Gaze Locking: Passive Eye Contact Detection for Human–Object Interaction

Brian A. Smith **Qi Yin** **Steven K. Feiner** **Shree K. Nayar**
Department of Computer Science, Columbia University
450 Comp. Sci. Bldg., 1214 Amsterdam Ave., New York, NY 10027 USA
{brian, qiyin, feiner, nayar}@cs.columbia.edu

ABSTRACT

Eye contact plays a crucial role in our everyday social interactions. The ability of a device to reliably detect when a person is looking at it can lead to powerful human–object interfaces. Today, most gaze-based interactive systems rely on gaze tracking technology. Unfortunately, current gaze tracking techniques require active infrared illumination, calibration, or are sensitive to distance and pose. In this work, we propose a different solution—a passive, appearance-based approach for sensing eye contact in an image. By focusing on gaze *locking* rather than gaze tracking, we exploit the special appearance of direct eye gaze, achieving a Matthews correlation coefficient (MCC) of over 0.83 at long distances (up to 18 m) and large pose variations (up to $\pm 30^\circ$ of head yaw rotation) using a very basic classifier and without calibration. To train our detector, we also created a large publicly available gaze data set: 5,880 images of 56 people over varying gaze directions and head poses. We demonstrate how our method facilitates human–object interaction, user analytics, image filtering, and gaze-triggered photography.

Author Keywords

Gaze-based interaction; passive eye contact detection; human–object interaction; user analytics; image filtering; gaze-triggered photography; human vision.

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces—Input devices and strategies; I.4.8 [Image Processing and Computer Vision]: Scene Analysis

INTRODUCTION

Eye contact is a major form of nonverbal communication and plays a crucial role in our everyday social encounters. We use it to seek information (e.g., to see how something we say is received) and regulate interaction (e.g., by signaling when it is someone else's turn to speak), among other uses [1, 19, 40]. Interestingly, while we have a hard time determining the exact angle at which someone else is looking, we seem to be very good at determining when someone else is looking *at us* (i.e., at or very near our eyes), and are acutely aware of it [15].

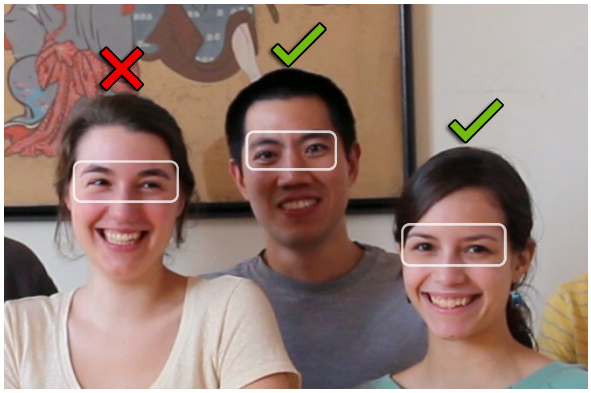

Figure 1. Gaze locking. We propose the idea of sensing eye contact directly from an image in a passive, appearance-based manner. The main idea is to focus on gaze *locking* (a binary problem) rather than gaze tracking (a continuous problem) and exploit the special appearance of direct eye gaze. Our approach can be used to facilitate a wide range of applications.

Most gaze-based interactive systems today rely on gaze tracking. They find the exact angle users are looking at instead of sensing eye contact directly. Although gaze tracking has been extensively studied and current methods [2, 3, 18, 24, 36] are highly accurate, they suffer from several limitations that restrict their practical use. They generally work only over short distances (often 80 cm or less) [2, 3, 18, 24, 36] or with direct head poses [2, 18, 36], or require active infrared illumination [2, 3, 24, 36], intrusive equipment (such as head-mounted cameras), or extensive calibration [3, 24]. One exception is Shell et al.'s system [31], which does in fact sense eye contact directly, but also requires active illumination. To address these limitations, we make the following contributions.

Gaze Locking

We propose the idea of sensing eye contact directly from an image in a passive, appearance-based manner (Figure 1). The main idea is to focus on gaze *locking* (a binary problem) rather than gaze tracking (a continuous problem) and exploit the special appearance of direct eye gaze. In addition to being passive (no special illumination or hardware required), our approach is non-intrusive, calibration-free, and robust to distance and head pose. Like Shell et al.'s active method [31], it can be used to allow humans to interact with computers, devices, and other objects just by looking at them.

UIST'13, October 8–11, 2013, St. Andrews, United Kingdom.

http://dx.doi.org/10.1145/2501988.2501994

Sample Detector

As a proof of concept, we demonstrate that even a simple and lightweight gaze locking system can yield accurate and robust results. Our sample detector uses very basic features—the eye area's pixel intensities—yet achieves a Matthews correlation coefficient (a measure of accuracy for binary classifiers) of over 0.83 at long distances (up to 18 m) and large pose variations (up to $\pm 30^\circ$ of head yaw rotation) without requiring calibration. This equates to a 92% accuracy on our training data set. It runs at over 20 FPS on a computer with an Intel Core i5-3470 processor, 8 GB of RAM, and an NVIDIA GeForce GTX 660M graphics card. A more advanced classifier could be used to improve accuracy even further.

Human Performance Evaluation

We performed a study to see how accurate people are at sensing eye contact and found several interesting results. For example, we found that people achieve MCCs of over 0.2 at distances of 18 m, and that their accuracy decreases roughly linearly over distance regardless of others' (horizontal) head orientations. We also found that people are often more accurate when they can only see one of the other person's eyes.

Gaze Data Set

To facilitate our human study and provide training data for our sample detector, we created a gaze data set of 56 people and 5,880 images, available at `http://www.cs.columbia.edu/CAVE/databases/columbia_gaze/`. It has more images and fixed gaze targets than any other publicly available gaze data set. To ensure robustness, our data set spans a variety of parameters: 5 head poses and 21 gaze directions per head pose. Our subjects were ethnically diverse and 21 of them wore glasses. We use our data set for gaze locking purposes, but it can serve as a very large resource for gaze tracking purposes as well.

Demonstration of Applications

Lastly, we show a few applications that gaze locking facilitates. First, since camera modules are becoming increasingly small and inexpensive to produce, we can allow any device to respond to eye contact by embedding a camera that serves as an eye contact sensor inside it. Our technique can thus serve as a backbone for allowing humans to interact with computers, devices, and other objects simply by looking at them. In addition, our method is passive and hence can be applied to any existing image. Therefore, it can be used to sort images on the web and on personal computers by their degree of eye contact, improving image search. Finally, we can incorporate a gaze trigger in cameras to capture group photos exactly when everyone in the group is looking straight back.

