Looking At You:

Fused Gyro and Face Tracking for Viewing Large Imagery on Mobile Devices

Neel Joshi[1], **Abhishek Kar**[1,2], **and Michael F. Cohen**[1]

[1]Microsoft Research
Redmond, WA 98052, USA
{neel, mcohen}@microsoft.com

[2]Indian Institute of Technology, Kanpur
Kanpur, Uttar Pradesh 208016, India
akar@iitk.ac.in

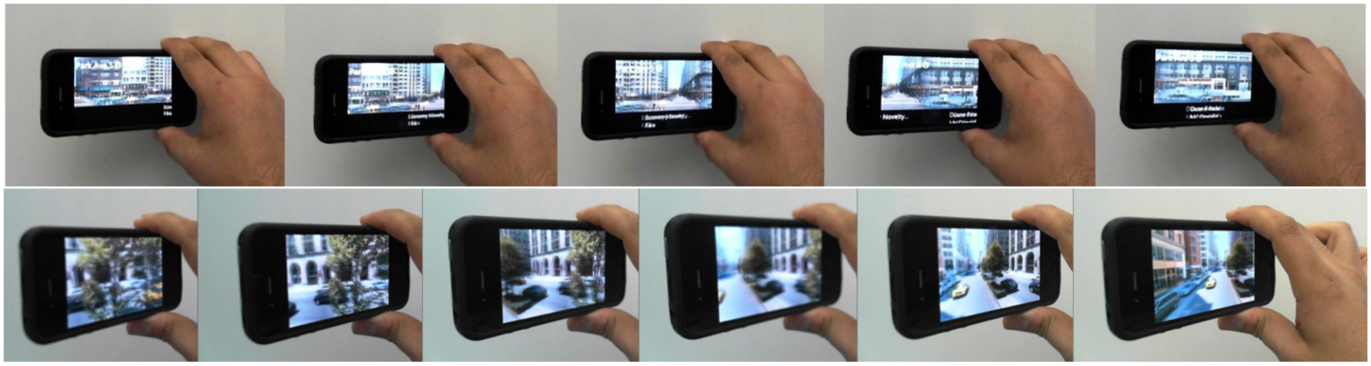

Figure 1. Top: Sliding along a multi-perspective panorama as the phone is tilted away from the viewer. Bottom: after moving the phone closer to zoom in, rotating the view of a 360° **panorama "bubble" in the street side imagery.**

ABSTRACT

We present a touch-free interface for viewing large imagery on mobile devices. In particular, we focus on viewing paradigms for 360 degree panoramas, parallax image sequences, and long multi-perspective panoramas. We describe a sensor fusion methodology that combines face tracking using a front-facing camera with gyroscope data to produce a robust signal that defines the viewer's 3D position relative to the display. The gyroscopic data provides both low-latency feedback and allows extrapolation of the face position beyond the the field-of-view of the front-facing camera. We also demonstrate a hybrid position and rate control that uses the viewer's 3D position to drive exploration of very large image spaces. We report on the efficacy of the hybrid control vs. position only control through a user study.

Author Keywords

Mobile device; sensors; navigation; viewing large imagery

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces - Interaction styles, Input devices and strategies;

INTRODUCTION

It is a fair guess that most viewing of photographs now takes place on an electronic display rather than in print form. Yet, almost all interfaces for viewing photos still try to mimic a static piece paper by "pasting the photo on the back of the glass", in other words, simply scaling the image to fit the display. This ignores the inherent flexibility of displays while also living with the constraints of limited pixel resolution. In addition, the resolution and types of imagery available continues to expand beyond traditional flat images, e.g., high resolution, multi-perspective, and panoramic imagery. Paradoxically, as the size and dimensionality of available imagery has increased, the typical viewing size has decreased as an increasingly significant fraction of photo viewing takes place on a mobile device with limited screen size and resolution. As a result, the mismatch between imagery and display has become even more obvious. While there are obvious limitations due to screen size on mobile devices, one significant benefit is that they are outfitted with numerous sensors including accelerometers, gyros, and cameras. The sensors, which currently are ignored in the image viewing process, can be used to provide additional image navigational affordances.

In this paper, we explore options for image viewing on mobile devices, leveraging the many sensors on the device. In particular, we use the low-latency gyros (fused with accelerometer and magnetometer readings) to sense changes in direction of the device as well as the front-facing camera to detect and track the 3D position of the viewer relative to the display, albeit with higher noise and latency. Fusion of these two sensor streams provides the functionality to create compelling inter-

CHI'12, May 5–10, 2012, Austin, TX, USA.

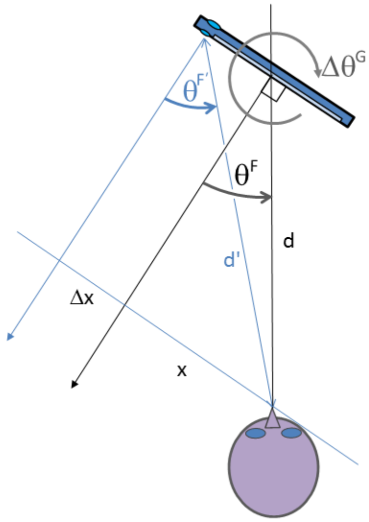

Figure 2. The face offset angle and distance is computed from a face tracked in a camera situated to the side of the display.

faces to a range of imagery. We have also chosen to focus on touch-free interaction (i.e., one hand holding the device with no finger-to-screen interaction). We do this to focus on the affordances of the sensor based interaction. The one handed interface frees the other hand for other tasks. In addition, the touch surface is freed up for other modes of interaction such as button selection without the need to disambiguate the intent of the touch gestures.

We demonstrate natural pan-tilt-zoom interfaces for many forms of complex imagery ranging from multiple images stitched to create a single viewpoint 360° panorama, multi-viewpoint image sets depicting parallax in a scene, street side interfaces integrating both multi-perspective panoramas and single viewpoint 360° panoramas. We also demonstrate touch-free panning and zooming over maps.

One aspect of large format and/or very wide-angle imagery is that there is a natural tension between a desire for direct positional control, i.e., a direct mapping of sensor output to position, versus rate control, mapping sensor position to velocity of motion across an image. For very large imagery, where only a small fraction is visible on-screen at any one time, positional control may require repeated clutching thus rate control has the advantage of requiring only a single sustained interaction to move large distances. Positionally mapping angle to view direction for panoramic imagery requires users spin in place which is difficult in many settings. This is also overcome with the hybrid approach. That said, positional control has greater precision within small regions and we do not want to lose this. We demonstrate a hybrid rate/position control through a single relationship between sensors and output that maintains the advantages of both.

