Perceptual User Interfaces using Vision-based Eye Tracking

Ravikrishna Ruddarraju†Antonio Haro‡, Kris Nagel‡, Quan T. Tran‡
Irfan A. Essa†‡, Gregory Abowd‡, Elizabeth D. Mynatt‡

†School of Electrical and Computer Engineering and ‡GVU Center, College of Computing
Georgia Institute of Technology, Atlanta, Georgia 30332. USA.
{ravigtri,haro,kris,quantt,irfan,abowd,myantt}@cc.gatech.edu

ABSTRACT

We present a multi-camera vision-based eye tracking method to robustly locate and track user's eyes as they interact with an application. We propose enhancements to various vision-based eye-tracking approaches, which include (a) the use of multiple cameras to estimate head pose and increase coverage of the sensors and (b) the use of probabilistic measures incorporating Fisher's linear discriminant to robustly track the eyes under varying lighting conditions in real-time. We present experiments and quantitative results to demonstrate the robustness of our eye tracking in two application prototypes.

Categories and Subject Descriptors

I.4.m [**Image Processing and Computer Vision**]: Miscellaneous; I.4.0 [**Image Processing and Computer Vision**]: General—*Image Processing Software* ; G.3 [**Probability and Statistics**]: Probabilistic algorithms

General Terms

Algorithms, Human Factors, Design

Keywords

Eye Tracking, Multiple Cameras, Fisher's Discriminant Analysis, Computer Vision, Human Computer Interaction

1. INTRODUCTION

Head pose and eye gaze information are very valuable cues in face-to-face interactions between people. Such information is also important for computer systems that a person intends to interact with. The awareness of eye gaze provides context to the computer system that the user is looking at it and therefore supports effective interaction with the user.

In this paper, we present a real-time vision-based eye tracking system to robustly track a user's eyes and head movements. Our system utilizes robust eye tracking data from multiple cameras to estimate 3D head orientation via triangulation. Multiple cameras also afford a larger tracking volume than is possible with an individual sensor, which is valuable in an attentive environment. Each individual eye tracker exploits the red-eye effect to track eyes robustly using an infrared lighting source. Although our system relies on infrared light, it can still track reliably in environments saturated with infrared light, such as a residential living room with much sunlight.

We also present experiments and quantitative results to demonstrate the robustness of our eye tracking in two application prototypes: the Family Intercom [15], and the Cook's Collage [19]. In the Family Intercom application, eye gaze estimation is used to make inferences about the desire of an elderly parent to communicate with remote family members. In the Cook's Collage application, eye gaze estimates are used to assist a user while cooking in a kitchen as well as to evaluate the effectiveness of the Collage's user interface. In both cases, our tracking system works well under real-world environments, subject to varying lighting conditions, while allowing the user interaction to be unobtrusive yet engaging.

2. RELATED WORK

A significant amount of past research in human factors suggests that humans are generally interested in what they look at [20, 5]. A significant relationship between eye gaze and attention has also been emphasized [3, 4]. In addition, there is evidence to suggest that eye gaze is a very important aspect of day-to-day perceptual-motor activities [2].

Such important observations from real scenarios has prompted researchers to develop and use eye tracking systems in human-machine interface applications. These eye-tracking systems are used to measure visual attention and to facilitate different kinds of user interaction tasks. Most of the technology used to locate and track eyes has been relatively cumbersome and invasive, requiring a user to wear a special apparatus. For example, Aaltonen *et al.* [1] use eye gaze to perform user interaction in basic PC interface tasks. Goldberg *et al.* [8] use eye movements to infer user intent in real-time interfaces.

Recently, there have been a few technologies that support

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

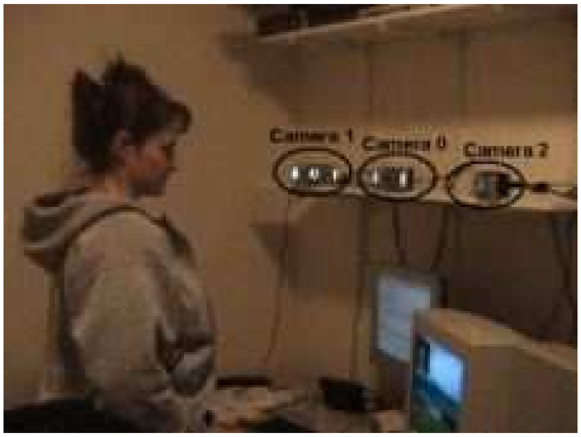


Figure 1: A three camera eye tracking setup.

non-invasive measurement of eye locations and apply them to interaction tasks. Jabrain *et al.* [11] use eye trackers to test the importance of eye gaze in a videoconferencing application. Stiefelhagen *et al.* [18] use eye trackers to estimate user attention in group meetings.