RELATED WORK

Gaze Estimation and Tracking

Not surprisingly, both gaze estimation and gaze tracking (gaze estimation at video rate) have been studied extensively in the past few decades. Hansen and Ji [17] and Morimoto and Mimica [26] provide excellent surveys of earlier work. Ideally, a gaze tracking system should be accurate, passive, non-intrusive, calibration-free, and robust to distance and head pose [26]. Unfortunately, current systems maintain accuracy at the expense of other qualities—they are predominantly active systems that work only at close range (80 cm or less).

For example, Morimoto et al.'s [24] and Beymer and Flickner's [3] feature-based techniques are accurate and robust to head pose, but are active, require calibration, and work only at close range. Baluja and Pomerleau's [2] and Tan et al.'s [36] appearance-based techniques are accurate and calibration-free, but are active, sensitive to head pose, and work only at close range. Hansen and Pece's [18] commercial off-the-shelf (COTS) system is passive and easy to calibrate, but is sensitive to head pose and again works only at close range. Nishino and Nayar's method [27] is passive and produces an image of what a person is looking at, but requires a high resolution image of the eye to provide useful results. Stiefelhagen et al.'s eye tracker [34] is passive and boasts an accuracy of up to 1.3°, but requires a fixed head pose and was only tested on four users at close range. They also developed an earlier gaze tracking system [33] that tracks only head pose and not eye gaze direction. Our gaze locking approach can be applied to any image (including those from COTS products or even from the Web) and is accurate, passive, non-intrusive, calibration-free, and robust to distance and head pose.

Lastly, Cadavid et al. [6] use spectral regression to passively detect whether an infant is looking at his or her parent's face, but their work differs from our gaze locking approach in three important respects. First, their system does not perform gaze locking since the camera is not worn by (and co-located with) the parent, but rather shows a third-party view. Second, neither the precision nor accuracy of their approach was directly tested. The parent's face subtends a large solid angle with respect to the infant (resulting in a large "sweet spot" for detection), and the system's underlying positive and negative samples were manually labeled by human coders viewing a simultaneous video recording showing both the infant's and parent's faces. We prove that our sample detector has a precision of 5°, allowing it to be used with very small objects. We also show in our human performance evaluation that, in the case of eye contact detection, humans themselves are actually inaccurate 9–40% of the time. Third, their work addresses only the specific application of infant–parent interaction for a single subject, and is not generalizable for human–object interaction with arbitrary humans and objects. We explore the concept of gaze locking, introduce several representative applications, and demonstrate that a very simple classifier can yield robust results with arbitrary humans and objects.

Gaze Perception

A number of studies evaluating people's perception of gaze have been performed, starting with Gibson and Pick's study [16] using six subjects and an in-person gazer. It concludes that, at 2 m, people sense eye contact when others are looking between the left and right edges of their face, but the authors urge a repeat of the experiment using a model (e.g., a set of photographs) as stimulus instead of a live person. Cline's in-person study [9] using a half-mirror incorporates several head poses and eye occlusion, confirming Gibson and

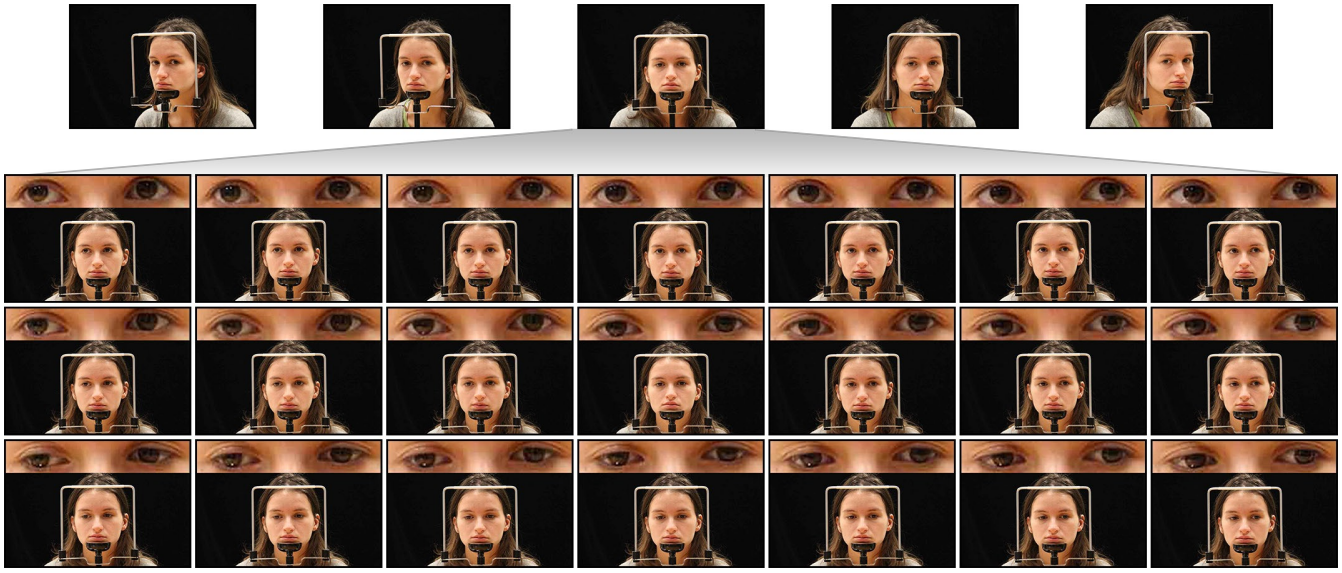

Figure 2. Sample data set images. Our gaze data set includes 56 subjects and five different head poses (shown on top): 0°, ±15°, and ±30° horizontally. For each subject and head pose, there are 21 different gaze directions (shown on bottom for the 0° head pose): the combinations of seven horizontal ones (0°, ±5°, ±10°, ±15°) and three vertical ones (0°, ±10°). For each of these, we also show a cropped area of the eye region.

Pick's eye contact results and measuring errors in the perception of gaze direction in general. Gamer and Hecht [14] explore the effects of distance, eye occlusion, and the presence of a second head as well, while Martin and Jones [21] examine the effects of distance and lighting intensity from a signal detection standpoint. Symons et al. [35] focus on triadic eye gaze acuity (the ability to judge where someone is looking in space) rather than dyadic eye gaze acuity, and verify that digital photographs are a good substitute for in-person gazers. Gemmell et al. [15] and Chen [8] explore how the design of videoconferencing systems can promote gaze awareness without using special-purpose hardware.

Gaze-Based Interactive Systems

The pursuit of more natural, ubiquitous user interfaces has been an important goal for the HCI community. A new class of user interfaces, called *attentive* user interfaces, aims to facilitate more social interactions between users and devices by treating users' attention as a valuable resource [38]. In doing so, they must (a) sense users' attention, (b) make inferences about what users want to do, and (c) negotiate "turns" amongst themselves, as Vertegaal and Shell describe [39].

