Supporting Multitasking in Video Conferencing using Gaze Tracking and On-Screen Activity Detection

Daniel Avrahami[1], **Eveline van Everdingen**[2], **Jennifer Marlow**[1]

[1] FXPAL
Palo Alto, CA 94304 USA
{daniel, marlow}@fxpal.com

[2] Vrije Universiteit Amsterdam
1081 HV Amsterdam, The Netherlands
e.a2.van.everdingen@student.vu.nl

ABSTRACT

The use of videoconferencing in the workplace has been steadily growing. While multitasking during video conferencing is often necessary, it is also viewed as impolite and sometimes unacceptable. One potential contributor to negative attitudes towards such multitasking is the disrupted sense of eye contact that occurs when an individual shifts their gaze away to another screen, for example, in a dual-monitor setup, common in office settings. We present an approach to improve a sense of eye contact over videoconferencing in dual-monitor setups. Our approach uses computer vision and desktop activity detection to dynamically choose the camera with the best view of a user's face. We describe two alternative implementations of our solution (RGB-only, and a combination of RGB and RGB-D cameras). We then describe results from an online experiment that shows the potential of our approach to significantly improve perceptions of a person's politeness and engagement in the meeting.

Author Keywords

Video conferencing; multitasking; gaze tracking; head-pose; intelligent camera switching.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The use of videoconferencing in the workplace has grown significantly over the last decade. This includes the use of dedicated systems, often in dedicated rooms (*c.f.,* [2]) and video conferencing from a personal computer (a desktop or a laptop). Prior work has shown that videoconferencing can be advantageous over audio-only communication ([3, 4]). One challenge for using videoconferencing in the workplace is that participants often need (or want) to engage with other tasks. However, multitasking during videoconferencing is typically regarded as rude or inappropriate. As a result, the need and desire to multitask often leads workers to opt for using an audio-only link, forgoing the benefits of videoconferencing. Since multitasking activities are often related to the meeting itself (for example, looking at a shared document or searching for relevant material), we argue that mitigating the negative perceptions of multitasking in videoconferencing could be beneficial to all parties in the conversation.

In a recent study, Marlow *et al.* [7] found that multitasking during videoconferencing was rated as significantly more polite and acceptable when it took place on the same screen as the videoconferencing than when it took place on a second monitor or on a phone. One explanation for this difference, proposed in [7], is that the ability to see the remote person's eyes, even when they were multitasking, made them appear more engaged. The importance of, and sensitivity to gaze in both face-to-face and video conferencing interaction has been demonstrated many times (*c.f.*, [12]). However, with the price of display technologies dropping, dual-monitor setups are becoming common. Thus, videoconferencing (and multitasking) in such setups needs to be supported.

To improve the communication between participants during videoconferencing, we present a solution that utilizes different cameras for each display. Our solution intelligently switches between the two cameras, based on a combination of head-pose tracking and on-screen events monitoring, in order to provide remote participants with the best head-on view of the user. We describe two implementations: one that relies on two standard RGB cameras, and one that uses one standard RGB camera and one depth camera (RGB-D). Using a controlled online study, we demonstrate the viability of our approach for improving impressions of engagement, politeness, and acceptability of multitasking behavior.

RELATED WORK

Prior work has looked at different methods to infer gaze behavior and adapt information presentation accordingly, both virtually and in collocated settings. Past work has used Kinect-based systems to suggest the best viewpoints from different cameras and angles while continuously displaying a target object on the screen [11] or to identify speakers and camera angles for a group of people having a face-to-face meeting [9, 10].

One impact that multitasking can have on video communication is that it potentially affects gaze. People are

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. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

IUI'16, March 07 - 10, 2016, Sonoma, CA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4137-0/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2856767.2856801

