Integrating Video with Artificial Gesture

Jóhann Ingi Skúlason
Reykjavík University
Menntavegur 1, 102, Reykjavík
johanns19@ru.is

Jón Skeggi Helgason
Reykjavík University
Menntavegur 1, 102, Reykjavík
jonh19@ru.is

Anna Sigridur Islind
Reykjavík University
Menntavegur 1, 102, Reykjavík
islind@ru.is

Steinunn Gróa Sigurdardóttir
Reykjavík University
Menntavegur 1, 102, Reykjavík
steinunngroa@ru.is

Hannes Högni Vilhjálmsson
Reykjavík University
Menntavegur 1, 102, Reykjavík
hannes@ru.is

Abstract

Modern video communication applications provide a richer communicative experience than traditional telephony. However, compared to in-person communication, video communication applications fall short, in part due to the lack of a shared space and a common frame of reference for non-verbal cues. Our hypothesis is that an application which emulates in-person communication in a shared space, while leveraging off video communication, results in a richer social experience. With that in mind, we have designed and developed an application that combines videotelephony and video game technology. Users control an avatar from a first person perspective in a shared 3D virtual environment while their webcams track and capture a live feed of their face that gets rendered onto their avatar. The avatars then use an automatic gesture system to produce essential non-verbal cues, for instance turn-taking cues, in the joint virtual environment. Early prototype testing without gesturing already showed promise, but user feedback led to the development of artificial gesture. This is ongoing research and the next steps are to conduct further extensive user testing, focusing on measuring the effectiveness of the new automatic gesture system.

1. Introduction

The demand for quality video communication increased drastically in 2020 due to the COVID-19 pandemic. Although the available video communication tools have been praised by many, they are not flawless. The goal of this project is to design, develop and test a new type of telecommunication software. The application we developed focuses on improving user experience by increasingly emulating in-person communication. The application places users as 3D avatars in a virtual environment where they can see and hear each other. Unlike traditional 3D avatar based interaction, a live video feed captured from each user's face is displayed inside the helmet of their avatar (see Figure 1). The benefits of the application are mainly that it i) provides users with a common space and freedom to move around; ii) allows for multiple simultaneous conversations in a video-based chat room and; iii) supports more natural group dynamics by enabling users to locate and approach each other using the direction and volume of their voices.

The results from testing an early prototype with 24 users, without any gesturing, were already encouraging. Participants enjoyed the freedom to roam and mingle, and the whole concept of an avatar-video hybrid made a strong positive impression. However, user feedback also indicated some difficulty in coordinating conversations. Based on those results, we decided to take the emulation of in-person communication even further, which has brought us to the work described in this paper. Since it has been demonstrated [2, 17] that automating useful non-verbal cues in online avatars can support communication, including coordination, we decided to add automated co-verbal gesture based on user voices. In particular we aim to support social awareness and turn-taking.

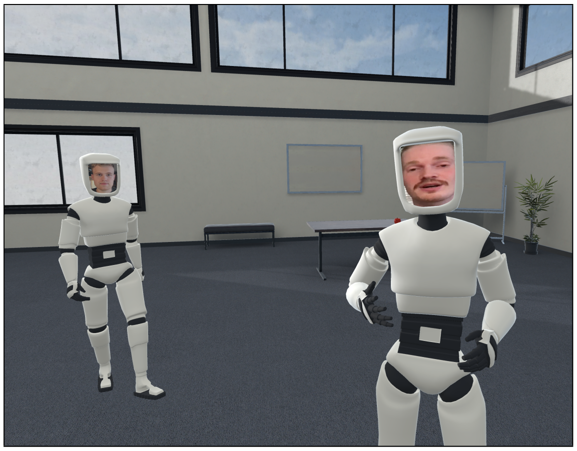

Figure 1. The user on the right is conversing with you, while the user on the left is on their way to join the conversation. Every user is seeing this from their point of view. The hand movements are fully automated.

The addition of video to conventional telephony has revolutionized the medium. Today, using video communication tools is a well established way of conducting relatively effective remote meetings. However, this technology has in fact not changed much over the last decades and we are steadily approaching a turning point where we need to ask ourselves: What are the next steps in this evolution?

It is an increasingly common opinion that virtual reality (VR) is the next logical step in the evolution of virtual communication [11]. Realizing that vision at a massive scale is however years down the road, considering how inaccessible VR hardware is for the general public, both in terms of costs and usability. The immersion that VR provides has brought well-deserved attention to VR but there is an important feature that is currently missing: The face. Seeing people's faces provides vital visual cues that both add depth to human communication and support its flow [4, 3]. Attempting to accurately simulate or even track facial expressions is a challenge that can lead to undesired results. Slightest imperfections can lead to a loss in translation due to the delicacies of facial expressions [15]. Moreover, conventional hardware is accessible to the general population and instead of relying on specialized equipment (such as for VR), we focus on creating a tailored desktop application that will run on regular laptops with a web cam.