Our technical contributions include the sensor fusion between the gyros and face tracking from a front-facing camera, a one-handed continuous zoom interface, as well as novel functional relationships between this sensing and the hybrid position/rate control of image viewing across numerous modalities.

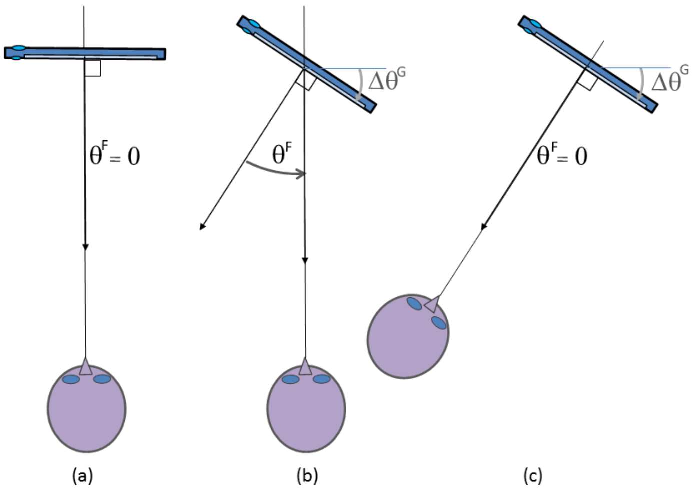

Figure 3. The gyro alone cannot distinguish between situations (b) and (c). The drift signal, θ^D, disambiguates these and brings the control in line with θ^F.

We report on a user study to assess the efficacy of the hybrid control vs. position only control, and also test against a standard one finger touch interface. We used the panoramic imagery setting since it is amenable to all three interfaces. We also test panning plus zooming in the hybrid vs. positional control. The results demonstrate no significant differences in speed and accuracy between the interfaces, thus making significant the other advantages for the hybrid control, e.g., it amenable to wider range of imagery, no need to spin in place, etc.

RELATED WORK

There is a long history of developing controls for viewing media that does not conveniently fit on a screen. This ranges from scrolling mechanisms for long text documents, to navigating over maps, to controls for panning and zooming across panoramic imagery. Scrolling mechanisms vary from the standard scroll bars, to radial widgets [18], and scrolling which adapts to the content [9]. Scrolling speed is sometimes coupled with zoom levels to simulate the experience of moving back from a document as the speed increases [8]. Interfaces to panoramic imagery either use positional control plus clutching or rate control based on changes in pointer position. Hinckley *et al.* [6] present a nice comparison and analysis for numerous scrolling techniques. Hinckley's book chapter [5] takes an even broader view of the subject and provides a nice overview of input technologies.

All of the above work assumes the affordance of either a pointing device such as a mouse or finger gestures with a touch sensitive device. We deliberately focus on touch-free navigation, leveraging the sensors on mobile devices. Touch-free navigation has been explored through gaze directed scrolling [13]; however, this requires more specialized sensors than can be found on mobile devices.

A number of papers have begun to focus on using the motion of a mobile device itself as the affordance for viewing imagery. *Boom Chameleon* [19] demonstrated a display tethered to a 6-axis arm that measured position and orientation of the display. This allowed the display to act as a virtual window to a 3D world. Many augmented reality applications, too numerous to list here, use a back-facing camera to replace the tethered arm. In many cases, they use fiducial marks visible in the scene to anchor the device's location, while some use features in natural scenes. Hua *et al.* [7] present efficient methods for using the camera to detect motion of the mobile device. They map detected motion to pointer motion in web pages, games, drawing interfaces, and even for such actions as picking up the phone to answer a call. In contrast to this body of work, our goal is to use the viewer's spatial relationship to the display (as opposed to the world position or motion) as an affordance for panning and zooming.

There has been some recent work that performs face tracking alone with similar goals to our own. Hannuksela *et al.* [3] describe a feature based face tracker for mobile devices, but do not describe any details of applications. Hansen *et al.* [4] use face tracking as an affordance for panning and zooming on images amongst other applications. In a related paper, Erikkson *et al.* [2] discuss using video tracking on more general terms. Face tracking alone has serious limitations including noise, latency, and limited field of view which we overcome through sensor fusion with gyros.

We also explore a hybrid position and rate controller to allow for larger scale exploration. A number of works have explored such hybrid approaches, typically by specifying different portions of input devices to each interaction type. For example, the *Rubber Edge* paradigm [1] uses the outer ring of a round area to create an elastic region for rate control while the center area provides positional control. We have no explicit visual device, but rather develop a continuous hybrid function that smoothly moves between positional and rate control as the angle between the device and the viewer changes.

A very relevant area is that concerning one-handed interfaces. Karlson *et al.* [11] have shown many cases in which one handed use is preferred by users for mobile device interaction. However, while one handed panning interfaces are studied with several de facto solutions [10], there are few widely used one handed interfaces that provide a continuous zoom control. The most widely used one-handed touch-based approach is the "double-tap" zoom in and "two-finger double-tap" zoom out popularized by the iPhone, which provide a non-continuous zooming by octaves (in contrast with the continuously zooming "two-finger" pinch that requires two hands for comfortable use).

Other one-handed methods include action-driven techniques such as the TapTap method that performs automatic zooming on targets [16] and others that use "rubbing" [14] or "rolling" [17] to indicate a zoom. These methods still cannot always easily distinguish pan actions from zoom actions and cannot easily support a continuous zoom. In contrast, it this our continuous panning is supported and panning and zooming are easily disambiguated.

Finally, a number of iPhone *apps* and online videos have appeared, which do not have related technical papers, that address some applications similar to our own. The *Tour Wrist* application (`http://www.tourwrist.com/`) allows exploration of 360° panoramas by physically spinning in place. They use a one-to-one mapping of device to viewing angle to produce a very effective positional control for such imagery. They do not have a hands-free zoom control. In addition, our interface does not require one to physically spin around which makes it possible to navigate the entire 360° space without standing. A very effective *Head Coupled Display* by Francone and Nigay is demonstrated in videos at `http://iihm.imag.fr/en/demo/hcpmobile/`. By tracking the head, they are able to create an effective virtual reality display of synthetic scenes. They do not show the use of face tracking for larger explorations or allow the user to move beyond the field-of-view of the device.