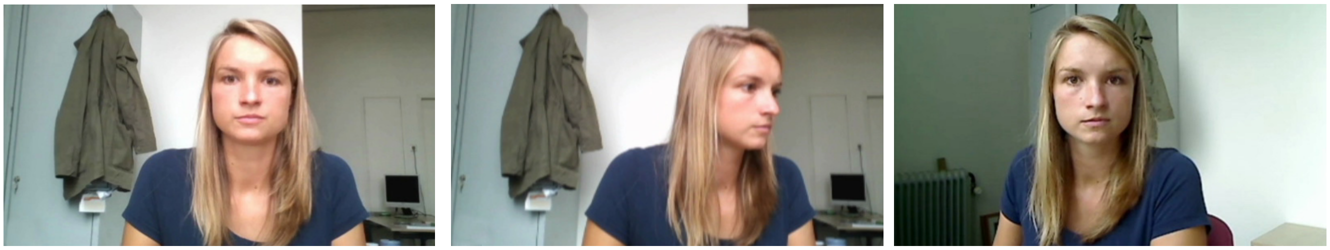

Figure 1. Looking at the primary monitor (left), looking at secondary monitor without camera switching (center), and looking at the secondary monitor with camera switching (right). While the background behind the user changes, the ability to get a head-on view of the face results in significantly higher ratings of politeness and engagement in the videoconference.

fairly accurate at detecting when eye contact/gaze is being directed at them, and this feeling of direct eye contact can build trust [1]. Prior work pertaining to video conferencing has addressed the role of correcting for gaze direction of a single person. The GAZE-2 system, for example, used multiple webcams on a single screen [12]. Other systems used a single camera and synthetically corrected a remote person's gaze (e.g. [5, 6]). Additional work on inferring engagement in conversation with a conversational agent also incorporated gaze detection mechanisms [8].

However, using video conferencing while interacting with two monitors is also a common scenario in distributed collaboration contexts [7]. In those cases, the range of possible head poses is often too great for synthetic gaze correction. Our emphasis is on following a user's gaze across different monitors and cameras rather than adjusting a single, relatively static eye-gaze view. Additionally, our solution uses information about the user's interaction with the computer. This allows increased reliability when reliance on computer-vision is insufficient.

IMPLEMENTATION

To improve video-conferencing experience in a dual-monitor setting, we implemented a solution that uses head-pose tracking and on-screen activity recognition for automatic camera selection. We make the assumption that each monitor already is, or can be cheaply, fitted with its own webcam. One important aspect of our solution is that while it aims to support multitasking it intentionally does not hide it (the background changes behind the user when the camera view changes, making multitasking clear).

The solution operation is illustrated in Figure 2. The On-Screen Activity Monitor continuously listens for mouse, keyboard and focus-change events. When started, the *Current Camera* is the camera of the primary monitor. On every iteration (typically every frame), the head-pose tracking module determines which monitor the user is looking towards (as illustrated in Figures 1 and 3). Possible return values are: Looking at the same monitor as the current camera (no need to switch), looking at the other monitor (recommend a switch), looking at neither monitors (no need to switch), and uncertain (*low confidence*).

If the result from the head-pose tracking module cannot determine which of the two monitors the user is looking at, the solution examines the location of the latest events recorded by the On-Screen Activity Monitor. Return values can be: the same monitor as the current camera (no need to switch) or the other monitor (recommend a switch).

Based on the return value, the solution will either attempt to switch to the other camera and set it as the Current Camera (see *Transition Smoothing* below), or stay with the same camera. Video from the current camera is then passed for transmission over the network to the remote videoconferencing participant.

Head-Pose Tracking Module

We created two alternative implementation of the head-tracking module that we describe next.

RGB-Only Solution

Our first implementation relies on two RGB cameras, one for each monitor. In this first implementation, face-detection on frames from each camera is performed using a Haar-Cascade classifier. If a face is detected in only one of the cameras, we assume that the user is looking towards that camera's monitor. If no faces are detected by either camera (or if faces of multiple people are detected), the module does not recommend switching. However, if a face is detected in both cameras, we estimate the head-pose in each camera using Haar-Cascade classifiers trained on facial features (specifically, eyes and noses). We use the geometric relationship between detected eyes and nose within the face's bounds to estimate which camera has a view that is closest to a head-on view of the face. If no clear winner is determined, the module marks its result as *low confidence*. The benefit of this solution is that it is cheaper and many existing monitors now come with a webcam built in.

RGB-D + RGB Solution

Our second implementation performs head-pose tracking using a single RGB-D camera, and only uses the second RGB camera to be able to send two views of the user to the videoconferencing application. The benefit of this solution is that it is more accurate and, thanks to using depth information, less susceptible to issues of illumination or problems with recognizing faces with glasses. We implemented this version in C# using the Intel RealSense F200 depth camera.

