Avoiding Nostril-cam and Postage-stamp People in Mobile Video Conferences

Mary Baker, Ramin Samadani, Ian Robinson
HP Labs
Palo Alto, CA 94304
{mary.baker,ramin.samadani, ian.robinson}@hp.com

Mehmet Yilmaz
Stanford University
Stanford, CA 94305
myilmaz@stanford.edu

Kean Wong, Matthew Hornyak
Hewlett-Packard
Sunnyvale, CA 94085
{kean.wong,matthew.hornyak}@palm.com

ABSTRACT
We would like to provide high-quality video conferencing so that people can communicate comfortably with each other anywhere, anytime. This is not a new goal, and there are now several applications such as Skype™ and FaceTime™ on mobile platforms that bring us closer to achieving anywhere, anytime video communications. Alas, these mobile conferences are often of poor quality due to the many challenges presented by mobile devices, such as constrained networks, limited processing power, small displays, and uncontrolled view angles and lighting. These challenges mean that simply porting existing desktop video conferencing solutions to portable devices does not provide the best user experience. Fortunately, these mobile devices also have many advantages that we can exploit to enable better-quality portable video conferences. In this paper we describe how we exploit the devices' mobility and embedded sensors to detect and fix two problems that are often ignored but that adversely affect user experience: bad view angles and too-tiny views of people and content, especially in multi-party conferences. Please note that what we describe is very much an experimental prototype and not a finished project.

Categories and Subject Descriptors
H.4.3 [**Communications Applications**]: Computer conferencing, teleconferencing, and videoconferencing; H.5.1 [**Multimedia Information Systems**]: Video; H.5.2 [**User Interfaces**].

General Terms
Design, Experimentation, Human Factors.

Keywords
Mobile applications.

1. INTRODUCTION
Video conferencing on different kinds of platforms still exhibits a wide range of quality and user experiences. At the high end of the quality spectrum are rooms specifically designed for video conferences, such as HP's Halo (recently sold to Polycom) and Cisco's Telepresence 3000. Most people who have used these rooms know that the experience can be very close to having the remote participants sitting across the table from you, even if they are on the other side of the world. Achieving this requires high quality video and audio (with excellent echo cancellation, for instance), but also requires control of the physical room itself. The walls and tables in all the rooms match each other in color and style so that all participants appear to be co-located. Table legs are placed so that you can only pull up a chair in places where the lighting and camera angle and microphone are well-positioned. The screens are placed so that the remote participants appear to be the right size for their apparent distance from you and appear to be sitting around the other end of your table. The experience can be good enough for long, but comfortable and productive conferences. Anyone who has experienced video conferencing of this sort may be very disappointed by conferences on other less controlled platforms. Unfortunately, these rooms are expensive and not available in most places. Where they are physically available they tend to be fought over and are rarely free when you need them.

At the other end of the quality spectrum we find mobile video conferencing. The quality issues are due in part to the constraints in processing power, networks, and display size of mobile devices, but are also due to having very little control over the physical environment. The lighting on people's faces may not be good; their backgrounds (both audio and visual) may be chaotic; the view angle of their faces may be odd; and in more than two-party conferences their faces may be too tiny and unnaturally arranged in boxes tiling the display. It's hard for us to know whether we appear and sound to the remote participants as we would like, and certainly they often don't look or sound to us as we (or they) might like. Some of these problems are also found in desktop video conferences, and we have addressed them there with techniques such as automatic relighting of faces, automatic cropping and framing to use more of the display on faces rather than backgrounds, and background substitution to create a sense of co-location. Alas, mobile devices tend to suffer from even smaller displays and more chaotic environments than desktops. Ideally we want the convenience of anywhere/anytime video conferencing that mobile devices provide but with a quality approaching, as much as possible, that of the carefully controlled video conferencing rooms.

2. THE PROBLEMS
There are many user experience problems for mobile video conferencing, and in this paper we address two that are not often considered, but that make a difference in the pleasantness and

MobiHeld'11, October 23, 2011, Cascais, Portugal.

Figure 1a: a good view angle.