MAPPING SENSORS TO IMAGE TRANSFORMATIONS

Despite the lack of many traditional affordances found in a desktop setting (large display, keyboard, mouse, etc.), mobile devices offer a wide variety of sensors (touch, gyros, accelerometers, compass, and cameras) that can help overcome the lack of traditional navigation controls and provide a richer and more natural interface to image viewing. Our set of applications cover a variety of image (scene) viewing scenarios in which the imagery covers either a large field of view, a wide strip multi-perspective panorama, multi-views, or a combination of these. In particular, we explore interfaces for 360° panoramas, multi-view strips exhibiting parallax, maps, and the Microsoft Bing for iOS Streetside interface that combines very long multi-perspective strip panoramas with single view 360° views. Some imagery from each of these can be seen in Figures 5, 6, and the title figure. A common aspect of all of these is that the imagery requires *exploration* to view the full breadth of the data.

The most obvious way to explore imagery that cannot fit in the display is to use touch sensing to mimic a traditional interface. We have become accustomed to sliding a finger to pan and performing a two-fingered pinch for zooming. These affordances have four main drawbacks, however. First, one's fingers and hand obscure a significant portion of the display. Second, it becomes difficult to disambiguate touches designed for purposes other than navigation, for example, a touch designed to select a link embedded with the imagery. Finally, using the touch screen generally requires two hands, particularly for zooming. We, instead, investigate the use of more *natural* interfaces involving touch-free, motion of the device itself for image navigation.

Hybrid Gyro Plus Face Tracking

In the real world, we move our gaze relative to a scene, or move an object relative to our gaze to fully explore a scene (or object). In both cases, our head is moving relative to the scene. If one considers an image as a representation of a scene on the device, tracking the head relative to the device as an affordance for navigation seems like a natural fit.

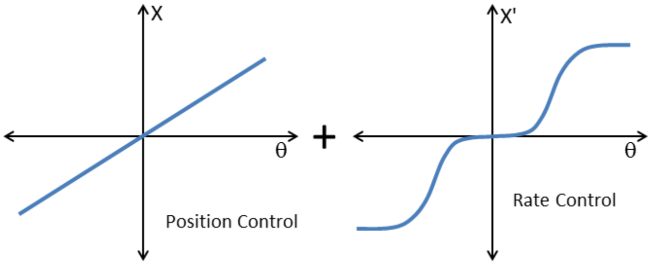

Figure 4. Hybrid mapping of angle to position and velocity

Face tracking alone can, in theory, provide a complete 3D input affordance, (x, y) position based on face location, and (z) depth based on face size. However, face tracking alone exhibits several robustness problems. Face tracking is costly and thus incurs some latency. In addition, the vision algorithms for tracking face position and size are inherently noisy as small changes in face shape and illumination can produce unexpected signals. This can be overcome somewhat through temporal filtering albeit at the price of more latency. Finally, face tracking is lost beyond an offset angle beyond the field of view of the front-facing camera (in our experiments we have found this limit to be about $\pm 15°$). Nonetheless, face tracking is unique in its ability to deliver a 3D signal that is directly relevant to image viewing applications.

Gyros provide a more robust and lower latency alternative for the 2D (x, y) angular position. For relative orientation, the gyros provide a superior signal, however they do drift considerably. We have commonly seen $5°$ drifts during a $360°$ rotation over 15 seconds. To compensate for this drift we use sensor fusion with an on-board magnetometer and accelerometers using the direct cosine matrix correction approach [15]. However, even with this correction, gyros alone cannot disambiguate between the cases shown in Figure 3 (b) and (c). In the first case, the user has rotated the device. In the second case, the user has rotated themselves carrying that same rotation to the device. Thus while the viewing angle has changed, the device angle has not. Our controls are based on viewing angle. Thus to achieve both robustness and liveness and reduce viewing angle ambiguity, we create a sensor fusion that is a hybrid of the gyro plus face tracking using the front-facing camera. We use the iPhone 4 for our experimental platform.

Face Tracker

A face is first located in the front-facing camera via a face finder based on the method similar to that of Viola and Jones [20], which returns a rectangle for the size and location of the face, i.e., $(position, scale)$. Given the field of view of the front-facing camera, $position$ is trivially transformed to horizontal and vertical angular offsets, $\theta_x^{F'}$ and $\theta_y^{F'}$. From here on, we will refer only to the more important horizontal offset, $\theta_x^{F'}$, and will drop the x subscript.

Horizontal Angle

Referring to Figure 2, there are two direct signals we track, $\theta^{F'}$, the angular offset of the face from the normal to the display (from the front-facing camera), and $\Delta\theta^G$ the change in rotation about the vertical axis tangent to the display (from the gyros). We estimate the distance d from the camera from face width. Given the fixed offset of the camera from the center of the display and $\Delta\theta^G$, we derive θ^F, the face's angular offset from the display center. We are now ready to compute the value, Θ, which is mapped to the position and rate control for our user interface.

$$\Theta_t = \alpha \cdot \Theta_{t-1} + (1-\alpha) \cdot (\theta_t^G + \theta_t^D) \tag{1}$$

Θ_t represents the value at time t we will map to our control functions. α serves to provide a small amount of hysteresis to smooth this signal. We have found a value of 0.1 to provide sufficient smoothing without adding noticeable lag. θ_t^G is the time integrated gyro signal, i.e., the total rotation of the device including any potential drift:

$$\theta_t^G = \theta_{t-1}^G + \Delta\theta_t^G \tag{2}$$

where $\Delta\theta_t^G$ represents the direct readings from the gyro. θ_t^D represents a smoothed signal of the difference between the face position, θ^F and the integrated gyro angle, θ^G. This quantity encompasses any drift incurred by the gyro as well as any rotation of the user himself (see Figure 3(c)). Since the face tracker runs slower than the gyro readings (approximately 10 Hz for the face tracker and 50Hz for the gyro), we record both the face position and gyro values each time we receive a face position. θ^D is thus defined by

$$\theta_t^D = \beta \cdot \theta_{t-1}^D + (1-\beta) \cdot (\theta_*^F - \theta_*^G) \tag{3}$$

where "$*$" represents the time of the most recent face track, and β serves to smooth the face signal and add hysteresis. We use a much higher value of $\beta = 0.9$ in this case. This produces a some lag time, which actually adds a side benefit we discuss in the context of the control mapping.