Several researchers have recently proposed stand-alone computer vision-based head pose tracking systems. Harville *et al.* [10] have used linear depth and brightness constraints along with twist mathematics to obtain 3D head pose. Matsumoto *et al.* [12] have used compact 3D face models to estimate head pose. Like Matsumoto *et al.* [12], Schoedl *et al.* [17], and Cascia *et al.* [7] use more complicated polygonal head models for tracking head pose.

Our work significantly differs from the head pose tracking performed in these works in two ways. First, unlike Aaltonen's and Goldberg's head mounted hardware, our system is completely non-invasive. Secondly, our system works in real-time and is significantly more robust than the commercial systems used by Jabrain *et al.*. Unlike the work of Matsumoto *et al.* and Schoedl *et al.*, our algorithms are simple enough that they run well on consumer-level computers without any need for special purpose hardware. Our method is unique in that it tracks using multiple cameras. This ability to track head pose and estimate eye gaze using multiple cameras is very important for large environments where the mobility of users is much greater than the desktop workspace.

3. VISION SYSTEM

Our system uses IBM BlueEyes infrared lighting cameras [13]. These cameras are used as sensors for our eye tracking algorithm and the tracked eyes are used in conjunction to estimate the user's head pose and eye gaze direction (see Figure 1). Each eye tracker utilizes a simple dynamics model of eye movements along with Kalman filters and appearance models to track the eyes robustly in real-time and under widely varying lighting conditions. The tracked head pose is used to estimate a user's eye gaze to measure whether a user is looking at a previously defined region of interest that the prototype applications use to further interact with the user.

Multi-camera IR-based eye tracking

We use several pre-calibrated cameras to estimate a user's head pose. For each camera, we use the tracked eye locations to estimate mouth corners. These two mouth corners and eye positions are then used as low level features between all cameras to estimate the user's 3D head pose. We use a combination of stereo triangulation, noise reduction via interpolation, and a camera switching metric to use the best subsets of cameras for better tracking as a user is moving their head in the tracking volume. Further details of our head pose estimation can be found in [16]. Our results in [16] show that our multiple camera system gives reliable head pose estimates for various users and can be used as a basis for human computer interaction.

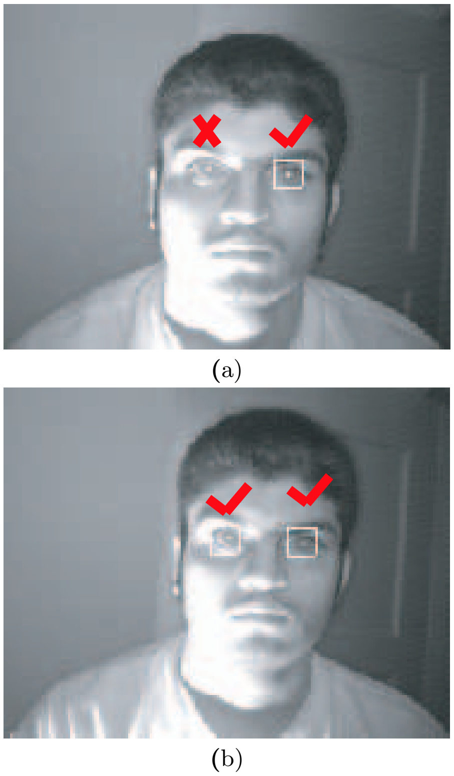
Multiple cameras provide both a large tracking volume as well as 3D head pose information. However, as a user moves in the tracking volume, it is possible that their eyes are no longer visible from some cameras. Our system detects when a user's head is moving away from a camera and uses a different subset of cameras not including it to estimate the 3D pose more accurately. This is done because cameras with only partial views of the face will have increased eye tracking errors since their appearance (and possibly motion) will no longer appear to be eyes. Removing these cameras from the 3D pose calculation is important because the 3D pose is sensitive to noise since we use a small number of cameras [16]. In practice, camera subset switching is not done very often, but must be done for certain head angles to avoid incorrect pose estimates.