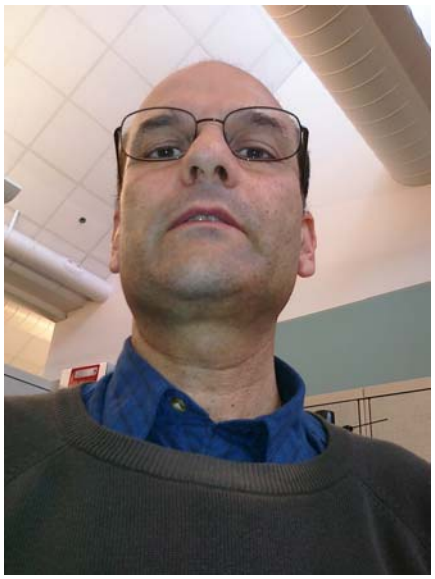

Figure 1b: a bad view angle.

effectiveness of the conference. The first problem is bad view angles of conference participants, and the second is tiny displays of participants and documents (or other content) in conferences that involve more than two or three people.

2.1 Bad View Angles

Users often hold portable devices in their hands or put them on tables, which means that the video camera orientation is not constrained or fixed, and this can result in a poor angle of view for video capture. Commonly, users position the phones too low, even if they are propped on a table or placed on a stand, resulting in a distorted and unappealing upward view of nostrils and chin, or what we call "nostril cam." We see these participants looking down their noses at us, with chins tucked toward necks so that even the slimmest participants have multiple chins (Figure 1). Instead, we would like to see people at their best and be assured others see us from a good view angle as well.

2.2 Postage Stamp People

For a comfortable user experience in video conferencing, it is important to see the other participants well enough to judge mood and expression, and to have them arranged such that they appear to maintain their positions with respect to each other during the conference [2][6]. For instance, on a large enough display, the local user might always see participant Alice next to and to the left of participant Bob. This allows the local user to participate with the same set of expectations regarding positioning as if he were at a physical table with Alice and Bob physically present. Likewise, the user should have a good view and comfortable sense of the position of any document or other shared artifact being viewed in the video conference. Portable devices such as smart phones and even tablets (depending on the number of participants) do not have large enough displays to render multiple remote participants and shared artifacts at sufficient size. To have a large enough image to judge mood and facial expression, we might only be able to view one remote participant at a time on a phone. Instead, we would like to be able to view other participants and documents with sufficient size and a comfortable sense of their positions.

3. OUR CURRENT SOLUTIONS

While the constraints of mobile devices lead to the user experience problems we describe above, we can also exploit the devices' features to help solve these problems. In particular, our current solutions exploit the mobility of the portable devices and their embedded sensors.

3.1 Bad View Angle Detection and Feedback

There are two parts to solving the bad view angle problem: the first is to *detect* the bad angle and the second is to *fix* it.

Detecting: We determine the angle of the phone dynamically during the conference using sensors that are already found in many smart phones and tablets. For instance, three-axis accelerometers are commonly used to sense orientation of the display screen (portrait or landscape), but in our prototype application we use them for a different purpose. The accelerometers sense gravity direction and the components of "gravity acceleration" in the local coordinates of the accelerometer, and this provides view angle information to answer the question: how close to the ideal angle is the device being held? For an upright user, the device is ideally held vertically in front of the user's face, as if he were talking face-to-face with the remote participants. For comfort, though, we tolerate some small angle from the vertical, for example 15 degrees from vertical (with respect to Earth's gravity). In a coordinate system where x and y are in the plane of the device, and z is the coordinate toward the viewer, we use the magnitude of the z component to detect the view angle. The x and y components are useful to ensure the display is not rotated.

Note that this current solution fails if the person holding the device is not upright. If the user is lying down or leaning back in a chair while video conferencing, a good view angle will be detected as a bad one. If front-facing depth cameras become available on mobile platforms, we can improve on our technique by sensing the relative angle between the mobile device and the person's face regardless of the person's position.