To date, however, these interfaces have been limited by current gaze tracking techniques, making it difficult to sense users' attention. Although sensing attention through eye contact alone would be ideal, many systems incorporate either gesture-based control [5], manual input [43], or intrusive head-mounted cameras [32] as a workaround. Shell et al., however, did in fact propose standalone eye contact sensors [31], but they use active infrared illumination. They also interfere with each other if placed within 80° of visual angle of each other. They extend Morimoto et al.'s PupilCam design [25] (which locates pupils by reflecting infrared light on them) by comparing the location of a corneal glint (i.e., the first Purkinje image) with that of the pupil reflection. Our gaze locking technique is completely passive, is not prone to interference, and is accurate at long range. We compare the two techniques directly in the Experiments section.

Omron's commercial OKAO Vision system [28] includes a passive gaze tracker, and Ye et al. [42] combine this with an active and intrusive head-mounted camera in order to determine mutual gaze (simultaneous eye contact) between the person wearing the camera and another person. However, Ye at al. find OKAO Vision's gaze tracker to be inaccurate and sensitive to head pose. Their resulting system has an MCC of 0.72, was only tested on one pair of subjects, and was not shown to work over long distances. Our gaze locking approach is accurate, passive, non-intrusive, robust to head pose, and achieves an MCC of over 0.83 at distances of 18 m.

GAZE LOCKING IN PEOPLE

People seem to have an uncanny ability to tell when others are looking at them. In this section, we describe a two-part experiment that we performed to show how accurate people really are. First, we created a gaze data set, then we asked a set of "players" to determine which of those images are gaze locking and which ones are not. Our experiment revealed some very interesting trends in human vision, and we used those to guide the design of our gaze locking approach.

Creating a Gaze Data Set

Data Set Statistics

Our data set contains a total of 5,880 high-resolution images of 56 different people (32 male, 24 female), and each image has a resolution of 5,184×3,456 pixels. 21 of our subjects were Asian, 19 were White, 8 were South Asian, 7 were Black, and 4 were Hispanic or Latino. Our subjects ranged from 18 to 36 years of age, and 21 of them wore prescription glasses.

	McMurrough et al.	Gi4E	Weidenbacher et al.	Our Data Set
# Subjects:	20	103	20	56
# Fixed Gaze Targets per Subject per Head Pose:	16	12	2-9	21
# Fixed Head Poses:	1	N/A	19	5
Head Pose Calibration?	N/A	No	Yes	Yes
Resolution (px):	768×480	800×600	1600×1200	5184×3456
Total # Images:	N/A	1,236	2,220	5,880

Figure 3. Gaze data set comparison. A comparison of our gaze data set with ones recently made for gaze tracking. McMurrough et al.'s data set is video-based and includes precise head pose measurements rather than simply calibrating head pose. The Gi4E data set does not stabilize subjects' head pose. Weidenbacher et al.'s data set offers a wide variety of fixed head poses, but many have only two corresponding gaze directions.

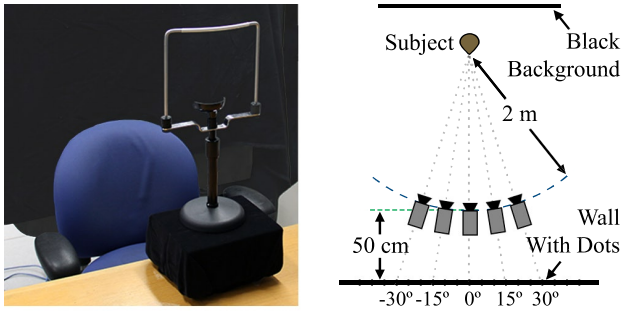

Figure 4. Setup for image capture. (a) Subjects were seated in front of a black background and used a chin rest to stabilize their face. (b) We captured images from five different camera positions asynchronously. Each position represented a (horizontal) head pose. The subjects focused on a grid of dots placed on the wall behind each camera location.

As shown in Figure 2, for each subject, we acquired images for each combination of five horizontal head poses (0°, ±15°, ±30°), seven horizontal gaze directions (0°, ±5°, ±10°, ±15°), and three vertical gaze directions (0°, ±10°). Note that this means we collected five gaze locking images (0° vertical and horizontal gaze direction) for each subject, one for each head pose. Figure 3 compares our gaze data set with data sets recently made by McMurrough et al. [23], Ponz et al. [29], and Weidenbacher et al. [41] for gaze tracking.

Collection Procedure

We recorded each image with a Canon EOS Rebel T3i camera and a Canon EF-S 18–135 mm IS f/3.5–5.6 zoom lens. As shown in Figure 4, subjects were seated in a fixed location in front of a black background, and a grid of dots was attached to a wall in front of them. The dots were placed in 5° increments horizontally and 10° increments vertically. There were five camera positions marked on the floor (one for each head pose), and each position was 2 m from the subject. The dots were organized in such a way that each camera position had a corresponding 7×3 grid.

The subjects used a height-adjustable chin rest to stabilize their face and position their eyes 70 cm above the floor. The camera was placed at eye height, as was the center row of dots. For each subject and head pose (camera position), we took three to six images of the subject gazing (in a raster scan fashion) at each dot of the pose's corresponding grid of dots. To ensure the subject was in focus, not blinking, and looking in the correct direction, we viewed each image at full resolution afterwards and kept the best one from each set of three or six.

Human Accuracy

Experimental Setup

After creating our gaze data set, we asked 52 "players" (27 male, 25 female) to play a computer-based quiz to determine which of those images are gaze locking and which of those are not. The players were all paid volunteers and were mostly university students. We asked the players to state whether or not the subject in each image was looking directly at him or her, a simple yes/no response. Each gaze-locking image was seen by an average of 8.8 players and each non-gaze-locking image was seen by an average of 3.96 players. Each player participated in one 40-minute session, viewing 440 images in the process.

The players viewed our images on a computer monitor, so we needed the subjects to appear at the same resolution as they would if seen in person for the results to be accurate. In addition, each image was captured at a distance of 2 m from the subject, so we created four more copies of each image to serve as a proxy for distances of 6 m, 10 m, 14 m, and 18 m. The appendix describes how we scaled the images—taking the acuity of the human eye into account—to solve both of these problems. We did not find a statistical difference in people's accuracy when we used a small sample of "true" 6-18 m images instead of our scaled images.

Human Accuracy Results

Figure 5 highlights some of our observations. We use the Matthews correlation coefficient (MCC) [22] to represent accuracy in Figure 5(a) and Figure 5(b) since it is widely used in machine learning for assessing binary classification performance with uneven class sizes. An MCC of 1.0 represents perfect classification, an MCC of -1.0 represents completely incorrect classification, and an MCC of 0.0 represents classification that is no better than random guessing. The MCC is not well-defined when only one class of data is used, so we use percentage accuracies in Figure 5(c).

