https://doi.org/10.2352/ISSN.2470-1173.2019.2.ERVR-184

Enhancing Mobile VR Immersion: A Multimodal System of Neural Networks Approach To An IMU Gesture Controller

Juan Niño, Jocelyne Kiss, Geoffrey Edwards, Ernesto Morales, Sherezada Ochoa and Bruno Bernier; Music Faculty Universite Laval and Center for Interdisciplinary Research in Rehabilitation and Social Integration; Quebec City, Canada

Abstract

An effect of immersion is a desirable quality in VR interfaces and its improvement has received a lot of attention in recent years. However, enhancing immersion by eliciting feelings of embodiment and presence through multimodal interactions often requires expensive and bulky dedicated tracking systems. This renders these interactions prohibitable for mobile VR platforms. Our work focused on developing a voice and motion gesture controller with tactile feedback to induce physical sensations and arouse the feeling of embodiment and presence in mobile VR platforms. A mobile virtual reality interface was designed to validate our system through a user experiment. The results indicate that the feelings of presence, embodiment and overall immersion in the mobile VR interface were increased by the use of our system.

Introduction

Immersive computer interfaces are dedicated to provoke the feeling of being in an alternate reality [1]. The feeling of immersion is mainly subjective, its impact and quality vary from one user to the other [2]. This feeling is affected by multiple factors which are difficult to reconstitute in an artificial environment. Shin [3] theorized the feeling of immersion in VR interfaces, he described 4 key components involved in the process of feeling immerse: embodiment, flow, presence and empathy. Embodiment could be defined as an impression of being in a body [4]. Flow describes a state of deep focus and enjoyment, usually triggered by the user's action in the interface [5]. Presence relates to the impression of being inside and exploring a virtual space [6]. Empathy is linked to the user's understanding/identification of virtual context [7]. Empathy can be aroused on the user through the use of virtual 3d avatars, as users can identify with them [8]. Flow experiences can be elicited through careful interface design [9]. Both empathy and flow rely mostly on the psychological stimuli depending on the quality of the interface, ergonomy and playability. Hence, conventional interfaces can arouse both components of immersion when designed correctly. In contrast, embodiment and presence are aroused in the user by a correlation between the physical stimulation, motion and point of view of his biological body and the observed stimulation, motion and point of view of the avatar's body [10]. An appropriate stimulation of embodiment and presence could synergize to enhance the overall immersion in the interface [11]. However, achieving it often requires a dedicated and expensive VR systems with embedded sensors and actuators, such as head and hand motion tracking systems, tactile actuators and spatialized sound, etc. These systems are not compatible with mobile VR displays due to computational power, electric consumption and portability constraints [12]. In the literature, Dias shows the technical restrictions of actual mobile gesture controllers in the context of immersive environment [13]. Considering all these limitations, we developed a VR mobile system with a low-cost gesture controller designed to stimulate the feeling of embodiment and presence, hence enhancing the system's immersive qualities. In this paper we will first present our solution for modeling the multimodal gestures of the user for smooth interaction with our VR interface. Secondly, we describe how the physical and logical components of our mobile VR interface work together to arouse embodiment and presence. Thirdly, we validate our system through a small user experiment and discuss its result. Finally, we provide a conclusion and discuss the perspectives for further research and development.

Method

In our work we focus on eliciting the feelings of embodiment and presence through an inexpensive, portable and non-computationally demanding technique for gesture interaction on mobile VR devices. Our system is designed to be mainly interacted through movement and sound, as sound is useful to produce the feeling of embodiment [14]. Our interface recognizes the gesture (vocal sound and movement) executed by the user and displays a smooth visualization of it during the execution. This visualization is partially controlled by the execution of the gesture to enhance the feeling of presence and embodiment. After the gesture is executed the appropriate interface action is triggered (such as open, close, jump, special attack, etc.). The feeling of presence is also enhanced by providing tactile, visual and auditory feedback to reinforce the correlation between the user's real actions and the avatar's virtual actions.