To summarize, Θ_t represents a best guess of the face position relative to the device even when the face is beyond the field of view of the device. Although face tracking is inherently slow and noisy, the gyro signal serves as a lively proxy with good accuracy over short time intervals. The face tracker is used to continuously *correct* the gyro input to bring it back in line with where the face is seen from the front-facing camera.

Distance

We use the face width in the camera's view as as proxy for the face's distance from the device. We use a time smoothed face size for this signal.

$$Z_t = \gamma \cdot Z_{t-1} + (1-\gamma) \cdot (1/FaceSize) \tag{4}$$

where $\gamma = 0.9$ to smooth over noisy readings albeit at some cost of latency.

Hybrid Position and Rate Control

Given the angular offset, Θ_t, we now are left with the mapping between this value and the controls for viewing the imagery. The simplest and most intuitive mapping is a *position*

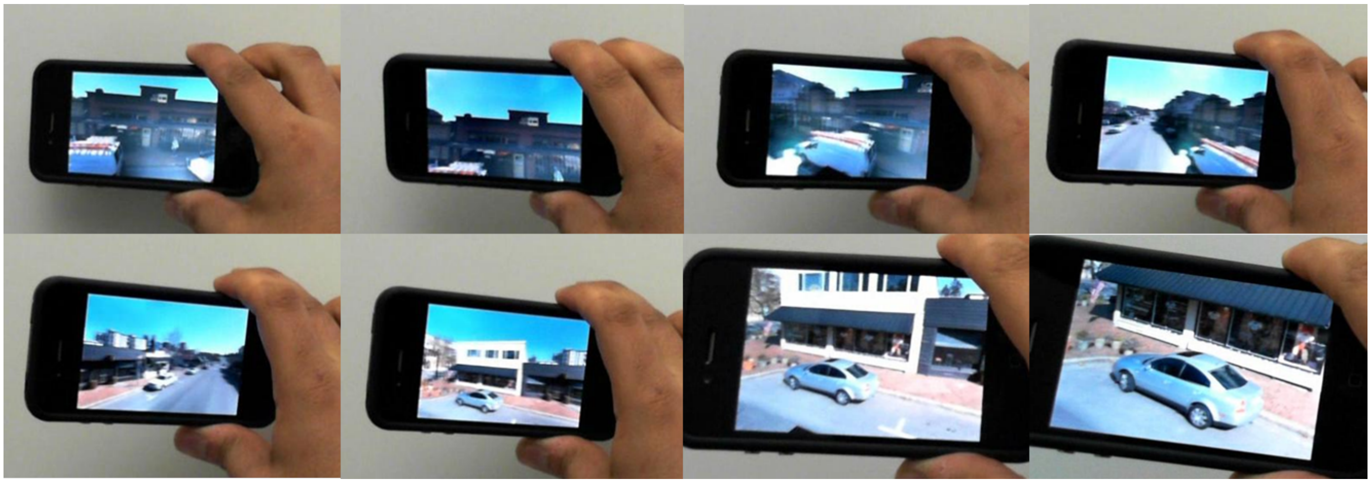

Figure 5. Frames depicting the changes in view angle and zoom for a 360° panorama as the viewer manipulates the device.

control, in which the Θ_t is mapped through some linear function to the position on the imagery (i.e., angle in a panorama, position on a large flat image, or viewing position in a multi-view parallax image set). Position mapping can provide fine control over short distances and is almost always the control of choice when applicable.

Unfortunately, such a simple mapping has severe limitations for viewing large imagery. The useful domain of Θ_t is between $\pm 40^\circ$ since beyond this angle the phone display becomes severely foreshortened and un-viewable. For 360° panoramas or very long multi-perspective images, this range is very limited. The alternatives are to provide *clutching* or to create a *rate* control in which Θ_t is mapped to a *velocity* across the imagery. Although rate controls provide an infinite range as the integrated position continues to increase over time, they have been shown to lack fine precision positioning as well as suffering from a tendency to overshoot.

We have chosen to formulate a hybrid position plus rate control to achieve the benefits from each. More specifically we sum two transfer functions. The position control provides a linear mapping from θ_t to a value x_t (See Figure 4 left).

$$x_t = m_t \theta_t \tag{5}$$

where m_t adjusts the slope to smoothly transition between the position and rate regimes as discussed below.

The rate control (Figure 4 right) formed by two sigmoid functions maps θ_t to a velocity x'_t. Near zero values of Θ_t, x'_t is also zero thus the positional control dominates. As Θ_t increases, some velocity, x'_t is added to the output soon dominating the positional control. The sigmoid functions limit x'_t to some maximum velocity, x'_{max}. Thus the final change in the output, ΔX, from time $t-1$ to time t is given by:

$$\Delta X_t = x_t - x_{t-1} + x'_t \cdot (T_t - T_{t-1}) \tag{6}$$

where T represents clock time. We will discuss the specific meaning of the output value, X, when we discuss applications.

We have found that when transitioning from rate control to position control that the desire to "brake" and stop the motion control quickly conflicts briefly with the position control. Thus, we adjust m_t to reflect this and smooth between these regimes. More specifically:

$$\begin{aligned} m_0 &= 1 \\ m_t &= \mu m_{t-1} + (1-\mu) m^* \\ m^* &= \max((1 - x'_t / x'_{max}), 1) \end{aligned}$$

where μ is a smoothing factor between regimes that we have set to 0.95. As a result, when a user is transitioning out of rate control, the sensitivity of the position control temporarily reduces.

An interesting aspect of the latency in the face tracking over the gyro signal is that quick rotations of the body such as depicted in Figure 3(c) do result in positional changes. We have found this has a side benefit of keeping the whole control feeling lively and responsive.

Zoom Control

In the panorama and street side applications, Z_t is linearly mapped to zoom level. We cap the minimum zoom level at a bit less than arm's length. The street side application has a fixed zoom level at which a mode change takes place between the multi-perspective panoramas and cylindrical panoramas. To avoid rapid mode changes near this transition point, we ease in a small offset to the zoom level after the mode switch and then ease out the offset after the mode switches back. In the map viewing application, Z_t is mapped parabolically to zoom level, to allow a greater range of zoom control with the same arm's length viewing area.