In Figure 5(a), we find that humans are indeed rather adept at determining when others are looking at them. At distances of 18 m, their MCC can still surpass 0.2. Moreover, even though the size of someone else's face decreases quadratically over distance, humans' gaze locking accuracy decreases only linearly with distance.

In Figure 5(b), we see that humans' accuracy is largely maintained across different head poses, even extreme ones such as ±30° to the side. Shechtman et al. [30] find that a 50° ocular duction (i.e., eye movement) is nearly impossible for many age groups, and we found during our experiments that people are uncomfortable moving their eyes 30° to the side. Hence,

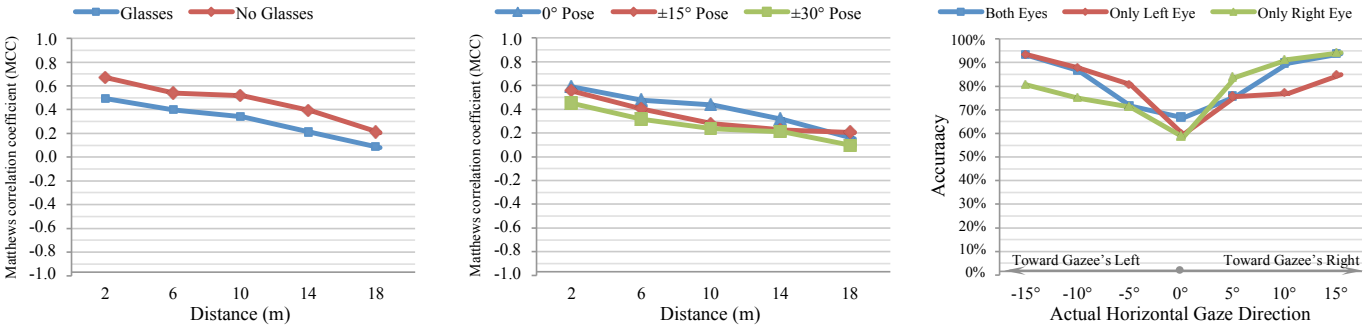

Figure 5. Gaze locking in people. (a) People are relatively accurate at sensing eye contact, even when the person gazing (i.e., the gazer) is wearing prescription glasses. At distances of 18 m, gazees still achieve MCCs of over 0.2 if the gazer is not wearing glasses. Here, the gazer is at a frontal (0°) head pose. (b) The gazee's accuracy decreases roughly linearly over distance regardless of the gazer's (horizontal) head pose. Head poses that are more off-center (such as ±30°) have slightly lower MCCs. (c) The gazees are least accurate when the gazer is actually looking at them (the 0° case)—that is, the false negative rate is higher than the false positive rate. Interestingly, if the gazer is looking away, the gazee is more accurate when he or she can only see one of the gazer's eyes (the blue line is not strictly above the red and green lines). Each accuracy measurement was calculated over all five distances and head poses. Here, we use percentage accuracies instead of MCCs because each horizontal gaze direction besides 0° is always non-gaze locking by definition and the MCC is not well-defined when only one class of data is used.

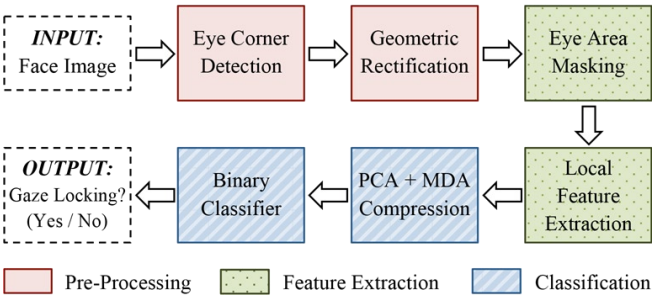

Figure 6. Gaze locking detector pipeline. Our gaze locking detector is comprised of three broad phases, shown here in different colors. In the first phase, we locate the eyes in an image and transform them into a standard coordinate frame. In the second phase, we mask out the eyes' surroundings and assemble pixel-wise features from the eyes' appearance. Finally, we project these features into a low-dimensional space, then feed them into a binary classifier to determine whether the face is gaze locking or not.

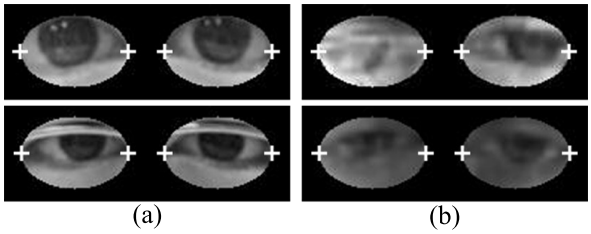

Figure 7. Rectified features and failure cases. (a) Examples of rectified and masked features. Each eye has been transformed to a 48×36 px coordinate frame. The crosshairs signify eye corners detected in the first phase. We mask each eye with a fixed-size ellipse whose shape was optimized offline for accuracy. (b) Two failure cases: strong highlights on glasses (top) and low contrast (bottom).

eye contact from even a ±30° head pose is unlikely to happen in everyday life.

Lastly, Figure 5(c) shows the results from an extended test that we performed with 10 players. In this test, the players viewed images whose left or right half was cropped off, showing only one of the subject's eyes, in addition to non-cropped images. Interestingly, we found that the players are often more accurate when they can only see one of the other person's eyes. This is the case when the subject is looking away and the visible eye is the one that is looking more off-center.

SAMPLE DETECTOR

Here, we show a simple, lightweight detector design that is nonetheless accurate and robust. It runs at over 20 FPS on a computer with an Intel Core i5-3470 processor, 8 GB of RAM, and an NVIDIA GeForce GTX 660M graphics card.

Given an image, the detector outputs binary decisions that indicate whether each face in the image is gaze locking or not. It is composed of three broad phases, shown in Figure 6. We describe each phase in the subsections below.

Pre-Processing Phase

In the first phase, we locate the eyes in an image and transform them into a standard coordinate frame. We find the eyes by taking the eye corner locations output from a commercial face and fiducial point detector [28]. At this point, eye shape varies greatly due to differences in head pose. To remove the influence of head pose, we rectify each eye via an affine transformation to a 48×36 px coordinate frame, then concatenate the two eye regions together to form a 96×36 px image. The low resolution of this image reflects the low resolution encompassed by objects at long distances, making our detector accurate over long distances. Figure 7 shows several examples of rectified features.

