Augmented Reality and Affective Computing for Nonverbal Interaction Support of the Visually Impaired

Deniz Iren*
Open Universiteit, NL

Krist Shingjergji
Open Universiteit, NL

Felix Böttger
Open Universiteit, NL

Corrie Urlings
Open Universiteit, NL

Jelle Meindert Osinga
Eindhoven University of Technology

Sjoerd van de Goor
Eindhoven University of Technology

Damian Bustowski
Eindhoven University of Technology

Juliette Passariello-Jansen
Eindhoven University of Technology

Roland Klemke
Open Universiteit, NL
Cologne Game Lab

Abstract

Nonverbal cues such as gestures and facial expressions are indispensable in human communication. However, such an essential aspect of social interactions is inaccessible to the sight impaired. This issue can be alleviated with the assistance of augmented reality and affective computing techniques embodied as wearable technology. Even though both augmented reality and affective computing have been studied comprehensively, the real-life deployment and utilization of these techniques are hindered by the limitations of wearable devices in terms of computational capabilities and battery life. This calls for a holistic approach to implementing lightweight and robust affective computing methods. In this study, we present a prototype that combines facial expression and gesture recognition that is optimized to function on battery-powered wearable devices. Additionally, the prototype embodies a haptic sleeve that communicates the detected facial expressions and gestures to the wearer.

Index Terms: CCS [Human-centered computing]: Accessibility—Accessibility technologies; CCS [Human-centered computing]: Human computer interaction (HCI)—Interaction techniques CCS [Computing methodologies]: Artificial intelligence—Computer vision

1 Introduction

Sight is arguably the most powerful human sense, and the role of vision in human cognition has been acknowledged with expressions deeply rooted in language and culture. One says "*I see*" to mean "*I understand*". Thus, it is not surprising to see that mainstream research on Augmented Reality (AR) focuses almost exclusively on advanced visualization. Nevertheless, the broader definition of AR also covers the augmentation of other senses.

A large component of human communication consists of nonverbal cues such as facial expressions, gestures, and body language. Unfortunately, such an essential part of human social interactions is inaccessible to the sight impaired. This issue can be partially alleviated with the assistance of AR backed by wearable technologies and artificial intelligence.

The technical aspect of this problem can be broken down into three parts. First, the capturing of the images toward the direction faced. Second, the detection of the target individual, and the recognition of nonverbal cues. Third, the conveying of the detected cues by means of sensory augmentation. The real-life implementation of such solution approaches faces several challenges; the computational capability and battery-life of mobile devices are limited, and movement of the wearable camera makes it difficult to analyze the video stream to recognize nonverbal cues. To overcome these challenges, a holistic engineering approach is required.

*e-mail:deniz.iren@ou.nl

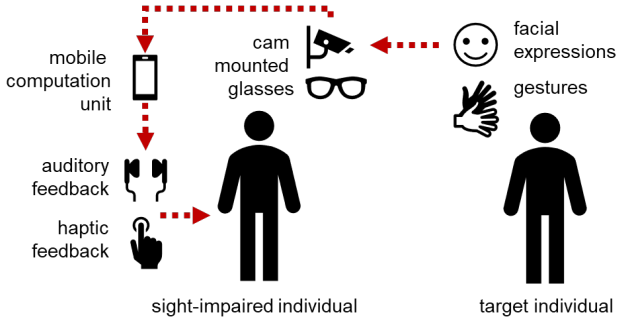

Figure 1: Conceptual representations of the proposed interaction system

This work-in-progress study addresses the design and development of affective computing models for detecting a variety of facial expressions and gestures of an individual: smile, frown, eyebrow raise, head nod, and head shake. Additionally, the integration with two sensory augmentation techniques is discussed and noted as future work. Fig. 1 depicts the overall conceptual representation of the proposed system.

The remainder of this paper is structured as follows: Section 2 highlights important literature on AR and affective computing. Section 3 describes the details of the prototype design. Section 4 covers a discussion on the interesting engineering challenges posed by the real-life implementation of AR and affective computing. Finally, section 5 concludes the paper.