Fixing: The second part of our solution is to fix the angle at which the device is held. We chose the approach of giving subtle

visual feedback that has the end effect of causing the user to change how he is holding the device. There are many ways to provide this feedback, and our goal is to do so in a way that is not intrusive but naturally causes the user to fix the angle of the phone. Our current method is to reduce the contrast or color saturation of the incoming video in a manner proportional to the wrong view angle. The participant naturally adjusts the device to a good angle because doing so restores the contrast or color. Since there are times when a bad angle is intentional or cannot be helped (e.g. when the user is lying down or walking while conferencing), it is important to have an easy way to disable the feedback, although we do not yet do this in our prototype.

We have previous successful experience, as evidenced by a formal user study [4], with providing such visual feedback to users in a desktop environment (temporarily displayed self-images blended with the remote view [7] that encourage users to stay within the frame of their desktop cameras). We have not yet performed a user study for the new approach to mobile view angle feedback, but informal tests on unsuspecting colleagues have been encouraging (please see below in the paper).

Instead of reducing contrast or color, we could permanently or temporarily display a small mirror window (self-image) to show the user what he looks like to the remote party. Unfortunately, our user study on the desktop informed us that many users are distracted by images of themselves, so showing the window by default is not a good general solution. Temporarily blending in the window worked well on the desktop, but on a very small display it can be hard for a user to see the nostril-cam problem unless we use almost the entire display for the blended image.

Another approach for detecting and solving the view angle problem is to use sophisticated computer vision and image warping techniques to rewarp, in software, the bad view into a better one. The computer vision approach has been previously attempted [3] for improving conferencing gaze-awareness. However, the authors of that paper report that the computer vision module was a bottleneck, being slow and inaccurate. For our application of correcting for bad view angle, an additional very difficult problem would be filling in occluded portions of the image during rewarping. For instance, a nose can occlude part of a person's face, and people are very sensitive to "mistakes" in rendering human faces. Our current approach to view angle detection, on the other hand, requires only simple processing of accelerometer readings. Our approach to providing user feedback requires only alpha blending, a simple operation available on GPU processors of today's cell phones and tablets.

Figure 2: even with a stand, the view angle of this guy isn't going to be good.

The biggest problem with our solution to bad view angles is that people's arms get tired. While the approach seems to work well for short conferences, holding even a light mobile device out in front of you for long time periods causes arm fatigue. The common solution to this is to use a phone stand such as the amusingly named iPlunge (http://www.worldwide-fred.com/iplunge.htm). Unfortunately, this requires a horizontal surface like a table, which is not always available, and unless the surface is high enough, the angle is still bad (Figure 2). Other solutions we have begun exploring involve "stick anywhere" backings for mobile devices so they can be handily attached to any nearby vertical surface, but we are a long way from getting this to work.

3.2 Better Sizing and Stable Positions

To solve the problem of postage-stamp-sized people in mobile multi-party conferences, we use the mobile device as a "window" into a larger virtual space and can exploit accelerometer, compass, gyroscope, and touch sensor data to navigate the space. Consider participant "Mary" in a multi-party conference. Her device receives video streams over the network, each of which provides video to display one or more remote participants. To provide the illusion that the remote participants are arranged in front of and facing Mary, as they would be if sitting across a physical table, we choose or calculate their arrangement and then use Mary's device as a window onto that arrangement. A similar process takes place for all other mobile participants. Assuming we have Mitch, David and Bruce as remote participants for Mary, we can arrange them in that order from left to right in a virtual space. If Mary holds her mobile display up and pans it from left to right (Figure 3) or rotates it from left to right around a vertical axis, she will see Mitch, then David and then Bruce as if they were in front of her but seen through the mobile display. We could also tilt the phone slightly to one side or another to move between participants. (We have implemented the panning and tilting options in our prototype.) Slides or other shared artifacts can similarly be positioned in the virtual space. Note that the degree of panning, rotating or tilting required to move through the virtual space is adjustable; a small movement can lead to a larger change of view. This is good, because requiring a large movement can result in bad view angles.

Using our solution also gives us an important advantage in the face of constrained networks and devices, since we can limit the number of simultaneous video streams that must be received and decoded by the device to just the visible participants and perhaps their neighbors. We require all the audio streams, but can be more selective about the video streams, which require more bandwidth and processing.