To recap, the goal of the proposed application is to increasingly emulate in-person communication, by combining modern videotelephony and online video game technology, using the same conventional hardware used in most videotelephony solutions. The application is meant to be both accessible and familiar.

The application functions much like a modern first-person video game. Users control a humanoid avatar in a 3D virtual environment from a first person perspective. A webcam feed is processed by a facial detection module and each user's face is cropped and rendered onto their respective avatar inside a space helmet. Voice audio, captured by the microphone, is perceived by each user as a spatialized sound to further emphasise the three dimensional experience.

The users control their avatars with the popular WASD control scheme and a mouse. The WASD keys on the keyboard are used in most modern first person PC video games and are utilized to move characters forward, backwards and side-to-side in 3D space. Users press a push-to-talk button to broadcast their voice audio, the audio data is simultaneously processed to generate automatic gestures for the speakers' avatar. Since the users real hands are occupied with controlling the avatar, the non-verbal gestures, that support the interaction, are automated and brought into the joint 3D space, as that has been shown to be an important aspect in previous research[2]. The gestures are a combination of hand movements and subtle head tilting. An example screenshot of the application in use can be seen in Figure 1

2. Related Work

The importance of facial expressions in human communication is well known [1], not only for displaying emotion [6], but also for supporting a number of linguistic and interaction coordinating functions [3]. When moving human communication from the physical setting to an online setting, the ability to see faces has contributed to the success of videotelephony [9], even though not all aspects of face-to-face communication are preserved. Communicating online has, especially in recent times, become an important aspect of our everyday life, both for socializing with family and as an integral part of modern work. For working together remotely, and for facilitating collaboration and communication at a distance in a professional manner, taking turns in a conversation is a particularly important aspect [16].

In addition to facial expressions, co-verbal hand gestures and body posture also convey important information, both related to the content and coordination of interaction [14, 10, 13]. Prior studies of successful coordination in face-to-face interaction have shown that even subtle hand movement, for example bringing hands up from a resting position into the so-called gesture space in front of one's torso, can indicate the intention to speak, while dropping them down again indicates the end of a speaker's turn [8, 5, 10]. In virtual communication, letting users control these gestures themselves can take the attention away from the actual conversation [7], since they are normally produced unconsciously. To address this in avatar-based online interaction, prior research has shown how automating avatar co-verbal gesture by monitoring user activity, can supply missing cues and improve online conversation without bur-

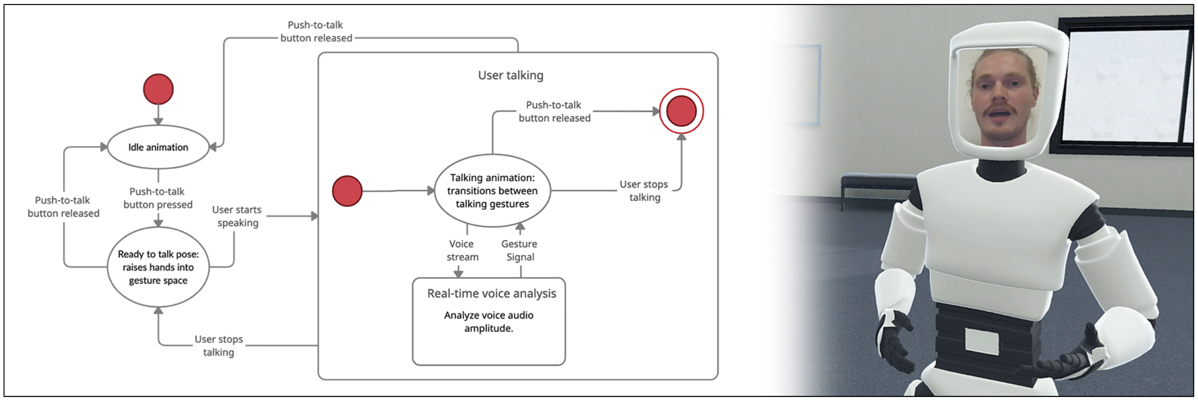
Push-to-talk button released
User talking
Push-to-talk button released
Idle animation
Push-to-talk button released
Push-to-talk button pressed
User starts speaking
Talking animation: transitions between talking gestures
User stops talking
Ready to talk pose: raises hands into gesture space
Voice stream
Gesture Signal
Real-time voice analysis
Analyze voice audio amplitude.
User stops talking

Figure 2. A high level nested state diagram for the automated gesture system. The red dots without an outline represent an entry and the outlined red dot an exit. The image on the right is a screenshot of the application in use. Gestures idle by the user's side until they indicate intention to speak, at which point hands are raised into gesture space. While talking, the voice analysis module issues gesture signals that create co-verbal emphasis motions, so called *strokes*.

dening the users [18, 17]. However, trying this approach in conjunction with a live video feed of the face, has not been tried before.