Feature Extraction Phase

The most difficult part of using the eyes' appearance for classification purposes is the inherent variance in the eyes' appearance. Both the eyes' shape and degree of openness significantly affect their appearance, even after performing the affine transformation in the first phase. Training our detector with a large number samples helps account for this, but we

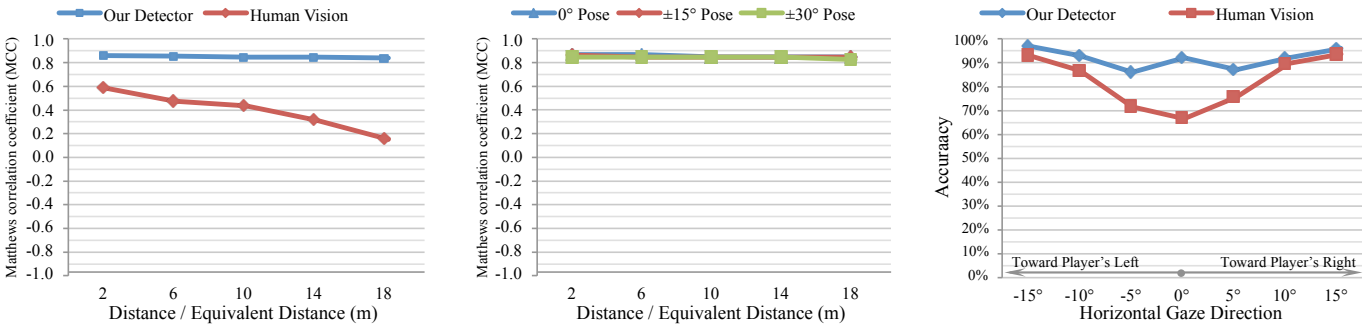

Figure 8. Gaze locking detector performance. As the appendix describes, we downsampled our detector's test images to match the resolution seen by the human fovea at the respective distances. (a) Our detector achieves MCCs of over 0.83 at a distance of 18 m, significantly outperforming humans' accuracy. The detector's accuracy is fairly constant over distance because our method uses features that are of very low resolution. The line representing human performance is an aggregation of the lines from Figure 5(a). (b) Our detector's accuracy is also fairly constant over a variety of (horizontal) head poses. (c) As with human vision, our detector's accuracy is worst when people are looking at or very close to the camera. Our detector significantly outperforms human vision nonetheless. Each accuracy measurement was calculated over all five distances and head poses. As with Figure 5(c), we use percentage accuracies here because the MCC is not well-defined when only one class of data is used.

take the additional step of masking out the areas around the eyes to remove the influence of their variances in appearance.

Our mask (Figure 7) is a fixed-size ellipse whose major axis lies on the line segment connecting the two eye corners. Choosing the size is nontrivial: a larger ellipse reveals more of the eye's surroundings and more information about gaze, but a smaller ellipse is more robust to noise from the surroundings. We used a brute-force search of all possible major and minor axis lengths offline to choose the best size. We chose the values that achieved the best accuracy in our set of training data, which is separate from our testing data.

After applying the mask, we concatenate the remaining pixels' intensity values into a high-dimensional feature vector, then normalize the feature vector to unit magnitude. This unit-magnitude feature vector is our final representation of the eyes' appearance.

Classification Phase

In the final phase, we project the high-dimensional feature vector onto a low-dimensional space via principal component analysis (PCA) [37] and multiple discriminant analysis (MDA) [11], then feed the projected vector into a support vector machine (SVM) [7] that we trained offline. The SVM decides whether the face is gaze locking or not. Since gaze locking is a binary classification problem rather than a continuous one, it is more robust to noise and requires fewer training samples. The Binary Classifier subsection describes the training process.

PCA + MDA Compression

Dimensionality reduction is a common task in appearance matching. It boosts classification speed, removes redundancies in the representations of features, avoids over-fitting, and reduces the effects of noise on classification. Hence, we use PCA to compress our feature vector to roughly 200 dimensions. Afterwards, we employ MDA to form a highly discriminative subspace, compressing our feature vectors even more. We find that a six-dimensional subspace, used to separate seven distinct classes of data, yields the highest accuracy. One class corresponds to gaze locking images and the rest correspond to non-gaze-locking images. This is likely because our training data set comprises seven horizontal gaze directions, one of them gaze locking (0°) and the rest non-gaze-locking.

Binary Classifier

In our sample detector, we use a linear SVM classifier [7] with default parameters (which includes a radial basis function kernel) to output our final binary decision. Even though we use an SVM, any binary classifier (e.g., LDA or neural networks) would work. The kernel allows input features similar to our positive training samples to be "lifted airborne," so to speak, separating them from ones near our negative samples that are still "on the ground." As was the case with the data set statistics described earlier, the number of positive and negative training samples we had was highly unbalanced (280 gaze locking images and 5,600 non-gaze-locking images), so we randomly perturbed our training data to generate 5,000 additional gaze locking samples and 15,000 additional non-gaze-locking samples. We accomplished this by making small, random adjustments to the resolution and detected eye corner positions of our training images.

EXPERIMENTS

We tested our gaze locking detector via leave-one-out cross-validation on a modified version of our gaze data set. The original data set, described earlier in the Creating a Gaze Data Set subsection, comprises five head poses and 21 gaze directions but was captured with a high-resolution camera from a distance of 2 m. The modified data set comprises five downsampled copies of each of the original data set's 5,880 images, where the resolution of each copy matches that seen by the human retina at distances of 2 m, 6 m, 10 m, 14 m, and 18 m. Equation 1 in the appendix describes how we downsampled the images. Note that even the 2 m image is downsampled by a factor of 58.0%. Our supplementary video shows how

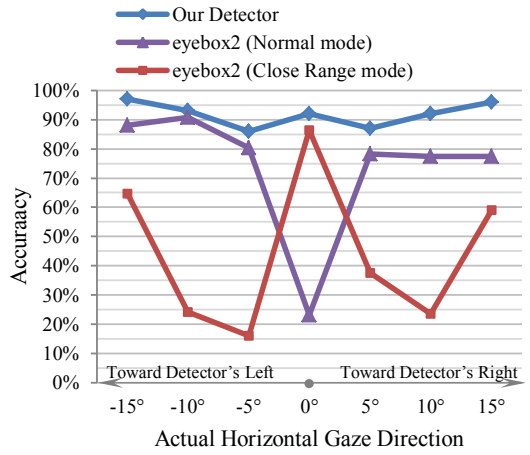

Figure 9. Comparison with an active system. A comparison of our sample detector with an eyebox2, which implements Shell et al.'s active approach to eye contact detection, in both Normal (6 m) and Close Range (2 m) modes. Though passive, our detector is more accurate than the eyebox2. The eyebox2's Normal mode seems to be tuned toward reducing false positives, and its Close Range mode seems to be tuned toward reducing false negatives.

our detector performs on raw footage from webcams and iPad video feeds.

Comparison with Human Vision

Figure 8 shows our sample detector's performance and compares it with human vision's performance (from Figure 5). We again use the Matthews correlation coefficient (MCC) [22] to represent accuracy since it is widely used in machine learning for assessing binary classification performance with uneven class sizes. Although our detector uses a very standard set of tools, it achieves an MCC of over 0.83 at a distance of 18 m, significantly outperforming the 0.15 MCC of human vision (Figure 8(a)). This high accuracy over long distances is a result of using very low-resolution feature vectors. As we see in Figure 8(b), this accuracy is maintained across different horizontal head poses, even fairly extreme ones such as $\pm 30°$.