Panning, turning and tilting use accelerometer, gyroscope and compass data, but we can use other kinds of sensor input as well. With a touch interface, we can scroll the display left or right to move to different remote users or slides or other shared artifacts. We could instead use a rear-facing camera to detect device motion (as the scene viewed by the camera changes) to augment or replace the gyroscopic or accelerometer inputs, or we could use the front-facing camera to detect the user's head position in the camera's field of view and use changes in head position to augment or replace the gyroscopic or accelerometer inputs.

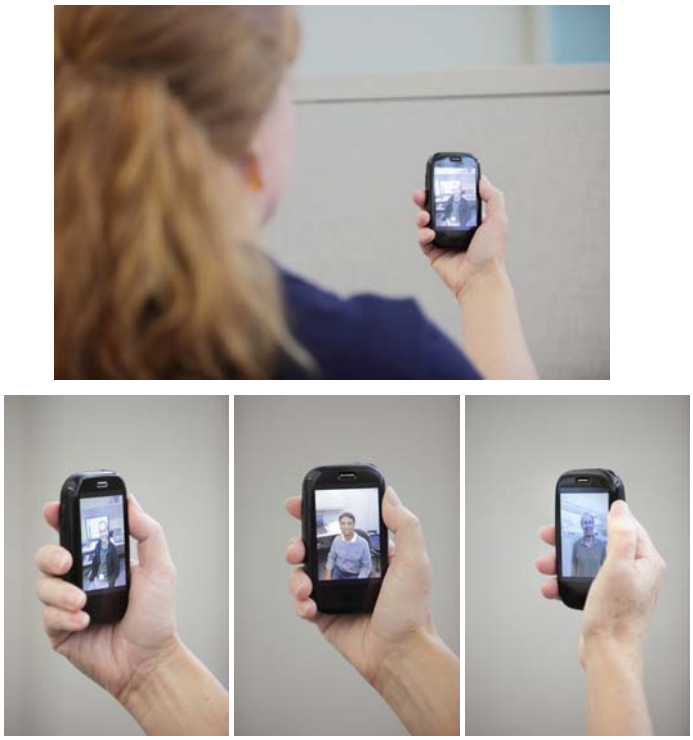

Figure 3: using the panning option to change between views.

Without enough display space to provide high-quality images of all the remote participants at once, the local user may not have enough information about which way to turn to view the current active speaker. Our current solution to this problem uses spatial audio cues through head phones (the voices of people positioned further to the left come proportionally more from the left ear phone, and vice versa). While this seems to be effective, the lack of visual positional orientation could prove disorienting for some people. In the future, we could also provide a small strip of thumbnail images at the top of the display (Figure 4), arranged in order of the remote participants, with the currently viewed participant highlighted (and also, perhaps, the active speaker if different). This provides the local user a quick understanding of the relative position of the person he is currently viewing, but he can also use the thumbnail strip to pan or scroll the display, or to touch a picture to move to the image of a particular remote participant. We do not automatically switch to the stream with the active speaker, since it proves disconcerting to users to lose control over whose face they choose to view.

4. SOME EVIDENCE FOR THE SOLUTIONS

As of this draft, the Palm phones we are using do not have front-facing cameras, and the phones and prototype tablets do not have gyroscopes or compasses, so we have been testing our ideas in advance of these expected features. To experiment with bad view angle detection and feedback, we have implemented an application on the Palm Pre Plus that proportionally *washes out* the video (reduces contrast) as the angle of the phone is tilted away from the vertical. Through this test we have determined that the precision of the accelerometer data is sufficient to provide responsive, smooth feedback. Informal testing with over ten users shows that the feedback is intuitive, since all but one user quickly adjusted the position of the phone to restore contrast to the video.

We have implemented a combination of demos to explore viewing a virtual space of participants through the mobile device. Using the touch interface, we can scroll left or right through images of participants arranged in a virtual space. Using the accelerometer, moving the phone quickly to the right or left pans from one image to the next in the indicated direction. Because we currently use only the accelerometer for this, it requires too much of a jerky motion to be a relaxing method of moving between images, but fusing the accelerometer data with gyroscopic and

Figure 4: mock-up of a possible multi-party UI.

compass data is likely to help. Tilting the phone to the right or left scrolls across the images to the left or right until the user returns the phone to a neutral position. In informal user testing, the tilt interface seems to be as easy a way to scroll through images as the touch interface, but it must be lightly applied so that it does not contribute to our other problem: bad view angles. We believe that once we have the required sensors available, panning or rotating the phone or using the touch interface will probably beat the tilt, but this must be verified through user testing.