Multimodal Gestures

Gestures are a form of conveying information through a combination of movement and voice [15]. Recently, speech and motion have been used to communicate commands to a robot [16], and limb pose estimation combined with tactile feedback has been employed for enhancing immersion in videogames [17]. Actually, the quality of perceived sensorial experience has been correlated with the feeling of immersion in multimodal VR interfaces [18]. For our purposes, we will define a gesture as an object containing an array of movement data points and an array of vocal sound data points. Considering that there is a finite number of user actions on most interfaces, one can associate each action with an individual gesture. Hence, most interfaces could be controlled with a finite number of gestures. Furthermore, gestures could be designed to be meaningful for the virtual action they represent, such as analogous motion or blowing, whistling, clapping or vowel sounds.

In this paper we develop a VR interface to validate our system. The interface will allow the user to control a virtual martial art student when producing the five fundamental gestures of shintaido (A,E,I,O,U) [19]. We chose to adopt the fundamental gestures of this martial art because they have a strong association between the movement and vocal sound of these gestures. The motion and sound of the "A" gesture are illustrated in figure 1

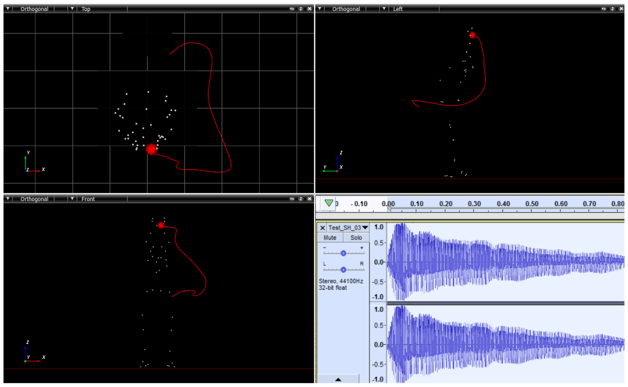


Figure 1. *Example of shintaido gesture; voice and orthogonal views of motion*

Ideal Model Playback

To provide the user with a smooth visualization of the gesture he is producing, we created a set of ideal gestures. Each ideal gesture contains the motion and vocal sound data corresponding to the "best possible" execution of the gesture. Each ideal gesture is played back for the user while he performs the corresponding gesture. By dynamically controlling the playback rate of the gesture to be proportional to the magnitude of the acceleration reported by the controller we intend to provide the user with a degree of control over the movement of the avatar. To create the set of ideal gestures for our interface, we used a motion capture system to record the movement and voice produced by a shintaido expert during each gesture execution. The motion data was exported to control a humanoid avatar. The audio data was exported as a wav file. The execution of a gesture during the motion caption session is shown in figure 2.

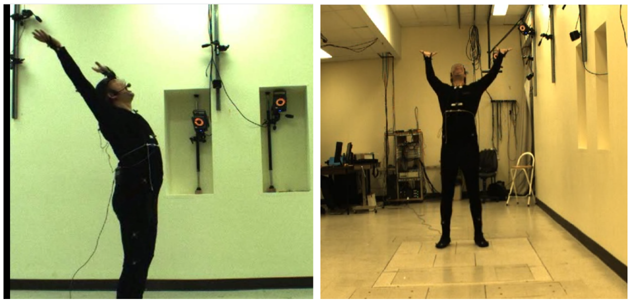


Figure 2. *Execution of "A" gesture during sound and motion capture*

The first section of the system is in charge of voice recognition. The sound associated to each multimodal gesture is detected on every frame. The system keeps updating the result of this classification until the gesture motion starts. The second section of the system provides a smooth premade animation of the gesture classified by the past section. The playback rate is directly proportional to the controller's acceleration to generate an impression of following the movement of the player. The third section of the system captures orientation and acceleration reported by the controller during motion. When the user has executed his movement, the captured gesture is classified and the appropriate response is triggered. The second and third sections of our system run in parallel to provide an impression of direct avatar control while the user executes the gesture. A schematic of the complete gesture controller system can be observed on figure 3.

Interaction system

A google cardboard was used to provide an inexpensive head mount for a smartphone. An inexpensive microphone headset was used to capture the sound associated to the gesture produced by the user. The motion gesture controller was developed based on a 9DOF inertial sensor (BNO055) controlled by an Arduino to wirelessly (bluetooth module HC05) report it's orientation and acceleration while providing tactile feedback (mini vibrator motor) in real time. While inertial sensors are inexpensive, mobile and require no setup, they cannot be configured to provide a position vector precise enough to naturally control an avatar during gameplay. To compensate for this drawback we developed a three section system for multimodal gesture interaction.