To use this solution, the system must first be calibrated; calibration allows the system to establish the geometric relationship between head position and orientation and the different monitors. In the calibration stage, the user is asked to look at, and click on circular targets shown at the corners of each monitor. The system records the head-position in 3D space and Eular angles of the head for each target. These are then used at runtime to estimate the user's gaze direction relative to each monitor. For example, for each monitor we compute a horizontal offset based on the head's distance from the camera, horizontal distance from the camera's center, and the yaw value as follows:

$$H_{Monitor} = X_{Head} + Z_{Head} \times \tan(Yaw_{Head})$$

The values are normalized such that a value of 0.0 corresponds to looking at the left edge of a monitor and 1.0 to looking at the right edge of the monitor. A negative value indicates that the user's gaze is to the left of the monitor, a value greater than 1 indicates that the user's gaze is to the right of the monitor. When calculated for both monitors, it is possible that the value for *both monitors* will be within the 0-1 range (or close to it), for example, when the user is looking towards an area close to both screens (see Figure 3c). In those cases, the module marks its result as *low confidence* and the system will rely on the On-Screen Activity Monitor.

Transition Smoothing

One issue that we discovered quickly is that frequent head movements or noise in the vision system can result in the video constantly switching between cameras, leading to poor user experience. We thus introduced a two-stage smoothing mechanism when deciding whether to switch between cameras. First, we only switch to a camera if the user looked towards its monitor for longer than a preset threshold (in our case, 750 milliseconds). Second, we do not switch away from a camera if it was recently switched to; in other words, we remain on a camera for at least a preset threshold (in this case, 1 second). While these thresholds can be adjusted, we found the introduction of this smoothing mechanism to greatly improve the usability of our solution.

One area for exploration in future work is to use knowledge about the application interacted with to predict the duration of the interaction. For example, an instant message window combined with keyboard events might suggest a longer gaze than, say, the appearance of an incoming-email notification.

EVALUATION

To test whether our approach is able to improve perceptions of a videoconference participant's behavior, we conducted a between-subjects online experiment following the paradigm from Marlow *et al.* [7]. In the study, participants watched one of two 1-minute clips of a videoconference meeting between two people (Person A and Person B) and then answered a series of questions about what they saw and heard. Instructions read, "*On the next page is a clip from a recorded videoconference business meeting between two coworkers, Person A and Person B, where they discussed an advertising campaign. You are asked to watch the video and answer a few questions about what you saw and heard.*"

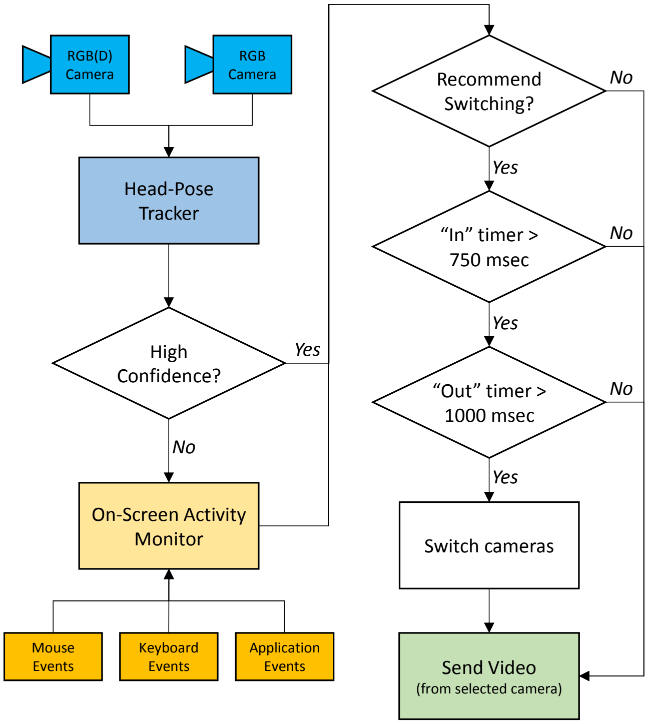

Figure 2. Decision flow of our system. Our system chooses a view of the user to send to remote participants based on head-pose tracking and on-screen activity monitoring.

In the clips, Person A describes three potential locations for an advertising campaign. Person B, who has a dual-monitor setup, listens, occasionally responding with short utterances such as "uh-huh." During the conversation, Person B shifts her gaze between her two monitors (from the primary to the secondary monitor and then back).

