FocalSpace: Multimodal Activity Tracking, Synthetic Blur and Adaptive Presentation for Video Conferencing

Lining Yao, Anthony DeVincenzi, Anna Pereira, Hiroshi Ishii
MIT Media Lab
75 Amherst St. E14-348P
Cambridge, MA 02142 USA
+1 87 253 9354
{liningy, anna_p, tonyd, ishii} @media.mit.edu

ABSTRACT

We introduce FocalSpace, a video conferencing system that dynamically recognizes relevant activities and objects through depth sensing and hybrid tracking of multimodal cues, such as voice, gesture, and proximity to surfaces. FocalSpace uses this information to enhance users' focus by diminishing the background through synthetic blur effects. We present scenarios that support the suppression of visual distraction, provide contextual augmentation, and enable privacy in dynamic mobile environments. Our user evaluation indicates increased memory accuracy and user preference for FocalSpace techniques compared to traditional video conferencing.

Categories and Subject Descriptors

H.5.2 [**User Interface**]: Graphic user interface (GUI), Screen design, user-centered design.

Keywords

Diminished reality; video conferencing; synthetic blur; focus; attention; focus and context; depth camera.

1. INTRODUCTION

During face-to-face conversations, without conscious thought, our eyes move in and out of different focal depths, fading out irrelevant background imagery. However, in the case of videoconferencing, this natural behavior is reduced by the inherent constraints of a "flat screen" [9]. The background, which can be distracting and contain unwanted noise, remains in focus.

While gaze and sound have been explored as potential cues [16][17] to prevent visual distractions and enhance focus in video conferencing, we were inspired by artists in cinematography who direct people's attention through Depth of Field (DOF). Previous research has shown that differences between sharp and blurred portions of an image can affect user attention [11]. In FocalSpace (Figure 1), focus is placed on pertinent information and the remainder is blurred, giving users visual indicators for selective attention.

To emphasize pertinent information, we constructed a dynamic layered space that allows participants to perceive different layers, including foreground and background, in different focus.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SUI'13, July 20–21, 2013, Los Angeles, California, USA.
Copyright © ACM 978-1-4503-2141-9/13/07...$15.00.

Figure 1. FocalSpace: active speaker in focus, inactive speakers dimmed 50%, and synthetic blur applied to background.

The focused foreground includes participatory and non-participatory individuals, activities such as sketching, and artifacts, including props and projection walls. Through depth sensing and hybrid tracking of multimodal cues, the system can dynamically identify the foreground and apply the blur filter. These types of visual effects are known as "Diminished Reality" [12].

Our contributions include:

• Application of activity detection to videoconferencing through multimodal tracking of audio, gesture and users' proximity to surfaces.
• Introduction of contextual synthetic blur to steer attention towards relevant content, in the spirit of "Diminished Reality".
• Proposed design scenarios including filtering visual detritus, augmentation with contextual graphics, and privacy in mobile environments.
• User evaluation that indicate increased memory accuracy and preference for FocalSpace techniques over traditional video conferencing.

2. RELATED WORK

2.1 Diminished Reality

Several different approaches of blur and focus have been used as Diminished Reality in data visualization and on screen display to direct people's attention. One example is a geographical information system with 26 layers that fades in and out through blur and transparency [2]. A digital chess tutoring system shows each chess piece in different blur level to indicate strategy step by step [11], and a file browser was developed to show the age of files through continuous blur [11].

2.2 Image Filters for Video Conferencing

"Multiscale communication"[14] was proposed with several video-based communication systems using image filters aimed at increasing engagement. In one example, blur was applied to the entire video, not portions of interest. In addition, image filters have been shown to effectively eliminate details while enhancing others. Blur has, for example, been explored to enhance users' sense of presence and portrayal [13].

2.3 "Focus + Context"

Some related work has demonstrated techniques for segmenting foreground from background in video conferencing, such as zooming [9] and gazing tracking and repositioning the focus [17]. Aforementioned systems treated speakers with gaze directed at them, either one or multiple, as foreground. Kinected Conference [4] introduced voice as a cue to trigger focus on speakers. Our system builds on this previous work, while introducing multiple cues such as gestures, proximity and voice, to track semantic activities beyond "talking heads". To enable the dynamic tracking, "Layered Space Model" is proposed.