2 Augmented Reality and Affective Computing

Affective computing aims at understanding and developing the technology that is capable of detecting, interpreting, and responding to human affect [20]. Human affect covers complex phenomena regarding the perception and manifestation of mood and emotions in humans. Felt affect goes hand in hand with observable physiological signals. Approaches for automatically recognizing human affect significantly differ based on the type of observed signals. For example, Face Expression Recognition (FER) models examine facial features to classify the displayed expressive cues on the human face [21]. Speech Emotion Recognition (SER), on the other hand, analyzes the

tone and content of human speech [2]. Moreover, Gesture Recognition (GR) attempts to detect gestures based on the nonverbal cues displayed by a variety of body parts such as hands and the head.

With the advances in artificial intelligence over the last decade, many novel approaches to recognizing human affects have been developed. FER domain primarily benefited from Convolutional Neural Networks (CNN) that were originally designed to address computer vision problems [12]. On the other hand, SER has been traditionally tackled using Recurrent Neural Networks (RNN) due to their ability to model long-range emotional context [11] while GR implementations mostly used Hidden Markov Models (HMM) and finite state machines [5].

Empowered by better datasets and advanced algorithms and architectures, recent studies on affective computing report promising results in terms of model performance and interoperability [22]. However, characteristics that are crucial to the deployment and real-life utilization of technology, such as model complexity and size, are often overlooked. This introduces a problem in the deployment of models that need to work on devices with limitations in terms of capability and connectivity. For instance, due to the limitations of battery and computation capabilities, running complex machine learning models on wearable devices is challenging [19].

Technologies that aim at assisting the sight-impaired in daily life comprise AR solutions that work on mobile and wearable devices. Such AR solutions primarily entail sensory substitution wherein auditory and haptic senses partially replace the function of sight. Several studies address the navigation problem [6, 18]. For instance, Albouys-Perrois et al. have designed a multisensory map to support individuals with low vision using tactile and audio feedback [1]. Liu et al. proposed a cognitive assistant that helps in navigation, obstacle avoidance, and scene formulation by means of replacing the visual sensory inputs with auditory ones [13]. Other examples are a simple mobile application that recognizes the objects on the camera feed and produces an audio output that names the object [15] as well as a system that recognizes obstacles using a 3D wearable camera on glasses and transmits the gathered information thought vibrations on a haptic feedback sleeve [23].

AR has also been used to support individuals with limited sight in their social interactions by visually enhancing some of the facial expressions of their hypothetical conversation partners [10]. McDaniel et al. used a haptic belt that informs the wearer regarding the direction and distance of the individuals within the visual field [17]. Such functionality is complementary to our proposed solution provided that it is extended to operate in 360 visual field, in order to inform the user about the location of potential conversation partners. In a more recent study, researchers implemented a prototype that uses a haptic belt to convey the inferred facial emotions [3]. This approach does not consider the context of the conversation. Thus, it runs the risk of misinforming the user about the inferred emotions.

Our approach differs from the solutions available in the existing literature in multiple ways, thereby making a contribution. First, instead of inferred emotions, our prototype provides the user with information about the detected gestures and facial expressions. The user has the flexibility to interpret the provided information based on the context of the conversation. Second, our GR model detects the rotation of facial landmarks instead of absolute spatial displacement. Thus, it is robust against the movements of the head-mounted camera that the user wears. Third, by using lightweight GR and FER models, the computational complexity is reduced. The combination of such lightweight models with adaptive conversation mode addresses the real-life challenge regarding the battery an computation limitations of wearable technologies.

3 Prototype Design

In this study, we designed a prototype in the form of a wearable device that aims at assisting visually impaired individuals in recognizing nonverbal cues during their social interactions. We considered the limitations imposed by wearable technologies and optimized the model performance and complexity accordingly. The prototype comprises three components: image capture, affect modeling, and AR feedback as depicted in Fig. 2.

