Video Conferencing in the Age of Covid-19: Engaging Online Interaction Using Facial Expression Recognition and Supplementary Haptic Cues

Ahmed Farooq[1,*], Zoran Radivojevic[2], Peter Mlakar[1], Roope Raisamo[1],

[1] Tampere Unit of Computer Human Interaction (TAUCHI), Tampere University, Finland
[2] Nokia Bell Labs, Cambridge, UK

[1]{FirstName.LastName}@tuni.fi
[2]{FirstName.LastName}@nokia-bell-labs.com

Abstract. More than 50 years since its mass market introduction the core user interfaces of Video Conferencing (VC) systems have essentially been unchanged. Relaying real time audio and video over distance is inherently productive. However, it lacks the sense of in-person interaction. With the current global pandemic, additional privacy concerns over the extended use of video and audio-conferencing systems, there is a need to redefine how VC Systems function and what information they communicate. To resolve these issues, we propose a VC system that utilizes facial recognition to identify and catalog participant's expressions and communicates their emotional states to other participants on the VC system using encoded haptic cues. In our testing we found that the approach was able to provide summarized haptic feedback of facial expressions and reduce the time it took for the participants to react to ongoing discussions without increasing mental or physical strain on the user.

Keywords: Human-Computer Interaction · Facial Recognition · Haptics · Human-systems Integration · Video Conferencing

1 Introduction

Online video conferencing is a key business tool for remote interaction. Since 1968, when AT&T introduced the concept, various systems have gained traction as viable, secure, and efficient ways to communicate in real time over distance. Subsequently, as communication infrastructure has improved, current video conferencing systems (VCS) have become very efficient at relaying visual and auditory information in high definition. However, interpersonal communication goes beyond a two-dimensional video feed of a communication partner. Febrianita and Hardjati [1] argue that without sufficient nonverbal cues such as facial expressions and body language, the effectiveness of in-person virtual communication can be substantially reduced. Van den Bergh and colleagues [2] suggest that soft skills acquired for in-person interaction may not be completely carried over to virtual interaction, especially in complex or emotional situations. As participants of such systems reside in different countries with different languages and cultural traits, the efficiency of communication may further be reduced. With the current global pandemic and the extended use of video conferencing systems to replace in-person interaction, there is a need to improve the technology and how we interact with it for enhancing user experience as well as providing a more personalized exchange between individuals and groups. For that reason, we propose a low cost video conferencing system similar to the one proposed by Myers and Secco [3], that utilizes facial recognition to identify and catalog its participant's expressions and communicate their emotional state using encoded haptic cues.

2 System Design

The setup was developed by training a neural network to identify facial expressions of video conferencing participants and by providing 3.5secs haptic feedback cues to communicate their facial expressions with respect to five common emotions (neutral, angry, happy, sad and surprised). After testing different APIs and open source libraries we developed the application setup on Tensorflow 2 in Python using Dlib and OpenCV [4]. The system employed a web camera to first extract the live image of the participant. By using a Haar cascade face detector (OpenCV) we used the input image to extract 51 salient points of the sampled face (nose position, eye shape, eyebrow shape, mouth shape etc.). These points were then normalized with the IMAGE_WIDTH, IMAGE_HEIGHT parameters and then matched to the data set and fed to the recognition model [5] (Tensorflow) within the neural network (Fig.1).

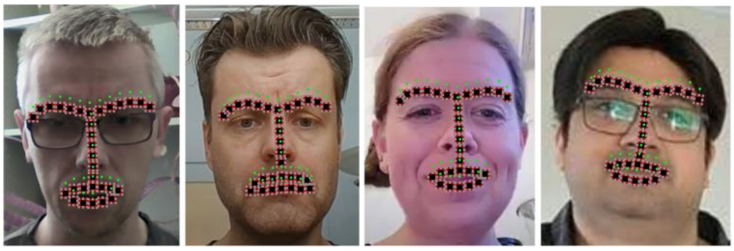

Fig. 1. (left to right) Angry, Sad. Happy, Neutral and Excited emotions recorded by the system

The generated output was a vector of 5 elements corresponding to a probability distribution of the 5 different emotional responses being identified: neutral, angry, happy, sad and surprised. These vector responses were accumulated over a period of time measured as Refresh_Time_Seconds parameter, which was set to 3.5secs for this study. An averaged response vector was computed, and the maximum average emotion was selected as the emotional response to be transmitted through the haptic feedback wristband where the emotional responses are stored as sound files.