As the view zooms in, the same angular change in viewer position maps to increasingly larger motions across the screen. As a result, at high zoom-levels, panning appears magnified as small hand motions map to large image displacements. To compensate, we perform a zoom-based scaling of the view control ΔX_t in Equation 6. Specifically, ΔX_t is scaled such

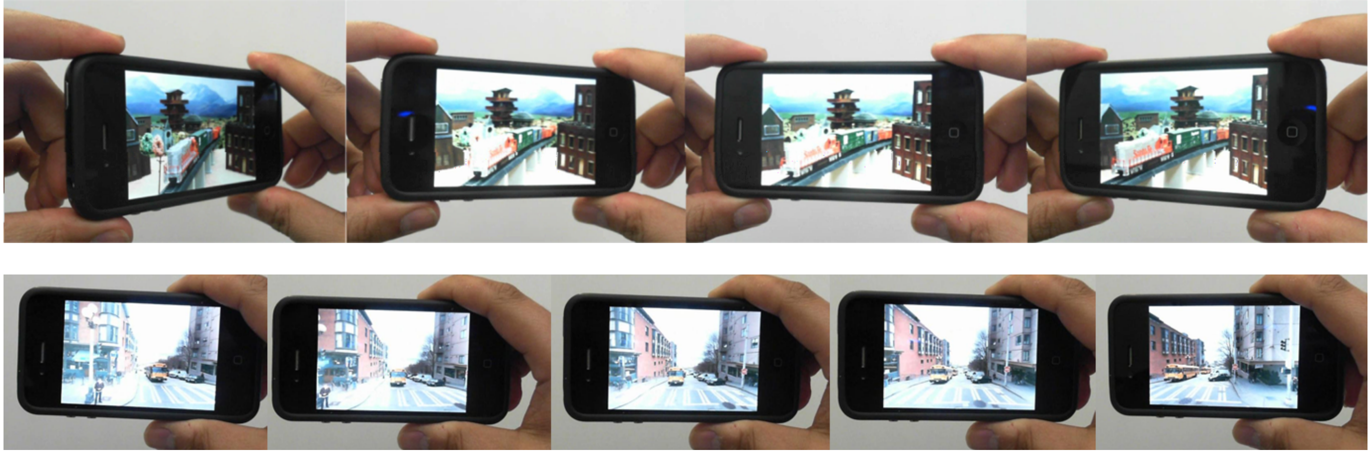

Figure 6. Frames depicting the changes in parallax as the viewer manipulates the device.

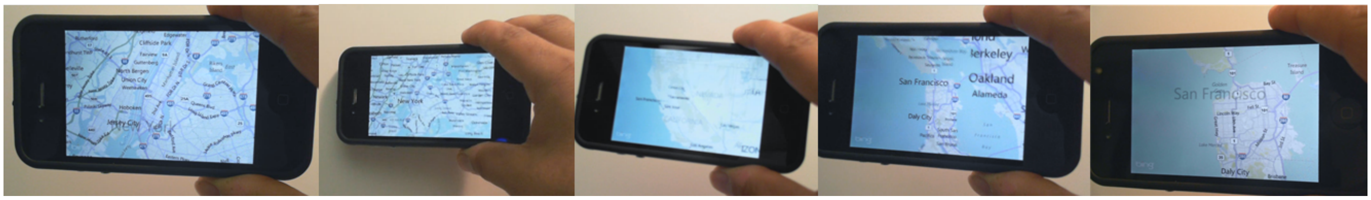

Figure 7. Frames depicting panning and zooming a map.

that as the distance between the viewer and device decreases, the change in the view velocity decreases:

$$\Delta X_t^{scaled} = \Delta X_t Z_t \tag{7}$$

APPLICATIONS

We have applied the interaction paradigm described above to a number of image viewing applications. These include wide-angle imagery such as 360° panoramas and *parallax* photos consisting of a series of side-by-side images. We have also built a new interface for street side imagery for the Microsoft Bing for iOS maps application. This includes integrated very long multi-perspective images and 360° panoramas.

Panoramas

Wide-angle and 360° panoramas have become a popular form of imagery especially as new technologies arrive making their construction easier. Sites, such as Gigapan (`http://www.gigapan.org/`) which hosts high resolution panoramas, and the *bubbles* of street side imagery found on sites such Google StreetView and Microsoft Bing Streetside are two examples.

By interpreting ΔX_t at each frame time as a change in orientation, and Z_t as the zoom factor, we demonstrate an interface to such imagery that does not require two-handed input or standing and physically turning in place. See the accompanying video for a dynamic version of the panorama viewer.

Parallax Images

By sliding a camera sideways and capturing a series of images one can create a virtual environment by simply flipping between the images. An automated and less constrained version for capture and display of *parallax* photos is demonstrated by Zheng *et al.* [21].

In this application, ΔX_t at each frame time represents a relative offset of the virtual camera. We demonstrate an interface to such imagery that creates a feeling of peering into a virtual 3D environment. In this case, the position control dominates. Figure 6 shows a number of frames from a session examining a *parallax photo*. Again, see the accompanying video for a dynamic version.

Street Side Imagery

A new interface to street side imagery was demonstrated in Street Slide [12]. The original imagery consists of a series of 360° panoramas set at approximately 2 meter intervals along a street. The Street Slide paradigm was subsequently adapted to create long multi-perspective strip panoramas constructed by clipping out and stitching parts of the series of panoramas. This new imagery is displayed in the Microsoft Bing for iOS mapping application on the iPhone. The application automatically flips between the long strip panoramas and the 360° panoramas depending on zoom level. The official iPhone app uses traditional finger swipes and pinch operations.

We have applied our navigation paradigm as a new user interface on top of the Street Side application. The title figure shows a number of frames from a session examining a street. Since there are two modes, the meaning of ΔX_t switches.