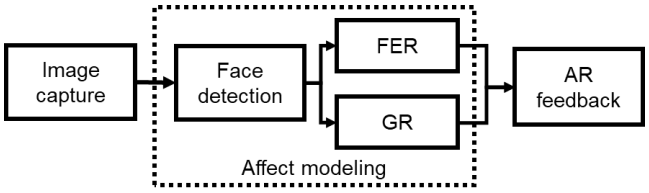

Figure 2: The process of the prototype operation

3.1 Image capture

The hardware elements of the image capture component are a camera mounted on a pair of glasses and a mobile computation unit. The camera is positioned toward the direction faced by the wearer. The images that are captured by the camera are transmitted to the mobile computation unit. In this study, we used a Raspberry Pi 4 with 4GB memory and a Logitech C270 HD 720p webcam. The camera was configured to operate at 30 frames-per-second and 640x480 resolution.

3.2 Affect modeling

Affect modeling component processes the images it receives from the camera. First, the face detection module analyzes the video stream one frame per second in search of a face. If a face is detected and it remains relatively close to the center of the frame for more than several seconds, the *conversation mode* is activated. If the face disappears or moves outside the center of the frame, the conversation mode is deactivated. Alternatively, the conversation mode can be switched on and off manually by the user. This provides flexibility when the user does not face the conversation partner directly or prefers not to use the system at all. FER and GR only work in conversation mode. In case there is more than one face in the frame, only the one that appears the closest to the camera and the center of the image is selected. Subsequently, the face detection model extracts facial landmarks; a (68x2) vector of coordinates. For face and facial landmarks detection, `dlib` is used [8]. Consecutively, the detected face image and landmarks are handed over to the GR and FER models that run simultaneously.

GR model detects two gestures; *head nods* and *head shakes*. These gestures are defined based on the rotational movements of three facial landmarks; 1: right ear, 2: nose center, and 3: left ear (see Fig. 3). Vertical rotation movements with alternating directions within a short time interval (e.g., two seconds) indicate a head nod. Vertical rotation movement is characterized by nose displacement that is significantly higher or lower than the average displacement of both ears. Similarly, rapid horizontal rotation movements with alternating directions indicate a head shake. Horizontal rotation movement is characterized by a simultaneous opposite change of two distances; d1: left ear to nose center, d2: right ear to nose center. To detect rapid vertical and horizontal rotation movements effectively, GR requires operating at 15 frames per second.

FER model recognizes three facial expressions; *smiling*, *frowning*, and *raised eyebrows* which are commonly associated with affective states of *joy*, *anger/frustration*, and *surprise* respectively. FER is performed based on the detection of facial muscle movements, i.e., Action Units (AU) [4]. Specifically, the presence of *AU6: Cheek Raiser* and *AU12: Lip Corner Puller* indicates smiling; *AU4: Brow Lowerer* hints at frowning, and *AU1: Inner Brow Raiser* and *AU2: Outer Brow Raiser* specify raised eyebrows.

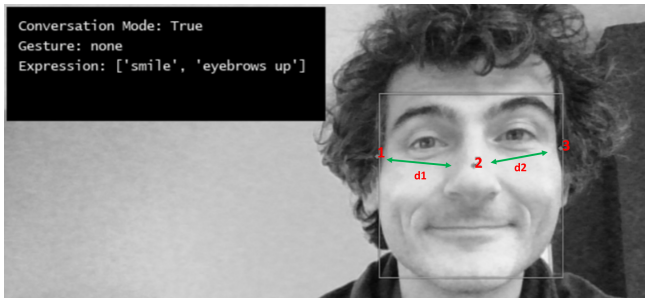

Figure 3: Screenshot of GR and FER model in action

Figure 4: The vibration motor array of the prototypical haptic sleeve

We designed a shallow CNN to detect the AUs, consisting of four convolutional layers with 32, 32, 64, and 64 filters respectively, and the ReLU activation function. A max-pooling layer followed the first three layers with a 2×2 filter, and the last one by a flatten layer. The final two layers are fully connected. The first fully-connected layer has 256 neurons, while the second has 12 sigmoid units representing the predictions of the 12 target AUs. We used a binary cross-entropy loss function and the Adam Optimizer. For training and testing the CNN, we gathered sample data from CK+ [14] and DISFA [16]; databases that consist of AU-labeled facial imagery. These databases include diverse imagery data in terms of age, gender, and ethnicity. CK+ has 593 videos from 123 participants, where each video sequence is the facial shift from a neutral expression to a targeted peak expression. We added the images that show the peak expression to our sample data. DISFA contains four-minute videos from 27 participants. From this data source, we sampled four positive and four negative images per AU, having 2,377 instances. The final set had 2,970 AU-labeled images. We performed stratified sampling with 2,376 as training and 594 as testing data.