To create distinct yet recognizable haptic feedback, we developed custom vibrotactile signals for each emotion response. The goal was to create natural tactile signals that can easily be identified by users with limited training. Therefore, we modulated natural auditory signals to simulate 3 core haptic primitives: human heartbeat, human scream, and a drum-bass combination. Each signal was divided into three segments, while adjusting the tempo and pitch of the second and third segment of the signal helped characterize the entire signal as positive or negative feedback. Using heartbeat as the base primitive, we created feedback for “neutral” and “surprised” emotions by increasing / decreasing the rate of the heartbeat (Fig.2 a & e). Similarly, we utilized the drum-bass beat as the primitive to modulate “happy” and “sad” emotions (Fig.2 c & d), and a modulated scream was used as a representation of the “anger” emotion.

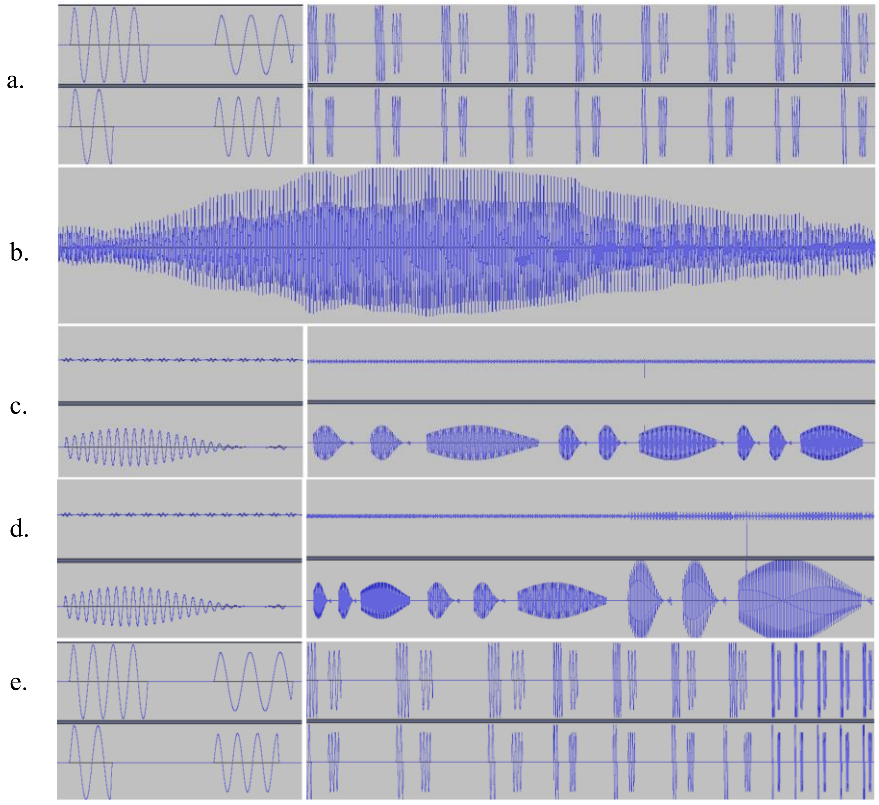

Fig. 2. Modulated feedback signals (from top to bottom) Neutral (a), Anger (b), Sadness (c). Happiness (d), & Excitement (e).

We applied these signals to the participant through a wearable palm device. The device was developed using a Tectonic TEAX14C02-8 voice coil actuator attached to the inside of the palm using a Velcro strap similar to Farooq et al., [6]. We piloted various adaptations of wristband and palm devices that could relay the custom designed signals, and concluded that the feedback parameters (duration, frequency, amplitude) were ideal for stimulating the inside of the palm. As all the 5 signals were of a similar duration (3.5sec) and required similar amplitude, we used a standard D-class amplifier with a peak amplitude of 5.8v.