2.4 Foreground Sensing Technology

One approach of foreground sensing technology is to pre-capture the background so that image elements that differ are calculated and considered as foreground [5]. However, it requires a pre-calibration process. Face recognition cannot detect other foreground elements beyond human faces. The availability of depth sensing devices [1]for situated environments and their expected imminent integration in mobile devices make the approach scalable and applicable to a wide range of usage scenarios.

3. SYSTEM DESCRIPTION

Our system was adapted to a traditional conference environment with simple readily available components (Figure 2). Three depth cameras are placed in front to optimally capture 3 sides of a meeting table in a conference room. The optional peripheral setup contains satellite-webcams pointing to predefined areas to access high-resolution images of the specified regions.

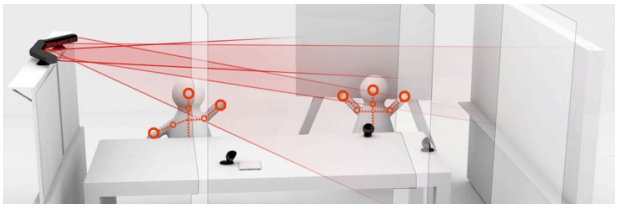

Figure 2. Three depth cameras and microphone arrays in front of a video screen

3.1 Hybrid Tracking of Multimodal Cues

Depth map and human skeleton data given by the Kinect camera are used to track different cues [10]. A microphone array embedded in the Kinect can track the audio cues from the horizontal sound angles of users talking in front of the device. Calibration matches each column of on-screen pixels with the sound angle. If pixels from a person match with the sound angle (10 cm buffer), the person is considered "active speaker" and brought into focus. To prevent unexpected transitions, such as from a natural pause in speech, there is a two-second delay before focus-to-blur transitions.

The system also detects certain gestures, such as "hands up". The hand raising detection is based on skeletal tracking and acceleration rate of the hand joint.

In order to detect how far away an active speaker is from a certain location, proximity cue was tracked using depth maps of the target location and the participants. If the average depth distance between the two was smaller than 20cm, we considered the participant to be approaching or working at a predefined location. Additionally, in order to detect participants' hands approaching an arbitrary object, such as a sketchbook, the object is color marked, thus the spatial location of the marked object can be tracked through depth camera. Proximity between participants' hands and the marked object is tracked in real time.

3.2 Image Filter

The tracked foreground participants are taken in and out of focus computationally. By applying the Fragment GLSL shader [15] to the background pixels twice, horizontally then vertically, Gaussian blur filter is generated for the video. The extent of blur is a user-adjustable parameter, currently allowing blur to be set in steps up to a ten-pixel radius. The system also allows multiple areas of focus and blur at the same time.

4. INTERACTION TECHNIQUES

4.1 Dynamic Layered Space

In order to segment the scene based on activity, we divide the remote space into two discrete layers (Figure 3), the foreground and the background. Conference participants and objects of interest exist within the foreground layer. Further, the foreground layer contains active and inactive foreground: pertinent images, such as the active speaker and active drawing surfaces, are considered active foreground. Active foreground is presented in focus. The remainder is inactive foreground that is focused, but dimmed 50%. The background layer contains the less relevant visual elements behind the foreground. Synthetic blur effects are applied to diminish the salience of the background.

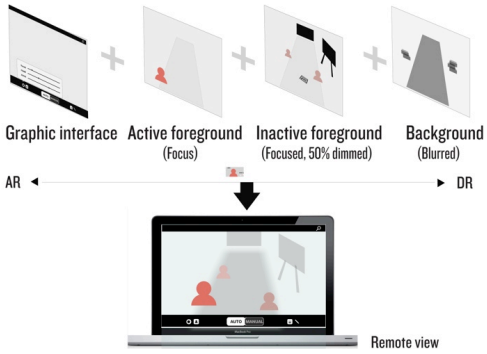

Figure 3. Dynamic Layered Space: Divided into foreground and background layers.