Figure 8(c) shows our detector's accuracy with respect to a person's actual gaze direction. As with human vision, our detector is least accurate when a person is looking at or very near the camera (i.e., at the borderline between gaze locking and "almost gaze locking"). However, even in this case, our detector is much more accurate than human vision (86% versus 67%). We use percentage accuracies instead of MCCs here because each horizontal gaze direction besides 0° is always non-gaze locking by definition and the MCC is not well-defined when only one class of data is used.

Comparison with an Active System

Here, we compare our sample detector's accuracy with that of the eyebox2 [12], a leading commercial implementation of Shell et al.'s active approach to eye contact detection. Recall that our approach is completely passive and hence does not use active illumination or special hardware like the eyebox2 does. The eyebox2 is specified to work best at a range of 5–10 m in Normal mode and 1.3–3.3 m in Close Range mode, so we asked six people to sit (indoors using a chin rest) 6 m in front of it in Normal mode and 2 m in front of it in Close Range mode. They stared at the eyebox2 and six dots placed horizontally around it (at $\pm 5°$, $\pm 10°$, and $\pm 15°$) for ten seconds apiece. To analyze the eyebox2's accuracy, we measured the proportion of time the eyebox2 claimed they were making eye contact.

Figure 9 shows the results. Our sample detector is more accurate than the eyebox2 regardless of the actual gaze direction. In Normal mode, the eyebox2 seems tuned toward reducing false positives—although we adjusted its illuminator position, threshold setting, and focus setting as the manual instructed, its output was very jittery in the gaze locking case. We also found its range for reliable tracking to be around 5–7 m, and that it does not work well for people with glasses (the participants represented in Figure 9 did not wear glasses). In Close Range mode, the eyebox2 seems to be tuned toward reducing false negatives—it usually claims that people are looking at it unless they are looking at least 15° away. Our gaze locking approach works for a greater range of distances (at least 2–18 m) without separate modes of operation, is more robust to eyeglasses, and can be applied to any image, including existing images.

Failure Cases

As with all appearance-based recognition systems, our approach can be prone to errors when a feature's visual appearance (in our case, that of the eyes) differs significantly from that of average features. For example, even though 22 of 56 subjects in our training data set wore glasses, our detector may not work well for all types of glasses over all head poses, given the large variety. This, however, is also true for active techniques that rely on reflections. Our approach is also prone to errors when the eyes are severely occluded (e.g., if a person's hair blocks an eye), when the illumination is extreme (e.g., strong highlights or profile illumination), or when there is very low contrast in the image. Figure 7(b) highlights two of our sample detector's failure cases.

APPLICATIONS OF GAZE LOCKING

We now demonstrate a few of the applications that gaze locking facilitates. For each of these applications, we recorded video feeds from the respective devices (iPads, webcams, and a DSLR camera) and ran our detector on them offline. Our accompanying video shows the output of our detector on the feeds and images from Figures 10–13.

Human–Object Interaction

Cameras are becoming increasingly small and inexpensive to produce. By embedding cameras in everyday devices and objects, the devices and objects can be selected or activated simply by looking at them.

As an example, Figure 10 shows a proof of concept system that we created with 3rd generation iPads. We process the videos from the iPads' built-in cameras (which have a 640×480 px resolution) to sense when they are looked at,

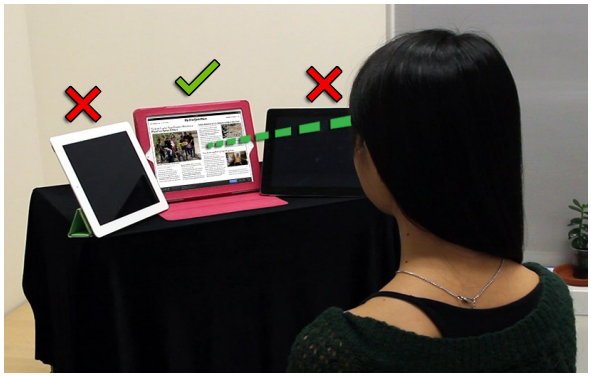

Figure 10. Human–object interaction. Our gaze locking approach allows people to interact with objects just by looking at them. In this proof of concept, we process the videos from the embedded cameras of three iPads to sense when the iPads are being looked at. Here, the woman is looking at the iPad in the middle. Since the iPads' cameras are on their extreme left, she was instructed to look at the iPads' left halves. Our accompanying video shows our detector's output on the actual video feeds.

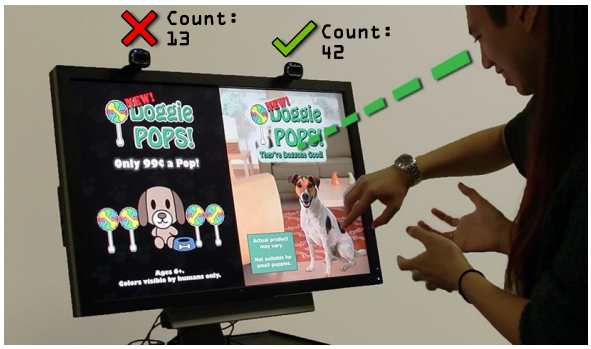

Figure 11. User analytics. Two ordinary webcams are placed above two ads for the same product. By counting the number of times each advertisement is viewed, we can gauge which one is more effective. The counts incremented when the viewers looked at the ads' top halves. Our accompanying video shows our detector's output on the actual video feeds.

then display relevant content such as news headlines or reading lists. As another example, a museum exhibit or department store item could be rigged with a small camera to inform passersby about them when they look at them.

Some smartphones (e.g., the Samsung Galaxy S III and the Samsung Galaxy S4) already include "smart pause" and "smart scroll" features to pause videos by looking away from the phone and scroll documents by looking up and down. However, we found that both features on the Galaxy S4 seem to work reliably only when a user moves his or her entire head, although the system also responded sometimes to large eye motions alone. Our technique can distinguish eye contact from a subtle $\pm 5^\circ$ gaze away, and it works over long distances.

User Analytics

Several commercial systems [12, 13] embed cameras in product displays or advertisements as a means of measuring consumer attention, but these systems employ active infrared illumination. As Figure 11 shows, our method offers a completely passive alternative that is robust to distance and does not require special hardware.