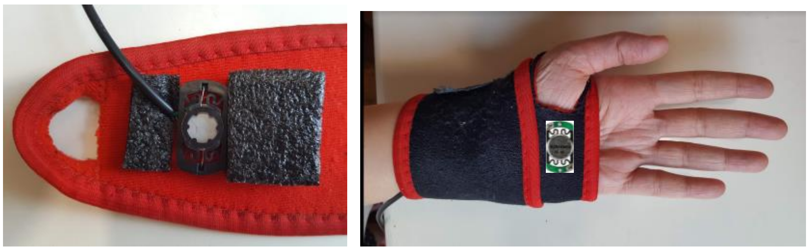

Fig. 3. Design and placement of the wearable palm device.

3 User Test

We conducted a 24-participant pair-wise study where two unfamiliar participants were acting as both as a presenter and as a listener (counter balanced) in two different sound isolated rooms. In the presenter mode the participants were asked to replicate 15 randomly generated facial expressions from five selected emotions (5X3) into the VC system which were recorded and played back to the listener. The listener was asked to identify each expression using three conditions: 1) by directly viewing the presenter's recorded video, 2) by listening to the presenter's recorded audio and 3) by only sensing the system's haptic feedback signals on their hand using the palm strap device. The duration of each feedback was 3.5sec and the participants were instructed to reply as soon as the feedback ended. At the end, the listener rated the task load (NASA TLX) and accuracy. We also measured the time it took for the listener to respond to each facial expression as well as their accuracy compared to the VC system.

Visual feedback was provided using Skype over a Samsung B2440L 24-inch monitor (1080p) with the VC emotion recognition software running in the background. Audio feedback was also provided using Skype where the listener's monitor was switched off and the participants were wearing noise canceling wired headsets. The presenter recorded the text message "*This is the presenter mode for feedback X*", where "X" was the number of feedback (1-15). The presenter was asked to convey their emotional state by amending their delivery of the text message, altering tone, annunciation, speed, and intensity. Haptic feedback was provided as five custom designed feedback signals shown in Fig.2 via the custom palm device (Fig.3). Once all the data for one participant was collected the presenter and listener switched roles.

4 Results and Discussion

Results of the NASA TLX questionnaire (Fig.4) showed that identifying emotions with audio only feedback was the most difficult task. Users rated audio-only modality as more mentally and temporally challenging compared to haptics only and visual only conditions. Results also showed that the participants found audio-only condition to require the most effort and it was more frustrating to manage compared to visual and haptic only conditions. Haptic only condition was rated as similar in frustration and effort to visual only condition, but more mentally and temporally challenging.

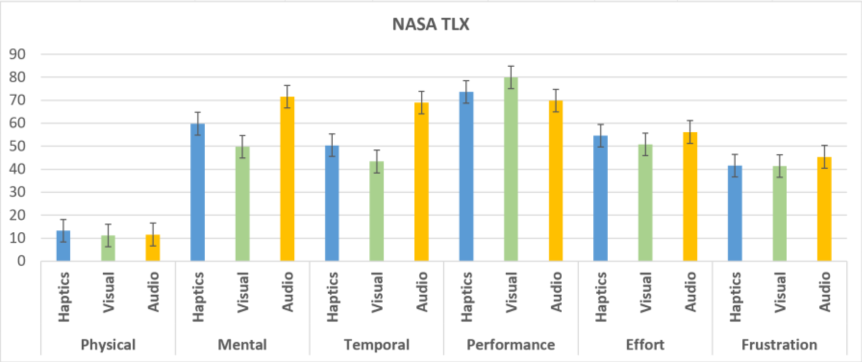

Fig. 4. Mental and Physical demand measured for each condition using NASA TLX.

Looking at the response-time measurements (Fig.5), we see that there were minor differences between the three modalities. However, there was a trend showing that the first task of each modality took longer to complete than the others. Audio only modality consistently remained the slowest across all tasks and conditions. This trend continues in recorded errors (Fig.6) where we see that participants made more mistakes for audio only modalities compared to haptic and visual only conditions.