The study was a between-subjects design in which participants were randomly assigned to one of two

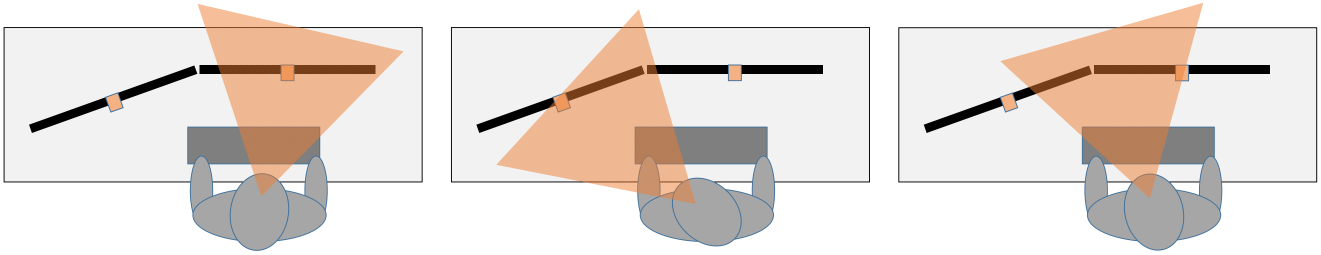

Figure 3. Looking twards monitor 1 (left), towards monitor 2 (middle) and towards the left edge of monitor 1 (right). In the last case, the head-pose tracker has low confidence, and the system relies instead on on-screen activity monitoring.

conditions: the *Dynamic View* condition, in which the view changed between the two cameras when Person B shifted her gaze to the secondary monitor, and the *Static View* condition, in which only the view from a single camera was used.

The study was conducted on the Crowdflower crowdsourcing platform. U.S.-based participants received 40 cents for their participation.

Measures

After watching the video clip, participants were asked to rate a series of statements on a 5-point Likert scale (with 1=completely disagree and 5=completely agree). These statements pertained to the Politeness, Acceptability, and Obviousness of the multitasking by Person B, and rating of Person B's engagement in the meeting.

Participants

138 unique individuals participated in the study. We excluded 32 participants who failed a comprehension question about the video and two participants whose completion time was outside the norm. Our final dataset included 104 respondents (55% women), 53 in the Dynamic View condition and 51 in the Static View condition.

RESULTS

We conducted an ANOVA with Condition (*Dynamic View* vs. *Static View*) as the independent variable and the align-ranked ratings [13] of Politeness, Acceptability of Person B's multitasking, and ratings of Person B's engagement in the meeting. We included Gender and Age as covariates.

The analysis shows that while multitasking was equally obvious to participants in both conditions (*M*=3.62 for Dynamic vs. *M*=3.73 for Static, *F*(1,102)=.03, *p*=.84), our camera-switching approach resulted in a significant improvement in ratings of perceived politeness, and acceptability of the multitasking, and higher ratings of engagement in the meeting. As shown in Figure 4, Person B's behavior in the Dynamic View condition was rated as significantly more Polite (*M*=3.86) than it was in the Static View condition (*M*=3.25) (*F*(1,102)=9.78, *p*<.01). Person B's behavior was also rated as significantly more Acceptable in the Dynamic View condition (*M*=3.75) than the Static View condition (*M*=3.23) (*F*(1,102)=6.63, *p*=.01) and Person B was seen as significantly more Engaged in the Meeting in the Dynamic View condition (*M*=3.56) than the Static View condition (*M*=3.07) (*F*(1,102)=5.99, *p*<.02).

We found supporting evidence looking at participants' open-ended responses to the question "Please describe in a few sentences what you saw person B doing." While many participants in the Static View condition described Person B as being distracted, unfocused, or preoccupied (e.g., *Person B was not paying attention to the person talking*), none of the responses in the Dynamic View condition contained negative assessments about Person B's behavior. Additionally, eight participants in the Static View condition referred to Person B's gaze direction, e.g. "*Listening to Person A and then looking off screen*" or "*looking off into another area.*" In contrast, 10 participants in the Dynamic View condition who referred to Person B changing her gaze were focused on the screens and camera angles (e.g., "*switching between two different monitors/screen angles*" and "*Person B was listening, and then looked into a different monitor before switching back*"). This suggests a better comprehension of the multitasking happening on a dual-monitor setup.