4.2 Multimodal Activity Detection of Semantic Events

Three cues, voice, gesture and proximity, are tracked to determine the active foreground (Figure 4). In addition, remote listeners can manually select the focus.

Audio Cue is used to detect the active speaker. The current speaker is typically the most prominent foreground in group meetings [17][2]. Once the current speakers are detected, they are placed into active foreground and automatically focused on.

Gesture is detected to understand user intent. The ability to track gestures enables the system to behave as a meeting leader, by tracking people raising their hands, and putting the waiting person into focus simultaneously with the active participant.

Since many kinds of communication make use of illustrations and presentation, the system also supports a Proximity Cue. The system tracks the active participants and activates focus on the corresponding surfaces, when users interact with a drawing surface or projection board.

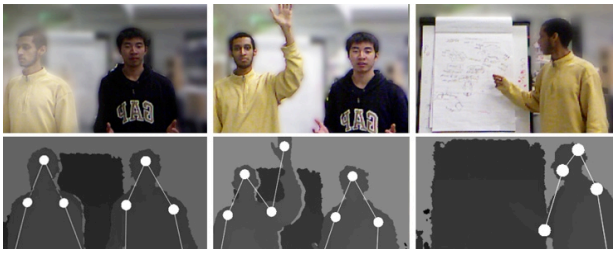

Figure 4. Voice, Gesture and Proximity Cue

In addition, participants who wish to define an area, person, or object of interest at their discretion, can use a user-defined selection mode. Objects are presented as object hyperlinks with common user interface behaviors such as *hover* and *click* states.

5. USER SENARIOS

5.1 Filtering Visual Detritus

Earlier work proposed the idea of utilizing blur effect on the video rather than the "talking heads" area to direct audience's attention [4]. Based on our Layered Space Model, FocalSpace interprets and updates the background in a more dynamic way. For certain scenarios, such as a busy working environment or noisy cafeteria, the background layer refers to the unwanted visual and auditory clutter behind speakers. While during other meetings involving frequent sketching and body movement, the background will exclude both participants and working artifacts, such as sketchbooks and whiteboards. By removing the unwanted background visual distraction we are able to increase communication bandwidth and direct participants' focus on the foreground activity.

5.2 Adaptive Presentation

FocalSpace diminishes unnecessary information and leave the space for adding additional more relevant information to the foreground layer.

Extending the former work of augmentation of "talking heads" with depth sensing [4], FocalSpace provides spatially registered, contextual augmentations of pre-defined objects. With an additional camera, we are able to display a virtual representation of a surface where a remote user is drawing. This technique allows remote users to observe planar surfaces in real time that would otherwise be invisible or illegible. We demonstrate sketching on flip charts, paper and a digital surface (Figure 5).

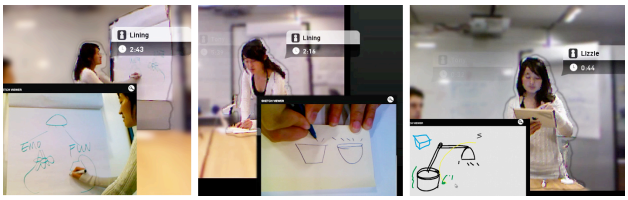

Figure 5: Contextual augmentation of local planar surfaces.

5.3 Privacy for Mobile Environments

We also developed a prototype to explore the FocalSpace concept to more flexible environment using mobile devices (Figure 6). Video stream is captured in front of a situated depth camera and sent over to the phone interface via screen sharing tool [8]. Given the ongoing development of wearable depth sensing technology [7], we envision scenarios where people sit in a coffee shop and send their video streams with blurred background to the remote side using their phones.

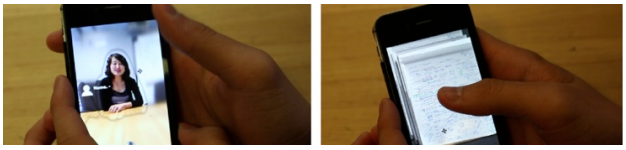

Figure 6: Focus is sent to the mobile interface via network.