Many commercial applications exist that bring users together in a virtual space. The applications that run on conventional hardware mostly rely on simple avatar models, often with a fixed set of animations. In these applications it is common that users either communicate through text or voice audio but lack video. The main strength of these applications is the shared space, which is great for multiplayer games, but subtle or nuanced communication can be a challenge. Regarding virtual reality (VR) based solutions, they are able to display the positions of users hands using tracked VR controllers. The resulting hand gestures are often both accurate and effective for communication, but the users' faces are not visible. The closest commercial solution to the one presented in this paper is from Mibo, which is a company that launched in October 2020. Mibo built an application that combines both video conferencing capabilities and movement in a virtual 3D world.[1] In Mibo the users' avatars are essentially a floating screen on top of a small square torso without hands and legs. The floating screen displays the users entire webcam feed, instead of tracking and cropping to the face. While the web cam feed can potentially show some hand movement, it is not easily visible from all angles. Mibo does not utilize fully animated 3D avatars, and therefore does not include the ability to have human-like animations and gestures within the shared communication space.

3. Approach and Implementation

Our application was implemented in the Unity 3D game engine, using multiple plugins, SDKs and custom code. Unity is a free development platform, that is especially useful for quickly building interactive 3D environments.[2] Furthermore, Unity offers an asset store where a variety of 3rd party plugins can be downloaded and integrated into an application, a feature that really helped us get off the ground.

For networking, an open source networking solution for Unity called Mirror was used.[3] This particular networking solution provides client and server code along with premade Unity game objects that handle for instance player connection, synchronization, packet transport and more. These components allow for a quick setup of a networked multiplayer game.

The avatar is a vital aspect of the application. The avatar model is a custom design and was modeled using the open source 3D modeling and animation tool Blender.[4] The avatar was rigged and animated using a combination of custom and pre-made animations. As mentioned earlier, the users control their avatar using WASD control scheme, made popular by first-person 3D games. The keyboard keys mimic the shape of the directional arrow keys, where ”W” is up, ”S” is down, ”A” is left and ”D” is right. One of the main benefits of this scheme is that users can use their left hand for controlling the movement of the character, leaving their right hand free to use the mouse for looking around, and thus orienting their avatar.

Facial video is captured from the user's webcam, and the video stream processed and cropped by a facial detection script from the OpenCV[5] open source computer vision library. The cropped video stream along with voice audio, captured from the microphone, is fed into an SDK called

[1] https://getmibo.com/

[2] https://unity.com/

[3] https://mirror-networking.com/

[4] https://www.blender.org/

[5] https://opencv.org

Agora.[6] Agora establishes a channel for video communication between users. Each processed frame is streamed to all users and rendered onto the face of each user's respective avatar. The gain and pan of the voice audio playback is programmed to function like 3D spatial audio, placing importance on users gathering together for conversations. Furthermore, the voice audio is the driver for the automated gesture system.

A high level state diagram for the automated gesture system can be seen in Figure 2. When the users wish to speak, they need to press a push-to-talk button and this button has two functions: i) it un-mutes the user's microphone and ii) it activates the gesture system. This might seem trivial but is in fact a truly powerful and intuitive signal to let others know about the intention to speak. When a user activates the gesture system, the avatar immediately raises it's arms into gesture space. Allowing the user to instantly signal to others in the space, a desire to enter the conversation. Our hope is that this will start addressing the turn-taking issue which is common both in contemporary videotelephony applications and many avatar-based solutions.

The talking animation follows the classic co-verbal gesture phases defined by [13] and further specified for Embodied Conversational Agents in [12]. When the user starts talking, the hands immediately go through the *preparation* phase, bringing them into gesture space, where they maintain the so-called *pre-stroke hold*. The system will then generate one or more emphasis motions, or *strokes*, each comprising of a *stroke-start*, *stroke* and a *stroke-end*. A new stroke can interrupt a previous stroke before its stroke-end, or commence once a *post-stroke hold* is reached, starting the stroke cycle again. When the user finishes talking, the hands go through a *retraction* phase, taking them back out of gesture space.