We tested the trained CNN model on the allocated test set which yielded an average F-1 score of 77.12. Specifically, the classification performance on the individual AUs was 77.13 and 80.54 for AU6 and AU12 which indicate smiling, 75.08 for AU4 which indicates frowning, and 73.10 and 79.73 for AU1 and AU2 which are related to raised eyebrows.

The algorithmic complexity of both GR and FER are constant. The size of the FER model weights is 13.5MB, and it takes 0.08 seconds to initialize the model. The preprocessing of a detected face and facial expression inference approximately takes 0.09 and 0.04 seconds respectively. The GR does not require initialization and it takes 0.09 seconds to analyze one image frame. The software implementation uses separate threads to run GR and FER which enables the entire system to operate fluently.

3.3 AR feedback

The prototype is capable of providing feedback via two alternative sensory augmentation methods; haptic and auditory. The haptic feedback device is a sleeve that contains a set of 24 Tectonic TEAX09C005-8 [9] vibration motors in a grid (see Fig. 4). These motors emit distinct patterns over the surface of the forearm. After undergoing a training process to learn which patterns correspond to the aforementioned gestures and facial expressions, the wearer can interpret the haptic feedback.

The auditory feedback method uses standard headphones. It emits sounds that indicate when the system enters conversation mode. However, auditory feedback runs the risk of causing information overload and confusing the user because the spoken conversation also takes place in the auditory channel.

4 Discussion

The holistic implementation of AR, GR, and FER as wearable technology poses interesting engineering challenges some of which are discussed in this section.

Wearable technologies that rely on computer vision embody a camera that is mounted on the head or shoulder of the user. This means the camera is not stationary and that it moves due to the body movements of the wearer. FER models use single frames to detect expressions. Thus, the movement of the camera does not pose a serious challenge. However, GR requires a series of consecutive frames and therefore it is potentially hindered by a moving camera. In our design, we developed a novel GR technique that formulates the vertical and horizontal rotational movement of the head using the relative distances of facial landmarks. Thus, our design is robust against the effects of changes in the camera position.

Mobile wearable devices require a battery to operate, and every computation drains the battery. Thus, avoiding unnecessary computation is essential to an effective and efficient design. In our prototype design, FER and GR are only required to operate when the user is engaged in a conversation with a partner. Thus, we implemented the conversation mode which is activated either manually by the user or once a face is detected in front of the camera and stays there for a while. In doing so, unnecessary FER and GR computation is avoided which considerably improves the battery life. Moreover, due to the simplicity of the used architecture, the FER model is lightweight. While this positively impacts the computational efficiency and battery consumption it potentially compromises the accuracy. The optimization of computation efficiency and accuracy needs to be studied further.

Another interesting challenge is the communication of detected gestures and facial expressions to the user. By design, GR and FER continuously provide data on the detected nonverbal cues. Such a flow of data may cause information overload and distract the conversation. To alleviate this challenge, time series smoothing and aggregation techniques need to be used to simplify the communicated message.

This study has theoretical and practical implications. Existing research mostly focuses on individual topics, such as AR, GR, or FER. On the other hand, this study describes a holistic design that combines all three together. Additionally, it addresses several engineering problems that are faced in the real-life utilization of such technologies, explores the potential bottlenecks and integration challenges, and demonstrates approaches to overcome them. Furthermore, the prototype demonstrates that such a system has the potential to improve the life quality of sight-impaired individuals by enabling them to perceive nonverbal cues which are a crucial part of human communication.

5 Conclusion

In this paper, we introduce a prototypical AR system that aims at assisting the visually impaired to perceive the nonverbal cues of their conversation partners. The prototype comprises FER and GR models that are optimized to function on a battery-powered wearable device. This work differentiates itself from existing research by considering the computation and battery limitations of wearable devices and optimizing the model complexity and performance accordingly.