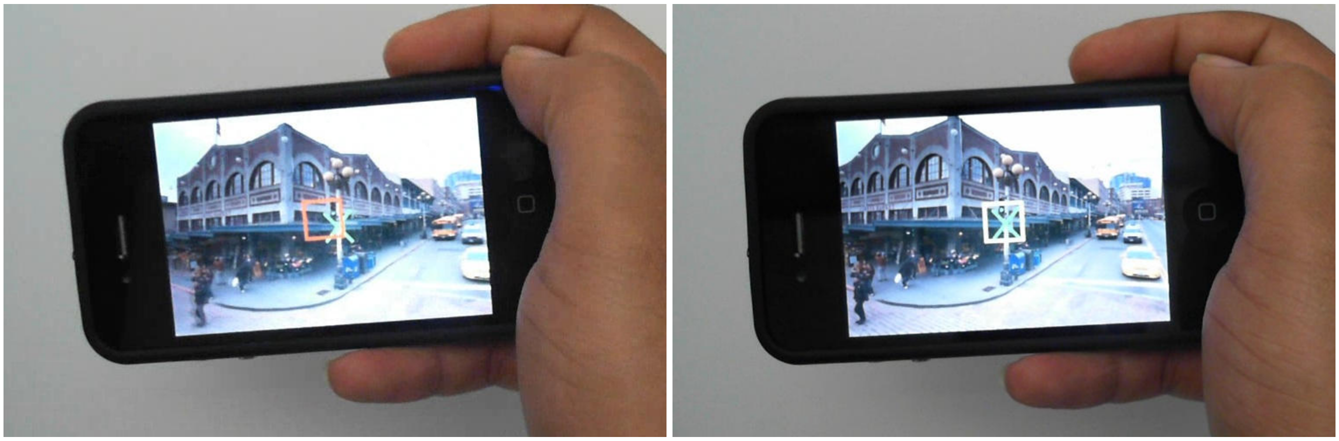

Figure 8. To evaluate user performance, subjects we asked to find a green X placed somewhere in the imagery and maneuver it into a square centered on the screen.

In *slide* mode, ΔX_t moves the view left and right along the street side. Z_t zooms the strip panorama in and out. At a given zoom level, the mode switches automatically to the corresponding 360° panorama at that location on the street. At this point, we revert to the panorama control described above. Zooming out once more returns to the slide mode. The accompanying video depicts a dynamic version including mode switches between strip panoramas and the more traditional circular panoramas. (Note that as we are running off of live, streaming data bandwidth and latency issues in the application sometimes lead to delays in imagery appearing.) Navigation now requires only one hand leaving the other hand free for unambiguous access to other navigation aids and information overlaid on the location imagery.

Maps

A pan and zoom interface provides a natural means for browsing map data. In Figure 7 and the accompanying video, we also show our interface used to explore a hierarchical map. The user navigates across the US starting from New York, zooming out and panning across the country, and then zooming back in to a view of San Francisco. This is achieved by simply moving the device away, tilting it “west” and pulling the device back towards the viewer.

EVALUATION

We conducted a study to gather feedback on our interaction techniques. The goal of the study was to assess the efficacy of two conditions of our one handed interface: hybrid position plus rate control vs. position only control, relative to a more familiar one-handed, one finger (on the same hand) position control. Our hypothesis is that these techniques provide an immersive, engaging, and intuitive interaction paradigm for large image navigation, without adding significant complexity over more traditional touch-based interfaces. In particular, in our study we quantitatively test the latter half of this hypothesis, by examining whether these techniques require more time to acquire a target position. The former part of the hypothesis is qualitatively examined through a short questionnaire. We also assessed the efficacy of our view based-zoom control under both the hybrid and position controls. We thus tested five conditions: finger based position control, touch-free position control, touch-free position+rate control, touch-free position and zoom control, and touch free position+rate plus zoom control.

Participants

We recruited ten volunteers from a large software company (5 male, 5 female, between the ages of 22 and 45) that were unaffiliated with this work.

Procedure

Participants were asked to perform timed tasks on each of the five conditions listed above. For each condition, the participant was asked to navigate over a single 360 degree panoramic image. (We chose to use panoramic imagery since it is amenable to all the interfaces, due to the fact that the angular position is naturally limited by its wrap-around nature.)

For each condition, the experimenter first demonstrated the interface to the participant, who was then allowed to practice with it until they indicated they were comfortable with the interface (generally less than 1 minute). The three non-zooming conditions were tested first (in random order) followed by the two zooming interfaces (in random order). Subjects were seated in a swiveling chair.

After the initial exploration period, the timed experiment began. The experiment involved finding a green X placed somewhere in the imagery (most often off-screen from when they start) and maneuvering it into a red square fixed in the center of the screen (see Figure 8). The X was placed at stratified random positions over 5 strata relative to the horizontal viewing angle at the beginning of each trial, i.e., 2 of the 10 X’s were placed plus or minus 36 degrees from the current view position, 2 were between 36 and 72 degrees away, 2 between 72 and 108, 108 and 144, and between 144 and 180 degrees away.

In the latter 2 zoomable interfaces, the subjects were also instructed to zoom in and out to get the X to be the same size

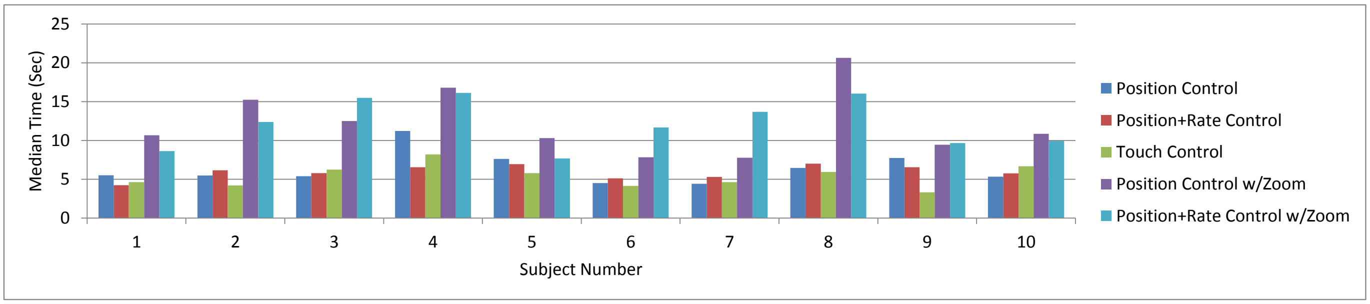

Figure 9. Median completion time per subject.

as the square. The width of the X was uniformly, randomly set to be in range of plus or minus 25% of the size in the non-zooming conditions. The goal was considered reached when for 0.5 seconds the X remained within the square plus or minus 10% of the square size. In the zoomable interface, the X must also be within 20% of the square size.

Each condition was tested 10 times consecutively, thus each participant performed a total of 50 tests. The time to goal was recorded, as were the device, face, and view positions at a 30Hz rate during each trial.