6. USER STUDY

As an initial evaluation of FocalSpace we performed a user study of one fundamental aspect of FocalSpace: blur and focus. We investigate how video conferencing with blur and focus compares to video conferencing without, referred to as traditional conferencing. In the study we test two alternative hypotheses:

H_1: FocalSpace increases participant content retention.

H_2: FocalSpace has increased user preference.

As previously highlighted, few user evaluations of video conferencing with diminished and augmented reality exist. We believe that an important first step is to quantify the advantages of FocalSpace and diminished reality. We believe that FocalSpace will increase user focus and therefore memory. Hence, we focus on user memory in our user study. In future work, we plan to follow up with investigations of our previously discussed interaction techniques.

6.1 Methods

In this user study, 16 participants, 8 female, watched two six-minute prerecorded videos emulating video conferences with and without FocalSpace effect. In order to keep the consistency of distraction level and content intensity, prerecorded videos were used instead of real time interactive video conferencing. Both of the videos showed conversations between two users with similar content and generality. Experimental conditions were counterbalanced. The FocalSpace video was recorded with the synthetic blur effect based on audio cues, with the blur level set to three-pixel radius. The traditional video had no special effects.

Content questionnaires evaluated accuracy of user memory. Seven multiple-choice questions were evenly distributed along the timeline of the conversation. The final score was the average percent correct per video. In addition, participants were interviewed on usability and user experience.

Differences in total percent correct score for both FocalSpace and traditional video were analyzed with a two tailed paired t-test. Error bars were reported as standard error of the mean (SEM).

6.2 Results and Discussion

Left of Table 1 shows that participants scored significantly more questions correct watching the FocalSpace conference compared to the traditional conference ($p < 0.01$). Participants retained more content information watching the FocalSpace conference system compared to the traditional system. Participants also skipped significantly ($p = 0.01$) more questions on the traditional video content questionnaire (26) compared to the FocalSpace questionnaire (7). A post-hoc analysis of the FocalSpace content questionnaire showed an increasing trend of questions answered correctly over time ($R^2 = 0.71$). This positive trend implies two key points. First, FocalSpace requires an adjustment period, as users missed more questions in the beginning. It also implies that if the experiment had continued for a longer duration, participants might have performed even higher on the FocalSpace content questionnaire.

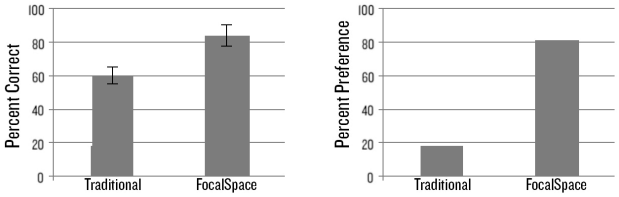

Table 1: (Left) Percent Correct; (Right) Percent Preference

In addition, participants reported their preferred video conferencing system (Right of Table 1). FocalSpace was significantly more preferred compared to a traditional system (p = 0.02). During the interview, seven participants explicitly mentioned that the background movements in the non-blurred video were "distracting." However, though the distractions were the same between videos, no participants mentioned the background movements as distracting in the FocalSpace video. Other comments included a preference towards the FocalSpace because he felt like he was "talking one on one."

Another discussion point of the interview was situations where people might find FocalSpace useful based on the test and previous personal video conferencing experience. Suggested scenarios include: "meetings in a chaotic and distracting environment," "long business meetings with heavy load for concentration," "interviews and lectures that are important and need focus," "larger group video conferencing when active faces are hard to identify," "meetings with participants who are non-native speakers or whose voices are weak," and "meetings with identical faces or voices in the same group." However, blur effect concerns were raised, mostly in casual chatting and complex remote collaboration. For example, one user mentioned that for personal chatting, the blur effect might lead to misunderstandings. People also worried about accidentally blurring important information in a dynamic creative environment. Concerns surround unwanted effects of blurring that could be mitigated in future design parameters of FocalSpace.