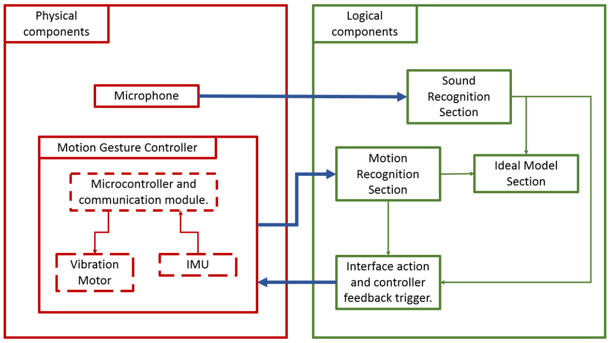

Physical components
Microphone
Motion Gesture Controller
Microcontroller and communication module.
Vibration Motor
IMU
Logical components
Sound Recognition Section
Motion Recognition Section
Ideal Model Section
Interface action and controller feedback trigger.

Figure 3. *Physical and logical components of the interaction system*

Gesture Recognition

Both the voice and movement recognition systems are based on artificial intelligence. Mulitlayer artificial neural networks (ANN) are appropriate for recognizing complex data patterns such as voice and movement while being fast and computationally inexpensive [20]. We implemented our own C# library for ANN construction, architecture and weight import, and classification in Unity3D.

The sound recognition system sections the data received from the microphone in 32 ms frames (512 samples with a 16 kHz microphone sample rate). Then we calculated the first 16 Mel-scale frequency cepstral coefficients of each frame. We chose this method of sound representation as it decreases the amount of data needed to represent sound and has been used in speech recognition before [21]. For our interface, 60 seconds of each vocal sound were performed by the shintaido expert. In addition, 60 seconds of noise and non-relevant sound were recorded. Each frame is labeled during recording as either "non relevant", "a", "e", "i", "o" or "u" and appended to a CSV database file. This database was used in a TensorFlow script to build and train an ANN with an input layer of 16 neurons, a single hidden layer of 11 neurons and an output layer of 6 neurons. The training was stopped when the accuracy reached 98%.

The movement recognition system receives an orientation quaternion and an acceleration 3D vector from the controller at 100 Hz. When each data point is received, the quaternion of orientation is applied to the acceleration vector. We call this vector oriented acceleration and it is relevant to the movement of the controller relative to its initial position and rotation. The magnitude of each oriented acceleration vector is used to detect the start and end of a user gesture. During the execution time of each gesture, the oriented acceleration is stored in a vector array. When the end of the movement is detected, the array's amplitude (acceleration magnitude) and length (time) are normalized. This allow for similar motions performed at different speeds to produce a similar vector array. Similarly to the sound recognition system, examples of each gesture were performed by the shintaido expert and labeled as either "non relevant", "a", "e", "i", "o" or "u" while being appended to a CSV database file. In addition to 20 iterations of each gesture, 100 non-relevant gestures were also recorded. This database was used in a TensorFlow script to build and train an ANN with an input layer of 120 neurons, three hidden layers of 80, 40 and 20 neurons respectively, and an output layer of 6 neurons. The training was stopped when the accuracy reached 98%.

Experimental Procedure

Participants (n=10) aged from twenty-five to forty-five were selected by opportunity sampling. The individuals agreed to participate in the experiment, to answer the questionnaires and to wear the controller and mobile VR device. Although this procedure provided a sample that is not representative of the general population, for a pilot study it was cost effective.

We started the experiment by testing the accuracy of the gesture detection. Each participant was showed a short video demonstrating the vocal sound and motion related to each gesture and then asked to perform it while wearing the motion gesture controller and the wireless microphone headset.