Dealing with infrared saturation

Our initial experiments with the system in a residential living room setting showed that infrared light coming from multiple windows during the day was affecting the tracking. The tracker was initially designed to work in indoor office environments with fluorescent lighting and limited daylight. The presence of almost omnidirectional infrared lighting complicates the tracking because there are many additional sources of infrared light besides the LEDs on the camera. Indeed, in an indoor residential setting, infrared light is picked up by the cameras as it is present everywhere during the day.

In the eye tracking sub-system, principal component analysis (PCA) was used to construct appearance models for the eyes [16]. However, PCA weighs all components of the feature vectors equally and cannot account for different noise contributions in different variables. The principal components cannot capture all variation in eyes and non-eyes lit differently throughout the day and from different windows. Multiple classes could be created, but it is unclear how to assign the training data to different classes or what the classes themselves should be to represent varying lighting.

Fisher's linear discriminant has previously been used to compensate for varying light conditions for improved facial recognition [6]. We use it to compute a classification score for candidate eyes versus non-eyes and have found it to yield improved classification over PCA for our data. Figure 2 shows an instance where PCA does not correctly classify an eye in the presence of infrared saturation while Fisher's discriminant does. To train the classifier, we use training data consisting of 64 eyes and 64 non-eyes. The discriminant seeks to find the projection matrix W for the training data



(a)

(b)

Figure 2: Comparison of PCA and Fisher's linear discriminant in eye tracking. (a) PCA: notice that one of the eyes is not detected due to infrared saturation. (b) Fisher's discriminant: both eyes are classified and tracked correctly.

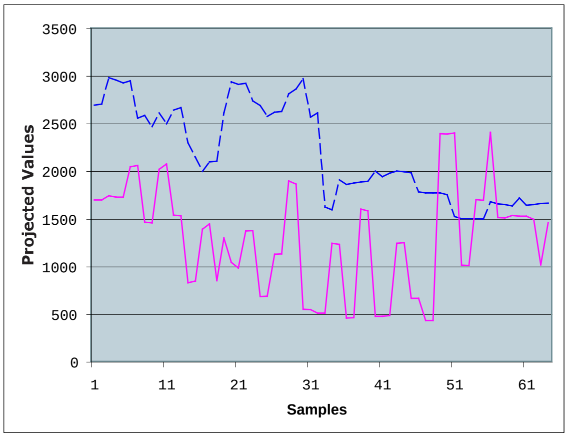


Figure 3: Classifying eyes and non-eyes using Fisher's discriminant. The dotted blue line represents the 'eyes' class and the red line represents the 'non-eyes' class.

x while maximizing $J(W)$:

$$y = Wx \tag{1}$$

$$J(W) = \frac{|\mu_{F_1} - \mu_{F_2}|^2}{\sigma_{F_1}^2 + \sigma_{F_2}^2} \tag{2}$$

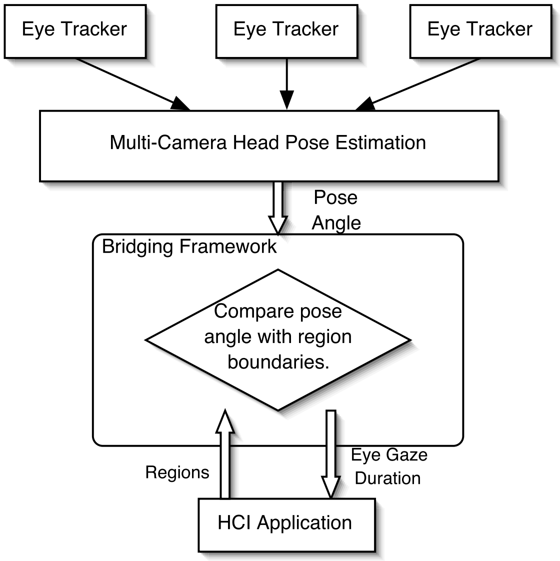


Figure 4: Data flow between the tracking system and applications. Three eye trackers are shown to estimate head pose, though more could be used.