In summary, we reject both the null hypotheses. FocalSpace has significantly higher participant memory retention and preference compared to traditional video conferencing.

7. FUTURE WORK

We are interested in exploring the use of gaze tracking for local video conferencing parties to detect focus areas. While our current system focuses on tracking cues and raising users' attention on certain areas, gaze tracking gives a stronger emphasis on areas that users have already been looking at, which can be a complementary approach to FocalSpace.

Our interviews indicated users' preference towards a more flexible environment, such as coffee shops or a home setting, using mobile devices. Additional development of enabling technology [7] is a necessary step in this direction.

In addition, by storing the rich, semantic information collected from both the sensors and user activity, we can begin to build a new type of search-and-review interface. With such an interface, conversation could be categorized and filtered by participant, topic, object, or specific types of interaction. Recalling previous teleconference sessions would allow users to adjust their focal points by manually selecting different active foregrounds, enabling them to review different perspectives on past events.

8. CONCLUSION

This paper presents an interactive video conferencing system called FocalSpace. By incorporating depth imaging into a teleconference system, we have demonstrated a method to effectively reduce perceptual clutter by diminishing unnecessary elements of the environment. We believe that by observing the space we inhabit as a richly layered, semantic object, FocalSpace can be valuable tool for other applications domains beyond video conferencing system.

9. REFERENCES

[1] Chatting, D. J., Galpin, J. S., and Donath, J. S. Presence and portrayal: video for casual home dialogues. In MULTIMEDIA '06. ACM, 395–401.

[2] Colby, G., and Scholl, L. Transparency and blur as selective cues for complex visual information. In SPIE 1460. Image Handling and Reproduction Systems Integration, 114.

[3] Criminisi, A., Cross, G., Blake, A., and Kolmogorov, V. Bilayer Segmentation of Live Video. In CVPR '06. IEEE Computer Society, 53-60.

[4] DeVincenzi, A., Yao, L., Ishii, H., and Raskar, R. Kinected conference: augmenting video imaging with calibrated depth and audio. In CSCW '11. ACM, 621-624

[5] Follmer, S. Raffle, H., Go, J., Ballagas, R., and Ishii, H. Video play: playful interactions in video conferencing for long-distance families with young children. In IDC '10. ACM, 49-58.

[6] Harrison, C., Benko, H., and Wilson, A. D. OmniTouch: wearable multitouch interaction everywhere. In UIST '11. ACM, 441-450.

[7] Hirsch, M., Lanman, D. Holtzman, H., and Raskar, R. BiDi screen: a thin, depth-sensing LCD for 3D interaction using light fields. In SIGGRAPH Asia '09. ACM, 159-165.

[8] iDisplay. Retrieved May 8, 2013, from SHAPE: http://www.getidisplay.com

[9] Jenkin, T., McGeachie, J., Fono, D., and Vertegaal, R. eyeView: focus+context views for large group video conferences. In CHI EA '05. ACM, 1497-1500.

[10] Kinect for Windows SDK. Retrieved May 8, 2013, from Microsoft: http://www.microsoft.com/en-us/kinectforwindows/

[11] Kosara, R., Miksch, S., and Hauser, H. Semantic Depth of Field. In INFOVIS '01. IEEE Computer Society, 97.

[12] Mann, S. "Through the Glass, lightly". IEEE Technology and Society, 31 (3). 10-14.

[13] McCay-Peet, L., Lalmas, M., and Navalpakkam, V. On saliency, affect and focused attention. In CHI '12. ACM, 541-550.

[14] Roussel, N., Gueddana, S. Beyond "Beyond Being There": Towards Multiscale Communication Systems. In MM'07. ACM, 238-246.

[15] OpenGL. Retrieved May 8, 2013 http://www.opengl.org/

[16] Okada, K., Maeda, F., Ichikawaa, Y., and Matsushita, Y. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In CSCW'94. ACM, 385-393.

[17] Vertegaal, R., Weevers, I., and Sohn, C. GAZE-2: an attentive video conferencing system. In CHI EA '02. ACM, 736-737.