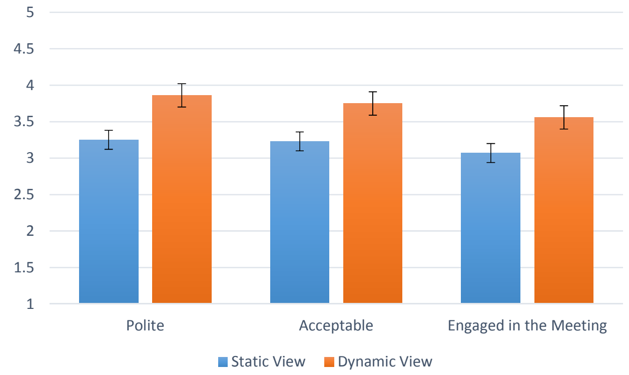

Figure 4. The effect of our system on ratings of Politeness, Acceptability, and Engagement in the Meeting.

DISCUSSION AND FUTURE WORK

In this work, we presented a video-conferencing approach designed for a dual-monitor setup, common in office environments, equipped with dual cameras. Our solution intelligently switches between the two cameras, based on a combination of head-pose tracking and on-screen events monitoring, in order to provide remote participants with the best head-on view of the user.

We evaluated our approach using an online experiment that shows that dynamically switching cameras during multitasking results in significantly higher ratings of perceived politeness, acceptability, and engagement in the meeting. Our solution does not hide the fact that multitasking is taking place – a user is seen turning away for a brief moment before the camera view switches, and the background behind the user changes. Our results thus improve perceptions of behavior, without obscuring it. One limitation of our experiment is that participants were placed in the role of non-participating meeting observer watching two strangers converse. It is important to note that, particularly in multi-person videoconferencing meetings, being a passive observer in a meeting is not uncommon.

In future work, we intend to examine the use of this approach across devices, such as switching between the cameras of a laptop and smartphone. Another area for future exploration is for improving the camera-switching decision process with additional semantic knowledge of the user's activity. For example, the type of application that the user is looking at, or interacting with on the second screen could be used to predict how long the interaction is likely to be, and decide whether to transition to the corresponding camera. Finally, detecting off-screen activities such as writing or sketching on a whiteboard and providing a best view is an interesting area for future work.

REFERENCES

1. Ernst Bekkering and J.P. Shim. 2006. Trust in Videoconferencing. *Communications of the ACM* 49, 7: 103–107.
2. Cisco Project Workplace: http://www.cisco.com/c/dam/assets/sol/tp/project-workplace
3. Owen Daly-Jones, Andrew Monk, and Leon Watts. 1998. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *International Journal of Human-Computer Studies* 49, 1: 21–58.
4. Wei Dong and Wai-Tat Fu. 2012. One piece at a time: why video-based communication is better for negotiation and conflict resolution. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM, 167–176.
5. Jim Gemmell, Kentaro Toyama, C. Lawrence Zitnick, Thomas Kang, and Steven Seitz. 2000. Gaze awareness for video-conferencing: A software approach. *IEEE Multimedia* 7, 4: 26–35.
6. Dominik Giger, Jean-Charles Bazin, Claudia Kuster, Tiberiu Popa, and Markus Gross. 2014. Gaze correction with a single webcam. *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, IEEE, 1–6.
7. Jennifer Marlow, Eveline van Everdingen, and Daniel Avrahami. Taking Notes or Playing Games? Understanding Multitasking in Video Communication. *Proceedings of CSCW 2016*, ACM, to appear.
8. Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. *Proceedings of the 15th international conference on Intelligent User Interfaces*, ACM, 139–148.
9. Abhishek Ranjan, Rorik Henrikson, Jeremy Birnholtz, Ravin Balakrishnan, and Dana Lee. 2010. Automatic camera control using unobtrusive vision and audio tracking. *Proceedings of Graphics Interface 2010*, Canadian Information Processing Society, 47-54.
10. Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. 2014. Overt or Subtle? Supporting Group Conversations with Automatically Targeted Directives. *Proceedings of the 19th International Conference on Intelligent User Interfaces,* ACM, 225-234.
11. Fumiharu Tomiyasu and Kenji Mase. 2015. Human-Machine Cooperative Viewing System for Wide-angle Multi-view Videos. *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion,* ACM, 85-88.
12. Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction. *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 521–528.
13. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 143–146.