After the first section, the participants were presented with a scene containing the avatar of a martial artist on a google cardboard device. Participants were asked to trigger an action in the interface (displaying the gesture label on the screen) in two different ways. In the first case, our interaction system was deactivated and the participants were asked to trigger an action in the interface by pressing the button corresponding to the label of the gesture (A, E, I, O, U). Additionally, they were asked to perform the selected gesture after pressing the corresponding key. Then, the related ideal gesture was displayed at a constant speed and the action was triggered on the system. In the second case, our interaction system was activated and the participants were asked only to perform the gesture. Our interaction system recognized the gesture and displayed the ideal gesture proportionally to the user's execution. After the gesture is executed the interface action is triggered and the controller vibration motor is shortly activated. After the completion of each session, the participants were presented with a small set of questions to evaluate from 0 to 10 (where 10 is maximum) the level of feeling of presence (Q1), embodiment (Q2) and overall immersion (Q3) that they experienced in the interface [1] as reported in table 1.

Results and Discussion

In this section, we present the results of the three different parts of the experiment. The 10 answers of every participant were averaged for each question and interpreted as an indicative description of the group's perception. In the first section of the experiment, the system was able to correctly classify every gesture produced by different participants, including "non relevant" gestures between experimental tasks. Every gesture was classified with a degree of confidence between 74.6% and 93%. Such a great level of accuracy could be associated with the fact that each gesture was simple and significantly different from one another. Gesture sets that contain more similar gestures might require a more complex ANN architecture and an extended training and validation set.

[1] A simplified demo of the interface is available at https://gesturevrcontroller.000webhostapp.com/

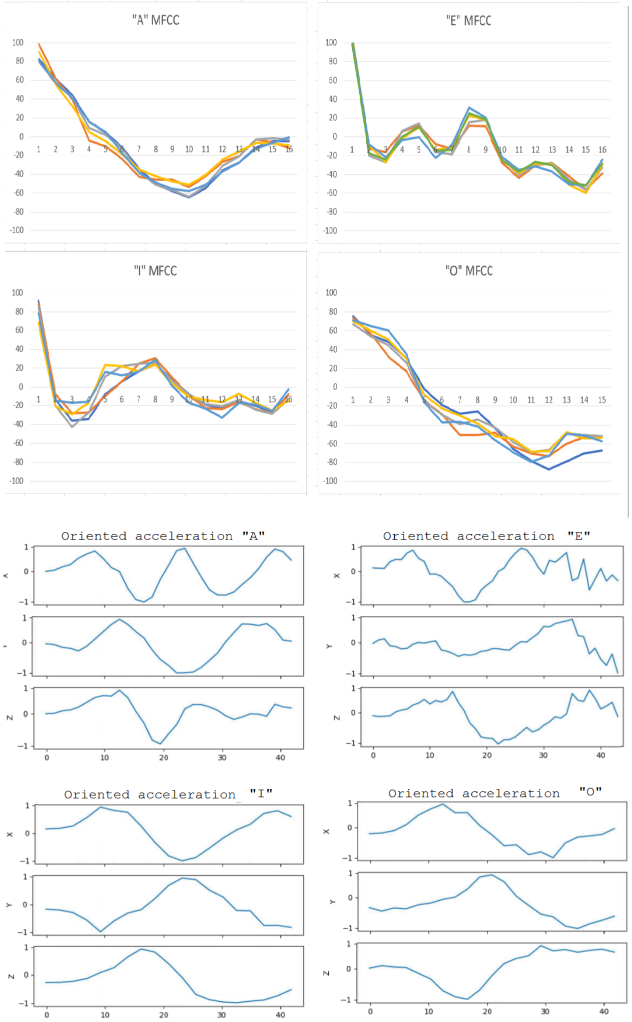


Figure 4. Plotted arrays of MFCC sound data (top) and normalized oriented acceleration (bottom)

Participants answers to questionnaire (averaged)

System state	Q1	Q2	Q3
Deactivated	56%	49%	57%
Activated	78%	81%	89%

For the second section, we directly compare the level of feeling of embodiment, presence and immersion reported by the users when they were asked to perform the task when the interaction system was deactivated or activated (first case and second case). The level of presence reported by the users was significantly higher (78%) when the interaction system was activated compared to when it was not (56%). Similarly, both the level of reported embodiment and overall immersion were higher when the interaction system was activated (81% and 89% respectively) compared to when it was deactivated (49% and 57% respectively).

Although the general validity of this study is limited by the size and duration of the pilot study, the observed results of providing tactile, visual and auditory feedback of the virtual avatar that corresponds to the user's physical actions matches the literature. This encourages further studies following the participants in a mobile VR game to confirm the results presented in this paper.