This paper presents a work in progress and therefore it is not without limitations. Even though the prototype has been tested, a scientific evaluation with the users is yet to be conducted. Also, our design decisions favored the simplest of models to decrease complexity and computational load. For instance, we used a primitive CNN architecture to promote lightweight model operation. Our observations indicate that wearable technologies could afford to operate more complex architectures that potentially yield improved accuracy.

We have clear goals for future work. First, we will thoroughly evaluate the components of the prototype. FER and GR need to be tested in both controlled and in-the-wild conditions, e.g., under different light conditions, and various usage scenarios. Second, we will conduct experiments with sight-impaired participants and evaluate the prototype comprehensively. Third, the mapping between the detected nonverbal cues and haptic feedback patterns needs to be further studied and evaluated in terms of user perception and ease of learning. Finally, we plan to benchmark various configurations of the prototype and measure and report complexity and accuracy to explore and highlight the optimal settings.

Acknowledgments

The authors wish to thank the brilliant members of Team HART [7] for their valuable efforts in creating technologies that actually improve human lives.

References

[1] J. Albouys-Perrois, J. Laviole, C. Briant, and A. M. Brock. Towards a multisensory augmented reality map for blind and low vision people: A participatory design approach. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14, 2018.

[2] S. Bromuri, A. P. Henkel, D. Iren, and V. Urovi. Using ai to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, 2020.

[3] H. P. Buimer, M. Bittner, T. Kostelijk, T. M. Van Der Geest, A. Nemri, R. J. Van Wezel, and Y. Zhao. Conveying facial expressions to blind and visually impaired persons through a wearable vibrotactile device. *PloS one*, 13(3):e0194737, 2018.

[4] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[5] R. El Kaliouby and P. Robinson. Real time head gesture recognition in affective interfaces. In *INTERACT*. Citeseer, 2003.

[6] S. Ferrand, F. Alouges, and M. Aussal. An augmented reality audio device helping blind people navigation. In *International Conference on Computers Helping People with Special Needs*, pp. 28–35. Springer, 2018.

[7] T. HART. Human augmentation research technology, Jan. 23, 2023 [Online].

[8] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[9] T. A. Labs. Teax09c005-8 technical specifications, Jan. 23, 2023 [Online].

[10] F. Lang, A. Schmidt, and T. Machulla. Augmented reality for people with low vision: symbolic and alphanumeric representation of information. In *International Conference on Computers Helping People with Special Needs*, pp. 146–156. Springer, 2020.

[11] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller. Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2021.

[12] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020.

[13] Y. Liu, N. R. Stiles, and M. Meister. Augmented reality powers a cognitive assistant for the blind. *ELife*, 7:e37841, 2018.

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101. IEEE, 2010.

[15] J. Y. Mambu, E. Anderson, A. Wahyudi, G. Keyeh, and B. Dajoh. Blind reader: An object identification mobile-based application for the blind using augmented reality detection. In *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, vol. 1, pp. 138–141. IEEE, 2019.

[16] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[17] T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan. Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. In *2008 IEEE international workshop on haptic audio visual environments and games*, pp. 13–18. IEEE, 2008.

[18] A. Neugebauer, K. Rifai, M. Getzlaff, and S. Wahl. Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study. *PLoS One*, 15(8):e0237344, 2020.

[19] A. Ometov, V. Shubina, L. Klus, J. Skibińska, S. Saafi, P. Pascacio, L. Flueratoru, D. Q. Gaibor, N. Chukhno, O. Chukhno, et al. A survey on wearable technology: History, state-of-the-art and current challenges. *Computer Networks*, 193:108074, 2021.

[20] R. Picard. Affective computing. perceptual computing section technical report. Technical report, TR 321. MIT Media Laboratory, 1995.

[21] K. Shingjergji, D. Iren, F. Böttger, C. Urlings, and R. Klemke. Interpretable explainability in facial emotion recognition and gamification for data collection. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8. IEEE, 2022.

[22] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022.

[23] M. Zahn and A. A. Khan. Obstacle avoidance for blind people using a 3d camera and a haptic feedback sleeve. *arXiv preprint arXiv:2201.04453*, 2022.