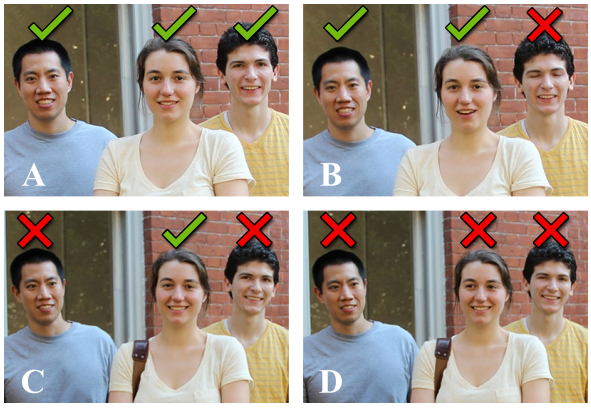

Figure 12. Image search filter. Our approach is completely appearance-based and can be applied to any image, including existing images such as ones from the Web. Hence, we can sort these images (A–D) by degree of eye contact to quickly find one where everyone is looking at the camera. These are actual decisions made by our detector.

Figure 13. Gaze-triggered photography. By incorporating a gaze locking detector in a consumer-level camera, the camera could automatically take a picture when the entire group is looking straight back, allowing the photographer to join the group and still capture a perfect photo. Our accompanying video shows our detector's output on the camera's feed.

As with the commercial systems, our method has a reasonably sized tolerance for what it considers gaze locking, allowing it to work for cameras placed adjacent to regions of interest as well. Our detector was trained to distinguish between 0° and $\pm 5^\circ$ horizontally and between 0° and $\pm 10^\circ$ vertically, so we estimate its tolerance to be roughly 2.5° in either direction horizontally and 5° in either direction vertically. This corresponds to a 8.7×17.5 cm target at a distance of 1 m and a 43.7×87.5 cm target at a distance of 5 m.

Image Filtering

Unlike active methods, our method can be used to detect eye contact in existing images such as ones from the Internet. There are billions of images on the Internet, and over 300 million photos are uploaded to Facebook alone each day [20]. Hence, as Figure 12 shows, we can sort and filter photos by degree of eye contact with our method to improve image search.

Gaze-Triggered Photography

With today's cameras, taking a group photo can be difficult since everyone must be looking at the camera at the right mo-

ment. With our technology, however, cameras can incorporate a gaze trigger that works as follows: the photographer would initiate the function, then join the group as the camera waits to see another face (the photographer's) enter the frame. As soon as this is detected, the camera would take a picture when the entire group is looking straight back. Figure 13 and our supplementary video demonstrate this concept.

Already, many consumer-level cameras (e.g., the Sony Cybershot W650) feature an anti-blink function that helps users capture photos when subjects are not blinking. Other cameras (e.g., the Canon PowerShot XS160 IS) also include face self-timers that release the shutter only when an additional face (the photographer's) enters the frame. By sensing eye contact instead of simply sensing blinking or the presence of faces, our technology can make cameras aware of when people are actually looking straight back.

DISCUSSION

Contributions

In this work, we have created a passive approach for sensing eye contact from a live camera or an existing still image or video recording and demonstrated several of the applications that it facilitates, such as human–object interaction and gaze-triggered photography. We also performed a study on how accurately humans can perform the same task, finding several interesting results. Lastly, we created a large gaze data set. Unlike existing gaze tracking approaches, our approach exploits the special appearance of direct eye gaze, making it largely robust to distance and pose, even though it is passive, non-intrusive, and calibration-free. Furthermore, it does not require any special hardware.

Toward embeddable gaze lockers

There is great potential for future work in gaze locking using embedded cameras. Our sample detector consists of fairly simple mathematical operations, so future efforts could create a "gaze locker"—a camera module with a system-on-chip for gaze locking. This gaze locker would be small, cheap, and computationally efficient. Systems-on-chip already exist in cameras for applications such as exposure compensation and image compression.

Moreover, our gaze locking approach is passive and, as a result, energy-efficient. Hence, gaze lockers may even be able to employ energy-harvesting techniques like RFID tags do. We could also use gaze lockers to aid the blind and deaf by sensing when others are looking at them. Furthermore, if we place gaze lockers in many objects, they would collectively form a cloud that serves as a ubiquitous gaze tracker that is accurate over distance. For all of these reasons, we believe gaze lockers could be the perfect platform for bringing gaze-based interactive systems into everyday use in the future.

ACKNOWLEDGMENTS

This research was supported in part by ONR Award No. N00014-11-1-0285. Brian A. Smith was supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

1. Argyle, M., and Dean, J. Eye-contact, distance and affiliation. *J. Sociometry 28*, 3 (1965), 289–304.
2. Baluja, S., and Pomerleau, D. Non-intrusive gaze tracking using artificial neural networks. Tech. rep., Department of Computer Science, Carnegie Mellon University, 1994.
3. Beymer, D., and Flickner, M. Eye gaze tracking using an active stereo head. In *Proc. CVPR 2003*, IEEE Press (2003).
4. Blackwell, H. R. Contrast thresholds of the human eye. *J. Opt. Soc. Am. 36*, 11 (1946).
5. Bolt, R. A. Put-that-there: Voice and gesture at the graphics interface. In *Proc. SIGGRAPH 1980*, ACM Press (1980), 262–270.
6. Cadavid, S., Mahoor, M., Messinger, D., and Cohn, J. Automated classification of gaze direction using spectral regression and support vector machine. In *Proc. ACII 2009*, IEEE Press (2009), 1–6.
7. Chang, C., and Lin, C. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. 2*, 3 (2011), 27:1–27:27.
8. Chen, M. Leveraging the asymmetric sensitivity of eye contact for videoconferencing. In *Proc. CHI 2002*, ACM Press (2002), 49–56.
9. Cline, M. G. The perception of where a person is looking. *Am. J. Psychol. 80*, 1 (1967), 41–50.
10. Curcio, C. A., Sloan, K. R., Kalina, R. E., and Hendrickson, A. E. Human photoreceptor topography. *J. Comp. Neurol. 292*, 4 (1990), 497–523.
11. Duda, R., Hart, P., and Stork, D. *Pattern classification*, vol. 2. Wiley-Interscience, 2001, 114–124.
12. eyebox2™ impressions™ for signage. **https://www.xuuk.com/Images/eyebox2productspecs.pdf**.
13. EyeTech™ VT2 XL. **http://www.eyetechds.com/vt2-xl.shtml**.
14. Gamer, M., and Hecht, H. Are you looking at me? measuring the cone of gaze. *J. Exp. Psychol. [Hum Percept.] 33*, 3 (2007), 705–715.
15. Gemmell, J., Toyama, K., Zitnick, C., Kang, T., and Seitz, S. Gaze awareness for video-conferencing: a software approach. *IEEE Multimedia 7*, 4 (2000), 26–35.
16. Gibson, J. J., and Pick, A. D. Perception of another person's looking behavior. *Am. J. Psychol. 76*, 3 (1963), 386–394.
17. Hansen, D., and Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell. 32*, 3 (2010), 478–500.
18. Hansen, D. W., and Pece, A. E. C. Eye tracking in the wild. *Comput. Vis. Image Und. 98*, 1 (2005), 155–181.
19. Kendon, A. Some functions of gaze direction in social interaction. *Acta Psychol. 26* (1967), 22–63.
20. Kiss, J. Facebook hits 1 billion users a month. **http://www.guardian.co.uk/technology/2012/oct/04/facebook-hits-billion-users-a-month**, Oct 2012.
21. Martin, W. W., and Jones, R. F. The accuracy of eye-gaze judgement: A signal detection approach. *Brit. J. Soc. Psychol. 21*, 4 (1982), 293–299.
22. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA-Protein Struct." 405*, 2 (1975), 442–451.
23. McMurrough, C. D., Metsis, V., Rich, J., and Makedon, F. An eye tracking dataset for point of gaze detection. In *Proc. ETRA 2012*, ACM Press (2012), 305–308.
24. Morimoto, C., Amir, A., and Flickner, M. Detecting eye position and gaze from a single camera and 2 light sources. In *Proc. ICPR 2002*, IEEE Press (2002), 314–317.
25. Morimoto, C., Koons, D., Amir, A., and Flickner, M. Pupil detection and tracking using multiple light sources. *J. Image Vision Comput. 18*, 4 (2000).