Conclusion and Perspectives

This paper presented our efforts to induce physical sensations to arouse the feeling of embodiment and presence in mobile VR platforms through a voice and motion gesture controller with tactile feedback. Combining elements that enhance this feelings allowed us to increase the level of immersion of VR interfaces. Our study suggests that the use of multimodal stimulation, vocal sound and motion recognition and a degree of control over the avatar's performance of the ideal gesture could increase the level of immersion on mobile VR interfaces.

The results also indicate that although inertial sensors do not provide a precise position for every point in time, by limiting the motions allowed by the interface, they can be used to control the interface through motion. Despite the fact that our system does not intend to substitute a dedicated VR tracking system, an inexpensive, portable and fast gesture control might be desirable for some mobile VR users even at the expense of complete freedom of movement offered by those system.

An option to improve the system would be to use a higher number of inertial sensors to provide more complex movement options. Furthermore, the system could be generalized to other gestures and be used in other contexts and trained by on-the-go with custom user gestures. A library of different disciplines might be built as an online didactic reference for the instruction and assessment of movement-based knowledge (i.e. dancers, orchestra directors, drummers, violinists, painters, surgeons and athletes, etc.).

We noticed how the participants were excited to use the system and learn new moves. Some were also excited when we mentioned the possibility of using this system to interact with others. Improving the quality of the multi-user interaction will be a target in the further development of our system.

References

[1] M. V. Sanchez-Vives and M. Slater, "From presence to consciousness through virtual reality," *Nature Reviews Neuroscience*, vol. 6, no. 4, p. 332, 2005.

[2] C. D. Reinhard and B. Dervin, "Comparing situated sense-making processes in virtual worlds: Application of Dervin's Sense-Making Methodology to media reception situations," *Convergence*, vol. 18, no. 1, pp. 27–48, 2012.

[3] D. Shin, "Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience?," *Computers in Human Behavior*, vol. 78, pp. 64–73, 2018.

[4] J. O. Bailey, J. N. Bailenson, and D. Casasanto, "When Does Virtual Embodiment Change Our Minds?," *Presence*, vol. 25, no. 3, pp. 222–233, 2016.

[5] C.-I. Teng, "Customization, immersion satisfaction, and online gamer loyalty," *Computers in Human Behavior*, vol. 26, pp. 1547–1554, nov 2010.

[6] R. P. McMahan, C. Lai, and S. K. Pal, "Interaction Fidelity: The Uncanny Valley of Virtual Reality Interactions," pp. 59–70, Springer, Cham, 2016.

[7] K. E. Stavroulia, E. Baka, A. Lanitis, and N. Magnenat-Thalmann, "Designing a virtual environment for teacher training," in *Proceedings of Computer Graphics International 2018 on - CGI 2018*, (New York, New York, USA), pp. 273–282, ACM Press, 2018.

[8] S. Paul, B. Mohler, and S. Paul, "Animated self-avatars in immersive virtual reality for studying body perception and distortions," in *IEEE VR Doctoral Consortium 2015*, pp. 1–3, 2018.

[9] K. Kiili, S. De Freitas, S. Arnab, and T. Lainema, "The design principles for flow experience in educational games," *Procedia Computer Science*, vol. 15, pp. 78–91, 2012.

[10] K. Kilteni, R. Groten, and M. Slater, "The sense of embodiment in virtual reality," *Presence: Teleoperators and Virtual Environments*, vol. 21, no. 4, pp. 373–387, 2012.

[11] C. M. Bachen, P. Hernández-Ramos, C. Raphael, and A. Waldron, "How do presence, flow, and character identification affect players' empathy and interest in learning from a serious computer game?," *Computers in Human Behavior*, vol. 64, pp. 77–87, nov 2016.

[12] Z. Lai, Y. Cui, Z. Wang, and X. Hu, "Immersion on the Edge: A Cooperative Framework for Mobile Immersive Computing," in *Proceedings of the ACM SIGCOMM 2018 Conference on Posters and Demos*, pp. 39–41, ACM, 2018.

[13] P. Dias, L. Afonso, S. Eliseu, and B. S. Santos, "Mobile devices for interaction in immersive virtual environments," in *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, p. 19, ACM, 2018.