where μ_{F_1}, σ_{F_1}, and μ_{F_2}, σ_{F_2} are the mean and standard deviation of the projected training images into two classes (eyes and non-eyes) onto a line (the Fisher space). To maximize this equation and compute J as an explicit function of W, the ratio between the between-class scatter (S_b) and within-class scatter (S_w) is maximized:

$$S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{3}$$

$$S_w = \sum_{i=0}^{N} [(x_i - \mu_1)(x_i - \mu_1)^T + (x_i - \mu_2)(x_i - \mu_2)^T] \tag{4}$$

$$W = \arg\max_{W} \frac{\|W^T S_b W\|}{\|W^T S_w W\|} \tag{5}$$

where N is the number of training samples and μ_1, μ_2 are the means of the training images into the eyes and non-eyes classes respectively. Once the projection matrix W is computed, all the training samples are projected into the discriminant space for scoring. For a given eye candidate region x, we first project the intensity values into the Fisher space resulting in x_p, its projection. We then compute the Mahalanobis distance between x_p and the means of the two classes in the Fisher space to compute scores that are used by the eye tracker, P_{eye} and P_{noteye} (in the notation of [9]).

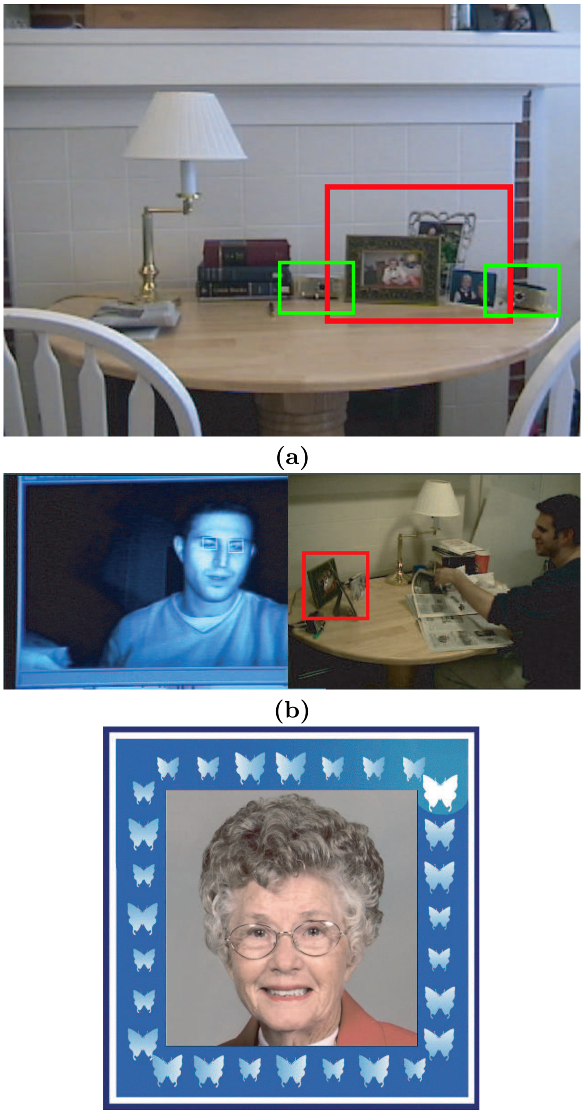
$$P_{eye}(x_p) = \|x_p - \mu_{F_1}\| \tag{6}$$

$$P_{noteye}(x_p) = \|x_p - \mu_{F_2}\| \tag{7}$$

Figure 3 shows the clear separation between the two classes in the discriminant space. We obtain this separation even in the presence of significantly varying infrared (sunlight) lighting conditions from multiple sources (windows).

Head pose and software application integration

We integrate the head pose and eye gaze estimates from the vision system as an noninvasive user interface for the



(a)

(b)

(c)

Figure 5: (a) Family Intercom setup. Red box: family pictures, green boxes: cameras, (b) subject using the Family Intercom, (c) A portrait of an elderly parent that can be used to monitor their eye gaze activity level.