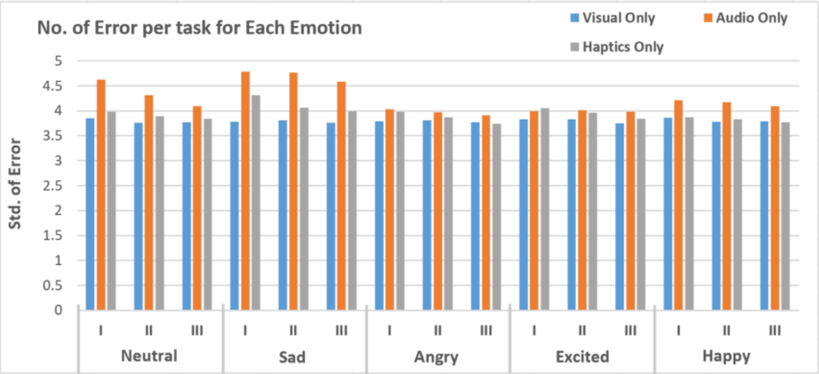

Fig. 5. Response time per task for each of the five emotions.

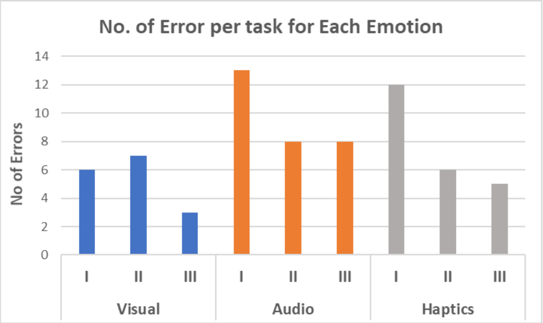

Fig. 6. Number of errors per task for each of the five emotions

If we consider the number of errors (Fig.6) we see that visual only task was the most accurate followed by haptics only, while audio only was not only slower (Fig.5) and difficult to perform (Fig.4), but produced the most errors (per task and in total). Interestingly, we observed more errors recorded for the first task in each modality condition for both audio and haptic. The results continued to improve during the study for both conditions. This indicates that there is room for learning for better performance.

5 Conclusion

We proposed a novel method of communicating over a video conferencing system where the presenter's facial expressions were used to encode emotional information and relay to the listeners using real-time vibrotactile feedback. A user study with 24 participants was conducted. Results showed that encoded haptic feedback helped participants to identify and respond to the presenter faster, with fewer errors and with no significant additional reported stress. Participants performed better at identifying facial expressions using visual-only condition and the VC system with encoded haptic feedback, over audio only condition. Moreover, the listeners identified the encoded haptic feedback signals with a high degree of accuracy and rated the overall system positively, classifying the periodical non-visual haptic information as a novel and informative aspect of the system. However, participants performance and rating of audio only condition was the lowest, illustrating that VC system without video input can be more frustrating and can lead to misunderstanding and misrepresentation.

References

1. Febrianita, Roziana & Hardjati, Susi. (2019). The Power of Interpersonal Communication Skill in Enhancing Service Provision. JOURNAL OF SOCIAL SCIENCE RESEARCH. 14. 3192-3199. 10.24297/jssr.v14i0.8150.
2. Canto S., Jauregi K., Van den Bergh H. (2013). Integrating cross-cultural interaction through video-communication and virtual worlds in foreign language teaching programs: is there an added value? In Proceedings of Huub.ReCALL: Journal of EUROCALL; Cambridge Vol. 25, Iss. 1, (Jan 2013): 105-121. DOI:10.1017/S0958344012000274
3. Myers K., Secco, E. L. (2020) A Low-Cost Embedded Computer Vision System for the Classification of Recyclable Objects. In: Soft Computing Research Society and Congress on Intelligent Systems (CIS) 2020, September 05-06, 2020, Virtual Format.
4. Gupta N., Sharma P., Deep V., and Shukla V. K. (2020) Automated Attendance System Using OpenCV.In proceedings of 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 1226-1230, doi: 10.1109/ICRITO48877.2020.9197936.

5 Reny J., A Convolutional Neural Network (CNN) Approach to Detect Face Using Tensorflow and Keras (May 30, 2019). International Journal of Emerging Technologies and Innovative Research, ISSN:2349-5162, Vol.6, Issue 5, page no.97-103., Available at SSRN: https://ssrn.com/abstract=3599641

6. Farooq A., Evreinov G., Raisamo R. (2020) Enhancing Multimodal Interaction for Virtual Reality Using Haptic Mediation Technology. In: Ahram T. (eds) Advances in Human Factors in Wearable Technologies and Game Design. AHFE 2019. Advances in Intelligent Systems and Computing, vol 973. Springer, Cham. https://doi.org/10.1007/978-3-030-20476-1_38