We additionally asked users to rate the ease of use of each condition on a 5-point Likert scale and provide general comments about the non-zoomable and zoomable conditions.

Results

The qualitative questions revealed that all participants rated the position only control and the position+rate control to be within one Likert scale level in terms of ease of use. Two found position only control one level easier, and one found position+rate control one level easier. Half indicated the touch control was easier than the gesture based interfaces but two of these participants also indicated some discomfort in using the finger control. They preferred the behavior of the touch control locking in place when the finger was lifted. The other two interfaces were harder to hold still when they were at the target. This observation is not unexpected. Although the ballistic phase of the target acquisition works well with our vision based interface, the correction phase (near the target) is exacerbated by the noise in the vision system, as opposed to the noiseless finger interface which simply stops when the finger is lifted.

A few specific comments were:

“I liked this one [position only] - it was quick to spin around on my chair, and fun to have an excuse to do so. If I were actually out in public, though, I might feel a little silly turning around like that.”

“[Position+rate] took some getting used to, but then I felt like I got the hang of it.”

“The motion methods [position and position+rate] were more fun than using a finger.”

“When I overshot in [position+rate], it was sometimes tough to back up a small amount.”

The ease-of-use questions showed that all subjects found the zooming interfaces more difficult than the non-zooming interfaces, as expected. One commented “Zooming was hard to grasp. I think having used *touch* as a means to zoom has sort of been baked in my behavior so moving the phone away or closer to me just felt foreign.”

Analysis

Statistical analysis of the timing data reinforces the qualitative user feedback and supports our hypothesis. The timing results were analyzed using a non-parametric Kruskal-Wallis test which tests for differences in medians rather than means of a normal distribution. As can be seen in Figure 10 the distribution of time to completion are significantly non-normal and contain a number of outliers, as periodically the user missed finding the target on a first pass.

For the non-zooming interfaces, all three interfaces, touch, position, and hybrid required similar median times to completion with no significant differences (Kruskal-Wallis test, $\chi^2 = 2.28$, $P = 0.32$). In observing the study, it appeared that all three interfaces were found intuitive after a very short introductory period. That said, the one-fingered touch interface was found very straining by the tenth target, the position control had everyone spinning 360 degrees multiple times, and the hybrid control allowed all to sit relatively still and accomplish the same tasks. The zooming interfaces were both equally fast (or slow) with no significant differences ($\chi^2 = 0.2$, $P = 0.65$) in the median times, but required approximately twice the amount of time. The increase in overall time relative to the non-zooming conditions is to be expected given that the zoom control added an extra degree-of-freedom and precision required for the target tracking. The latency in the zooming (with no low-latency sensors to compensate) caused some oscillation in the target acquisition which caused considerable (audible) frustration at times. We hope to have a faster implementation to overcome the latency.

Interestingly, we found a significant difference in completion time across subjects for the non-zooming conditions

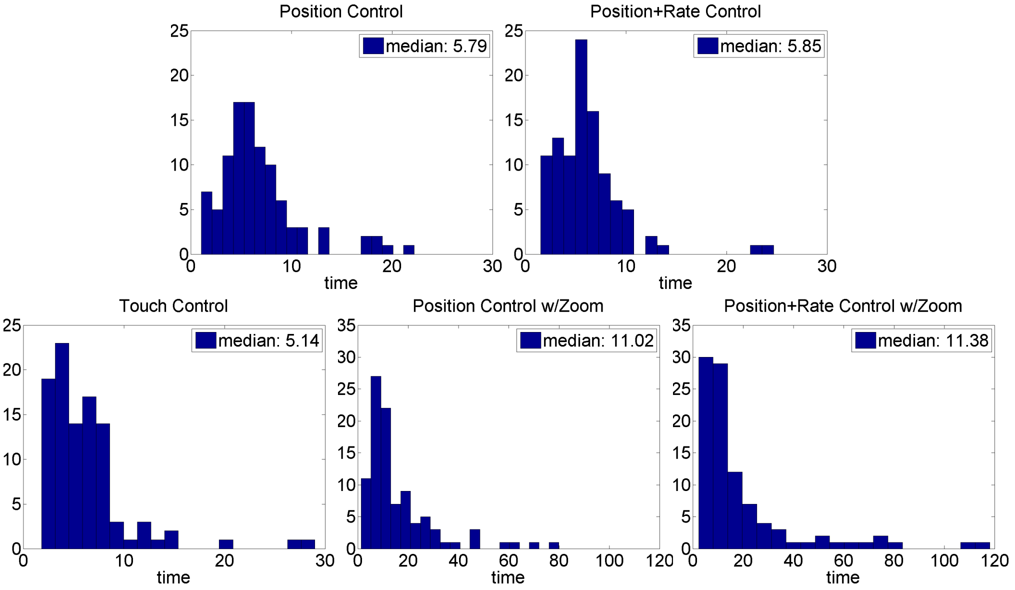

Figure 10. Distributions of completion time per condition.

($\chi^2 = 35.43$, $P < 0.001$) and no significant difference for the zooming conditions ($\chi^2 = 12.71$, $P = 0.18$), which suggests that zooming control was equally difficult (or easy) across users, possibly due to the previous lack of exposure to this type of interface control. This could be partially explained by a learning effect on the zooming conditions that shows a potentially significant decrease in time as more trials were performed ($\chi^2 = 14.6$, $P = 0.1$) while there was no significant difference in the non-zooming conditions ($\chi^2 = 12.51$, $P = 0.19$). A further study on the learning affect appears warranted. We found no significant differences between gender groups for both non-zooming ($\chi^2 = 0.62$, $P = 0.43$) and zooming ($\chi^2 = 0.22$, $P = 0.64$) sets of conditions.

In summary, the position+rate controls appear to be approximately equivalent in usability to the position only controls. The fact that the position+rate control is the only one capable of navigating very large imagery in a touch free fashion is encouraging. Clearly, such interfaces need to be further examined in specific scenarios such as maps and other very large imagery.

Plots showing per-subject and across subject completion times are shown in Figures 9 and 10. Figure 11 shows median completion times across the trials.

CONCLUSION AND FUTURE WORK

We have demonstrated a robust sensor fusion approach for estimating a viewer's position relative to a mobile device. A noisy face tracker is coupled with a more responsive gyro signal to provide a robust natural input that extends beyond the field-of-view of the front-facing camera. We have shown a hybrid position and rate control mechanism to map the sensor input to viewing large imagery. The hybrid control provides a touch-free interface for browsing a variety of media, and avoids many problems such as the need for clutching and/or a need to spin in place to traverse large scenes. Applications include viewing 360° panoramas, sets of parallax photos, and long multi-perspective street side panoramas. A user study confirmed there is no appreciable difference in efficacy between the hybrid control and either a one finger control or a position only control.

