mello, M. et al. (eds.) Affective Computing and Intelligent Interaction (ACII 2011), vol 6974, pp. 367–376. LNCS, Springer, Berlin, Germany (2011)
4. Nijholt, A., Reidsma, D., van Welbergen, H., op den Akker, R., Ruttkay, Z.: Mutually coordinated anticipatory multimodal interaction. In: Esposito, A. et al. (eds.) Verbal and Nonverbal Features of Human–Human and Human–Machine Interaction, vol 5042, pp. 70–89. LNCS, Springer, Berlin, Germany (2008)
5. Yang, B., Wu, J., Hattori, G.: Facial expression recognition with the advent of face masks. In: 19th International Conference on Mobile and Ubiquitous Multimedia, pp. 1–3. ACM, New York, USA (2020)
6. Carbon, C.-C.: Wearing face masks strongly confuses counterparts in reading emotions. Front. Psychol. **11**, 566886 (2020). https://doi.org/10.3389/fpsyg.2020.566886
7. Scheirer, J., Fernandez, R., Picard, R.W.: Expression glasses: a wearable device for facial expression recognition. In: CHI '99 Extended Abstracts on Human Factors in Computing Systems (CHI EA '99), pp. 262–263. ACM, New York, USA (1999)
8. Masai, K., Sugiura, Y., Ogata, M., Kunze, K., Inami, M., Sugimoto, M.: Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In: 21st Conference on Intelligent User Interfaces, pp. 317–326. ACM, New York, USA (2016)
9. Mai, C., Rambold, L., Khamis, M.: TransparentHMD: revealing the HMD user's face to bystanders. In: 16th International Conference on Mobile and Ubiquitous Multimedia (MUM '17), pp. 515–520. ACM, New York, USA (2017)
10. Chan, L., Minamizawa, K.: FrontFace: facilitating communication between HMD users and outsiders using front-facing-screen HMDs. In: 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17). Article 22, pp. 1–5. ACM, New York, USA (2017)
11. Takemura, M., Ohta, Y.: Diminishing head-mounted display for shared mixed reality. In: Proceedings International Symposium on Mixed and Augmented Reality, pp. 149–156. Darmstadt, Germany (2002)
12. Hartmann, J., Yeh, Y.-T., Vogel, D.: AAR: augmenting a wearable augmented reality display with an actuated head-mounted projector. In: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20), pp. 445–458. ACM, New York, USA (2020)
13. Jansen, P., Fischbach, F., Gugenheimer, J., Stemasov, E., Frommel, J., Rukzio, E.: ShARe: enabling co-located asymmetric multi-user interaction for augmented reality head-mounted displays. In: 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20), pp. 459–471. ACM, New York, USA (2020)
14. Frueh, C., Sud, A., Kwatra, V.: Headset removal for virtual and mixed reality. In: ACM SIGGRAPH 2017 Talks (SIGGRAPH '17), Article 80. ACM, New York, USA (2017)
15. Li, H., Trutoiu, L., Olszewski, K., Wei, L., Trutna, T., Hsieh, P.-L., Nicholls, A., Ma, C.: Facial performance sensing head-mounted display. ACM Trans. Graph. **34**(4), 47 (2015)
16. Olszewski, K., Lim, J.J., Saito, S., Li, H.: High-fidelity facial and speech animation for VR HMDs. ACM Trans. Graph. **35**, 6, Article 221 (2016)
17. Suzuki, K., Nakamura, F., Otsuka, J., Masai, K., Itoh, Y., Sugiura, Y., Sugimoto, M.: Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head-mounted display. In: IEEE Virtual Reality, pp. 177–185. Los Angeles, CA, USA (2017)

18. Hickson, S., Dufour, N., Sud, A., Kwatra, V., Essa, I.: Eyemotion: classifying facial expressions in VR using eye-tracking cameras. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1626–1635. Waikoloa Village, HI, USA, (2019)
19. Burgos-Artizzu, X.P., Fleureau, J., Dumas, O., Tapie, T., LeClerc, F., Mollet, N.: Real-time expression-sensitive HMD face reconstruction. In: SIGGRAPH Asia 2015 Technical Briefs (SA '15). Article 9, pp. 1–4. ACM, New York, USA (2015)
20. Orts-Escolano, S. et al.: Holoportation: virtual 3D teleportation in real-time. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16), pp. 741–754. ACM, New York, USA (2016)
21. Rekimoto, J., Uragaki, K., Yamada, K.: Behind-the-mask: a face-through head-mounted display. In: Proceedings of the 2018 International Conference on Advanced Visual Interfaces (AVI '18), Article 32, pp. 1–5. ACM, New York, USA (2018)
22. Chiba, M., Yamada, W., Manabe, H.: Transparent mask: face-capturing head-mounted display with IR pass filters. In: The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings (UIST '18 Adjunct), pp. 149–151. ACM, New York, USA (2018)
23. Zhao, Y., Xu, Q., Chen, W., Du, C., Xing, J., Huang, X., Yang, R.: Mask-off: synthesizing face images in the presence of head-mounted displays. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 267–276 (2019)
24. Numan, N., Haar, F., Cesar, P.: Generative RGB-D face completion for head-mounted display removal. In: IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 109–116 (2021)