[14] B. Spanlang, J.-M. Normand, D. Borland, K. Kilteni, E. Giannopoulos, A. Pomés, M. González-Franco, D. Perez-Marcos, J. Arroyo-Palacios, and X. N. Muncunill, "How to build an embodiment lab: achieving body representation illusions in virtual reality," *Frontiers in Robotics and AI*, vol. 1, p. 9, 2014.

[15] P. Bernardis and M. Gentilucci, "Speech and gesture share the same communication system," *Neuropsychologia*, vol. 44, no. 2, pp. 178–190, 2006.

[16] S. Kalaiarasi, Y. Trivedi, S. Banerjee, N. Chaturvedi, and U. G. Scholar, "Gesture Control Robot using Accelerometer and Voice Control for the Blind," *International Journal of Engineering Science*, vol. 16800, 2018.

[17] Z. Lv, A. Halawani, S. Feng, H. Li, and S. U. Réhman, "Multimodal

Hand and Foot Gesture Interaction for Handheld Devices," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, pp. 1–19, oct 2014.

[18] M. Carrozzino and M. Bergamasco, "Beyond virtual museums: Experiencing immersive virtual reality in real museums," *Journal of Cultural Heritage*, vol. 11, no. 4, pp. 452–458, 2010.

[19] D. Franklin, "The Meaning of Tenshingoso.," *Journal of the U.S. Shintaido Movement Issue*, no. 10, 2001.

[20] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[21] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using MFCC," in *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July*, pp. 28–29, 2012.

Author Biography

Juan Niño is doctorate student at the Music Faculty of Laval University in Quebec City. He is affiliated to the Center for Interdisciplinary Research in Rehabilitation and Social Integration were he continues to developed research about the stimulation of different modalities in computer interfaces for communication, rehabilitation, arts and technology.

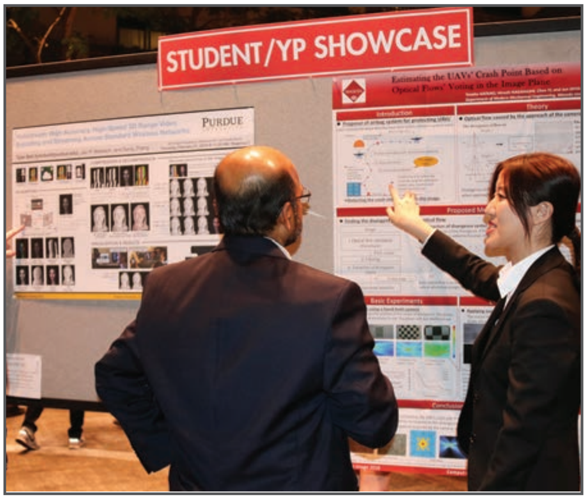
JOIN US AT THE NEXT EI!

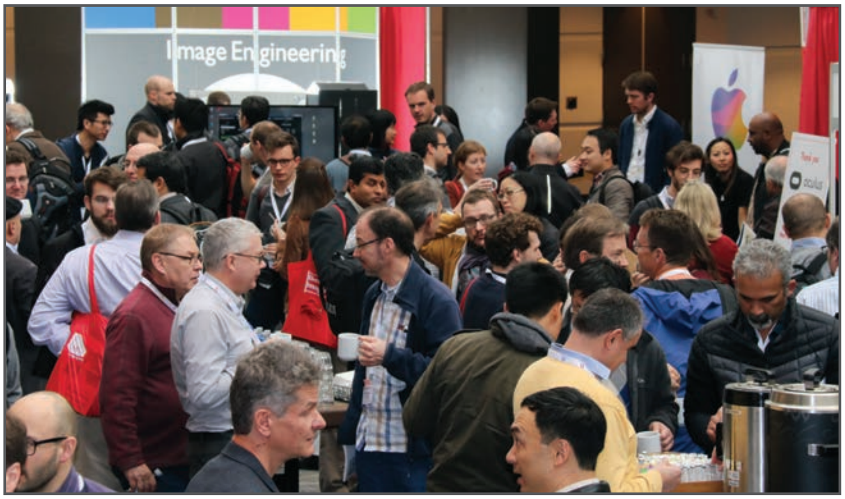
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





AR is a fundamental and transformative interface and platform.
20+ years from my initial working demo to today's AR displays
Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org