26. Morimoto, C., and Mimica, M. Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Und. 98*, 1 (2005), 4–24.
27. Nishino, K., and Nayar, S. The world in an eye. In *Proc. CVPR 2004*, IEEE Press (2004), 444–451.
28. Omron. OKAO vision. **http://www.omron.com/r_d/coretech/vision/okao.html**.
29. Ponz, V., Villanueva, A., and Cabeza, R. Dataset for the evaluation of eye detector for gaze estimation. In *Proc. UbiComp 2012*, ACM Press (2012), 681–684.
30. Shechtman, D., Riordan-Eva, P., and Hardigan, P. Maximum angle of ocular duction during visual fixation as a function of age. *Strabismus 13*, 1 (2005), 21–26.
31. Shell, J. S., Vertegaal, R., Cheng, D., Skaburskis, A. W., Sohn, C., Stewart, A. J., Aoudeh, O., and Dickie, C. ECSGlasses and EyePliances: using attention to open sociable windows of interaction. In *Proc. ETRA 2004*, ACM Press (2004), 93–100.
32. Smith, J. D., Vertegaal, R., and Sohn, C. ViewPointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis. In *Proc. UIST 2005*, ACM Press (2005), 53–61.
33. Stiefelhagen, R., Yang, J., and Waibel, A. A model-based gaze tracking system. In *Proc. IEEE Intelligence and Systems 1996*, IEEE Press (1996), 304–310.
34. Stiefelhagen, R., Yang, J., and Waibel, A. Tracking eyes and monitoring eye gaze. In *Proc. PUI 1997* (1997).
35. Symons, L. A., Lee, K., Cedrone, C. C., and Nishimura, M. What are you looking at? acuity for triadic eye gaze. *J. Gen. Psychol. 131*, 4 (2004), 451–469.
36. Tan, K.-H., Kriegman, D. J., and Ahuja, N. Appearance-based eye gaze estimation. In *Proc. WACV 2002*, IEEE Press (2002), 191–195.
37. Turk, M., and Pentland, A. Face recognition using eigenfaces. In *Proc. CVPR 1991*, IEEE Press (1991), 586–591.
38. Vertegaal, R. Attentive user interfaces. *Comm. ACM 46*, 3 (2003), 31–33.
39. Vertegaal, R., and Shell, J. Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects. *Soc. Sci. Inf. 47*, 3 (2008), 275–298.
40. Vertegaal, R., Slagter, R., van der Veer, G., and Nijholt, A. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proc. CHI 2001*, ACM Press (2001), 301–308.
41. Weidenbacher, U., Layher, G., Strauss, P.-M., and Neumann, H. A comprehensive head pose and gaze database. In *Proc. IE 2007*, IET Press (2007).
42. Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G. D., and Rehg, J. M. Detecting eye contact using wearable eye-tracking glasses. In *Proc. UbiComp 2012*, ACM Press (2012), 699–704.
43. Zhai, S., Morimoto, C., and Ihde, S. Manual and gaze input cascaded (MAGIC) pointing. In *Proc. CHI 1999*, ACM Press (1999), 246–253.

APPENDIX

Scaling Data Set Images

In our human test, the players viewed our images in front of a computer screen, but we needed the subjects in the images to appear as they would in person for the results to be accurate. We also needed to represent a variety of distances (2 m to 18 m) properly for both the human and sample detector experiments. Hence, we took the parameters of our camera, computer monitor, and even the acuity of the human eye into account to scale the images accordingly and display them at the proper resolution. The calculations are described here.

Human Test

An "eye pixel" is the smallest area of the human fovea that can distinguish a point or a line pair [4, 10]. In our human experiments, if the player were to view the subject directly from a distance of d_o, the subject's face would subtend E of the player's eye's pixels in width, where:

$$E = \frac{w}{2d_o \tan(\theta_e/2)}. \tag{1}$$

w is the width of the subject's face (usually ≈14 cm). θ_e is the angular resolution of the human eye fovea, and is roughly 0.3 arc-minutes (or 0.005° per eye pixel) [4, 10].

When the subject is captured by a camera instead, his or her face subtends C camera pixels in width, where:

$$C = \frac{P_c w}{w_c(u/f - 1)}. \tag{2}$$

P_c is the camera's horizontal pixel count, w_c is the width of the image sensor, f is the camera's focal length, and u is the distance from the subject to the camera.

Then, if the player views the captured image on a screen, the subject's face would subtend S eye pixels in width:

$$S = \frac{\alpha C w_s}{2d_s P_s \tan(\theta_e/2)}, \quad \text{where} \quad \alpha \leq 1. \tag{3}$$

α is the factor by which the image dimensions are scaled on the screen (1 represents 100%), P_s is the screen's horizontal pixel count, w_s is the screen's width, and d_s is the distance from the player to the screen.

For the player to view the subject on the screen without a loss in resolution compared to seeing the subject in person, both C and S must be greater than or equal to E. This was true in our configuration, in which u = 2 m, f = 85 mm, P_c = 5184 px, d_c = 22.3 mm, P_s = 2560 px, d_s = 664 mm, D = 508 mm, and the corresponding (d_o, α) pairs (i.e., the scale factors we used to represent each distance) were (2 m, 19.5%), (6 m, 6.5%), (10 m, 3.9%), (14 m, 2.8%), and (18 m, 2.2%).

Sample Detector Experiments

When testing the accuracy of our sample detector, we wanted the detector to "see" each image as if they were being viewed by a person at the appropriate distance d_o. That is, the subject's face in each image should be of the same resolution that the human retina would see it at a distance of d_o. Hence, we can simply downsample our data set's images to a resolution of E via Equation 1 to get versions of them that correspond to different distances d_o.