We also demonstrate the ability to zoom the imagery based on the distance of the user from the device based only on the front-facing camera. Although the system incurs some latency, users were able to acquire targets both in space and zoom level with almost no training.

There is clearly an enormous open area for further exploration and evaluation. Many other types of media are amenable to interfaces such as the one we describe. In particular, we hope to extend our interface to more mapping applications.

Also, it is our belief that a view-dependent interface provides many intangible benefits such as providing a better sense of "immersion" when viewing a scene. Due to the direct interplay of a gesture and change of view in a touch-based method, it is very hard to disambiguate the sense of moving around in a virtual world versus the sensation of physically moving

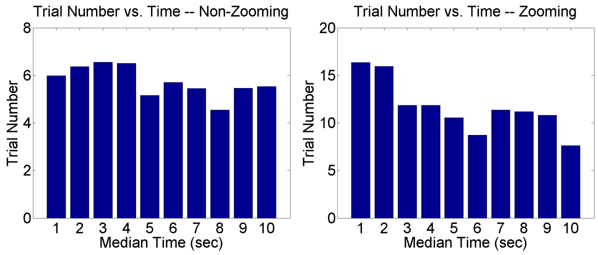

Figure 11. Median completion time per trial.

the world. We are very interested in exploring the tension between these two paradigms and how they are affected by both view- and touch-based interfaces. Mobile devices offer a wealth of untapped sensor data for integration with user interfaces to many applications. We are very excited to continue exploring this new space.

REFERENCES

1. Casiez, G., Vogel, D., Pan, Q., and Chaillou, C. Rubberedge: reducing clutching by combining position and rate control with elastic feedback. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, ACM (New York, NY, USA, 2007), 129–138.
2. Eriksson, E., Hansen, T. R., and Lykke-Olesen, A. Movement-based interaction in camera spaces: a conceptual framework. *Personal Ubiquitous Comput. 11* (December 2007), 621–632.
3. Hannuksela, J., Sangi, P., Turtinen, M., and Heikkilä, J. Face tracking for spatially aware mobile user interfaces. In *Proceedings of the 3rd international conference on Image and Signal Processing*, ICISP '08, Springer-Verlag (Berlin, Heidelberg, 2008), 405–412.
4. Hansen, T. R., Eriksson, E., and Lykke-Olesen, A. Use your head: exploring face tracking for mobile interaction. In *CHI '06 extended abstracts on Human factors in computing systems*, CHI EA '06, ACM (New York, NY, USA, 2006), 845–850.
5. Hinckley, K. Input technologies and techniques. In *The Human-Computer Iteraction Handbook*, A. Sears and J. A. Jacko, Eds. Addison Wesley, 2008, 161–176.
6. Hinckley, K., Cutrell, E., Bathiche, S., and Muss, T. Quantitative analysis of scrolling techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, CHI '02, ACM (New York, NY, USA, 2002), 65–72.
7. Hua, G., Yang, T.-Y., and Vasireddy, S. Peye: toward a visual motion based perceptual interface for mobile devices. In *Proceedings of the 2007 IEEE international conference on Human-computer interaction*, HCI'07, Springer-Verlag (Berlin, Heidelberg, 2007), 39–48.
8. Igarashi, T., and Hinckley, K. Speed-dependent automatic zooming for browsing large documents. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, UIST '00, ACM (New York, NY, USA, 2000), 139–148.
9. Ishak, E. W., and Feiner, S. K. Content-aware scrolling. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, UIST '06, ACM (New York, NY, USA, 2006), 155–158.
10. Karlson, A. K., and Bederson, B. B. Understanding single-handed mobile device interaction. Tech. rep., HCIL-2006-02, 2006.
11. Karlson, A. K., Bederson, B. B., and Contreras-Vidal, J. L. Understanding One-Handed Use of Mobile Devices. In *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, J. Lumsden, Ed. Information Science Reference, 2008, ch. VI, 86–101.
12. Kopf, J., Chen, B., Szeliski, R., and Cohen, M. Street slide: browsing street level imagery. *ACM Trans. Graph. 29* (July 2010), 96:1–96:8.
13. Kumar, M., and Winograd, T. Gaze-enhanced scrolling techniques. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, ACM (New York, NY, USA, 2007), 213–216.
14. Olwal, A., Feiner, S., and Heyman, S. Rubbing and tapping for precise and rapid selection on touch-screen displays. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, ACM (New York, NY, USA, 2008), 295–304.
15. Premerlani, W., and Bizard, P. Direction cosine matrix imu: Theory. `http://gentlenav.googlecode.com/files/DCMDraft2.pdf`.
16. Roudaut, A., Huot, S., and Lecolinet, E. Taptap and magstick: improving one-handed target acquisition on small touch-screens. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '08, ACM (New York, NY, USA, 2008), 146–153.
17. Roudaut, A., Lecolinet, E., and Guiard, Y. Microrolls: expanding touch-screen input vocabulary by distinguishing rolls vs. slides of the thumb. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, ACM (New York, NY, USA, 2009), 927–936.
18. Smith, G. M., and Schraefel, M. C. The radial scroll tool: scrolling support for stylus- or touch-based document navigation. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*, UIST '04, ACM (New York, NY, USA, 2004), 53–56.
19. Tsang, M., Fitzmzurice, G. W., Kurtenbach, G., Khan, A., and Buxton, B. Boom chameleon: simultaneous capture of 3d viewpoint, voice and gesture annotations on a spatially-aware display. *ACM Trans. Graph. 22* (July 2003), 698–698.
20. Viola, P., and Jones, M. J. *Robust Real-Time Face Detection*, vol. 57. Kluwer Academic Publishers, Hingham, MA, USA, May 2004.
21. Zheng, K. C., Colburn, A., Agarwala, A., Agrawala, M., Salesin, D., Curless, B., and Cohen, M. F. Parallax photography: creating 3d cinematic effects from stills. In *Proceedings of Graphics Interface 2009*, GI '09, Canadian Information Processing Society (Toronto, Ont., Canada, Canada, 2009), 111–118.