The strokes are activated by a parallel system that analyzes the voice audio. It essentially issues interrupts to the gesture animator, telling it to generate new strokes. This results in the hands moving in a rhythmic manner following the speakers voice. Accompanying the hand movements are subtle head tilts that have a slight chance of activating during strokes. The shape of each gesture is procedurally generated at run-time using a combination of inverse kinematics and key-frame animations, allowing for a diverse series of gestures that don't look repetitive.

4. Conclusions and Future Work

As seen in Figure 1, our approach has resulted in a working communication system that combines avatars and video in a fairly natural fashion. This is ongoing research, and our immediate next step is to conduct new extensive user tests focused on measuring usability, user experience and the effectiveness of the new automatic gesture system. Based on those results, we plan to continue looking into ways to effectively bring features of in-person communication into lightweight communication tools, building on established psycho-social theories.

In terms of immediate improvements to the application, the next steps include implementing a series of new features related to communicative intent monitoring and the non-verbal behavior generation. For instance, we plan to add a system for generating non-verbal listening cues, because clear backchannels from listeners, such as head nods, are important for speakers, and currently those are not animated by the avatars. These behaviors could potentially be triggered by further visual analyses of the listener faces. We also plan to implement more powerful voice analysis and better synchronization with the gesture generation, as currently we are only using the amplitude of the voice to know when a user is speaking. The system has already been built with prosody tracking in mind, which would then directly trigger strokes on detected pitch accents. Finally, we would like to experiment with a keyword detection system, that might prove effective for generating more specific gestures and head movements, such as nodding upon agreement or pointing when objects in the virtual space are referred to.

References

[1] RJR Blair. Facial expressions, their communicatory functions and neuro–cognitive substrates. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):561–572, 2003. 2

[2] Justine Cassell and Hannes Vilhjalmsson. Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. *Autonomous Agents and Multi-Agent Systems*, 2(1):45–64, 1999. 1, 2

[3] Nicole Chovil. Discourse-Oriented Facial Displays in Conversation. *Research on Language and Social Interaction*, 25(1991/1992):163–194, 1991. 2

[4] Frith Chris. Role of facial expressions in social interactions. *Philosophical Transactions B*, 364(1535):3453–3458, 2009. 2

[5] Starkey Duncan. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(Journal Article):161–180, 1974. 2

[6] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, Oct. 1987. 2

[7] Magy Seif El-Nasr, Katherine Isbister, Jeffery Ventrella, Bardia Aghabeigi, Chelsea Hash, Mona Erfani, Jacquelyn Morie, and Leslie Bishko. Body buddies: social signaling through puppeteering. In *International Conference on Virtual and Mixed Reality*, pages 279–288. Springer, 2011. 2

[6] https://www.agora.io

[8] Charles Goodwin. *Conversational Organization: Interaction between speakers and hearers*. Number Book, Whole. Academic Press, New York, 1981. 2

[9] Anna Sigridur Islind, Ulrika Lundh Snis, Tomas Lindroth, Johan Lundin, Katerina Cerna, and Gunnar Steineck. The virtual clinic: two-sided affordances in consultation practice. *Computer supported cooperative work (CSCW)*, 28(3):435–468, 2019. 2

[10] Adam Kendon. *Conducting Interaction: Patterns of behavior in focused encounters*. Studies in International Sociolinguistics. Cambridge University Press, Cambridge, England, 1990. 2

[11] Seungwon Kim, Mark Billinghurst, and Kangsoo Kim. Multimodal interfaces and communication cues for remote collaboration, 2020. 2

[12] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn Thorisson, and Hannes Vilhjalmsson. Towards a Common Framework for Multimodal Generation in ECAs: The Behavior Markup Language. volume 4133, Berlin Heidelberg, 2006. Springer. 4

[13] David McNeill. *Hand and Mind*. Number Book, Whole. The University of Chicago Press, Chicago and London, 1992. 2, 4

[14] Deepika Phutela. The importance of non-verbal communication. *IUP Journal of Soft Skills*, 9(4):43, 2015. 2

[15] Angela Tinwell, Mark Grimshaw, Debbie Abdel Nabi, and Andrew Williams. Facial expression of emotion and perception of the uncanny valley in virtual characters. *Computers in Human Behavior*, 27(2):741–749, 2011. 2

[16] Helena Vallo Hult, Anna Sigridur Islind, and Livia Norström. Reconfiguring professionalism in digital work. *Systems, Signs & Actions*, 12:1–17, 2021. 2

[17] Hannes Vilhjalmsson. Animating Conversation in Online Games. *Lecture Notes in Computer Science*, 3166(International Conference on Entertainment Computing):139–150, 2004. 1, 3

[18] Hannes Högni Vilhjálmsson. Automation of avatar behavior. *J. Tanenbaum, MS el Nasr, and NM, editors, Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters*, pages 255–266, 2014. 3