5. COMPETITIVE APPROACHES

Video conferencing has existed for several decades but has suffered from repeated failure to be adopted [1]. Many technological improvements (including higher processing power, better inexpensive displays, higher network bandwidth and lower latency) have combined to increase recent use of video conferencing applications. Yet the user experience problems we commonly see in today's video conferencing applications indicate that we are only beginning to harness the potential of the medium and that there are many further opportunities to improve it.

In regards to the particular problems we address in this paper, we are unaware of anyone else offering the set of features we describe. For addressing bad view angles, our most obvious competitor is the industry-standard small mirror window overlaid on the display (e.g. in Apple's FaceTime™) in which the user can see what he looks like to the other participants. This is an appealingly simple solution, which is why it is so common, but as previously described, we have found through formal user testing [4] that most users generally prefer not to see themselves all the time in a mirror window. They prefer such feedback to be visible only when something is wrong [7]. If the feedback is always available, they either find themselves distracted by it, or else it becomes background noise so that they do not actually catch the cue that something needs fixing. We can temporarily blend in the mirror window while the view angle is degraded, but the window's use is particularly problematic on small displays where

either it is too tiny to provide much feedback or it consumes a large percentage of the display.

Most mobile phone video conferencing solutions do not yet offer multi-party conferences, but those that do, such as Damaka™, are limited to a small number of participants (usually four) who then become postage-stamp-sized people on a smart phone display. Additionally, they tend to be displayed in a checkerboard pattern which does not comfortably match how people arrange themselves in physically co-located meetings. Tablets provide more display space, but even on tablets seeing too many tiled images may appear unnatural and they may be too small to see expressions clearly.

6. STATUS AND NEXT STEPS

Our demos show new ways to use mobile device sensors to provide natural, simple solutions for two fundamental problems of small devices: bad view angles and small display areas. We were planning to take advantage of the new sensors on upcoming webOS handsets and tablets to complete the solutions we describe above and then begin user testing. Unfortunately, a few days before the final draft of this paper, our company decided to remove itself from the area of consumer-facing electronics and so it appears unlikely that we will be able to pursue this project further. We hope that others are inspired to take on the kinds of user experience problems we describe here so that we may all enjoy higher quality mobile video conferencing in the future.

7. ACKNOWLEDGMENTS

Thank you to April Mitchell, Dan Gelb, and Kar Han Tan for their help and photographer services.

8. REFERENCES

[1] Edgido, C. 1988. Videoconferencing as a Technology to Support Group Work: A Review of its Failure.

[2] Finn, K., Sellen, A. and Wilbur, S. (editors). 1997. *Video-Mediated Communication*. L. Erlbaum Associates Inc.

[3] Gemmell, J., Toyama, K., Zitnick, C.L., Kang, T. and Seitz, S. 2000. Gaze Awareness for Video-Conferencing: A Software Approach. In *IEEE Multimedia*, 7, 4.

[4] Mitchell, A., Baker, M., Wu, C., Samadani, R. and Gelb, D. 2010. How Do I Look? *HP Labs Technical Report HPL-2010-175*, HP Laboratories.

[5] Tanguay, D., Gelb, D. and Baker, H. 2004. Nizza: A Framework for Developing Real-time Streaming Multimedia Applications. *HP Labs Technical Report HPL-2004-132*, HP Laboratories.

[6] Wilbur, S. and Whittaker, S. 1993. Conversations over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. In *Human-Computer Interaction*. Vol. 8, 389-428, L. Erlbaum Associates, Inc.

[7] Wu, C., Samadani, R., Mitchell, A., Baker, M. and Gelb, D. 2011. Visual Framing Feedback for Desktop Video Conferencing. In *IEEE Int. Conf. on Image Processing* (Brussels, Belgium, Sept. 2011).