HCI applications by treating the vision system as a server. The vision system can be queried by applications to find out how many eyes are in the scene, what the estimated head pose is and whether there is any overlap between application defined regions of interest and the user's head position. Treating the vision system as a service abstracts the technical details of the tracking and allows HCI researchers to focus on using the data provided by our system to support user interactions more effectively.

Figure 4 pictorially depicts our system and how it is used by software applications. Multiple eye trackers are connected to a server that merges the eye position estimates to calculate a head pose vector. The head pose vector is then made available for applications to access, as well as the eye positions that were found. We also allow applications to query the history of continued gaze towards a particular region as this may be of interest to the application as well.

The applications can also request that the data be filtered, in case some higher-level statistics are of interest. For example, the Family Intercom application seeks to understand trends in gaze duration of family pictures over several days. As a result, the server averages the gaze data to provide statistics that are of greater relevance to the application. On the other hand, the Cook's Collage is a real-time system, where the gaze direction is used to make inferences about the user's distractions from cooking, so raw data is made available instead.

4. EXPERIMENTS

We have experimented with two applications to support user interaction through estimates of eye gaze. Knowing the direction of a user's gaze allows us to construct attentive applications that are otherwise intrusive. Both of our applications consist of displays mounted in a residential setting, with cameras near the displays. Users need only be in front of the displays for user interaction to take place. The setups are shown in Figure 5 (Family Intercom) and Figure 7 (Cook's Collage).

Family Intercom

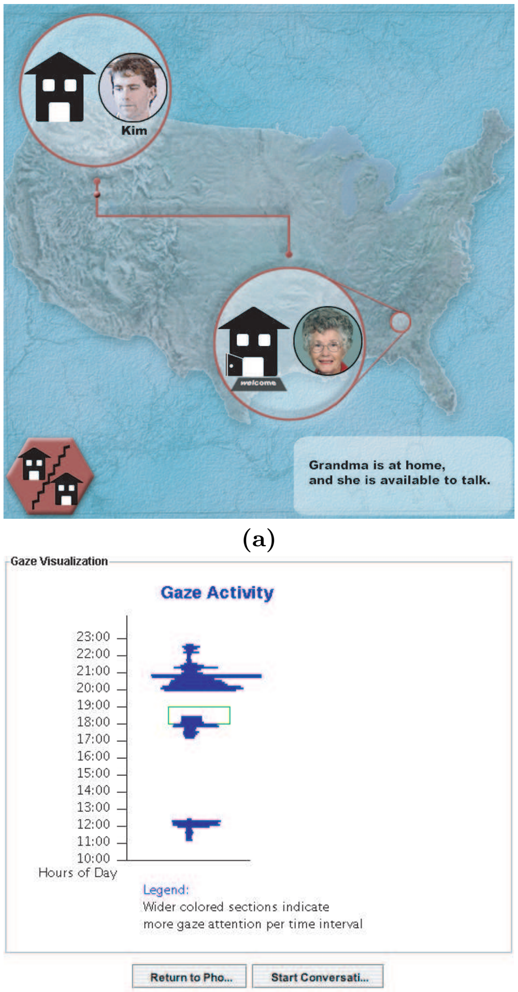
The Family Intercom project investigates context-aware family communication between homes [15]. The intent is to provide interfaces that facilitate a person's ability to decide whether a proposed conversation should be initiated or not. The user-gaze interface shown in Figure 6 provides feedback to the caller to help them determine whether it would be appropriate to initiate a voice conversation.

In one home, the vision-based eye tracking system tracks user gaze towards a collection of framed family photographs. Figure 5 shows the setup of the eye trackers on a common household table in an elderly parent's home. The red box shows the picture frame, and green boxes show the cameras used for eye tracking. In the second home, the remote panel is based on the Digital Family Portrait [14] and displays a portrait and a qualitative estimate of activity for the family member pictured from the first home. Figure 6b shows the interface at a remote family member's house. When a family member notices the digital portrait of their family, they simply touch the portrait to create a connection. The remotely collected eye-gaze data is displayed to provide context for the caller to gauge a time when the remote member desires family conversation.

The visual gaze tracker conveys statistics of the callee's eye gaze towards the family pictures to the caller and facilitates the appropriate social protocol for initiating conversations between the users. In previous prototypes [15] [14], only the room location of the callee was available via radio-frequency identification (RFID) tags worn by the family member. The RFID tracking provided coarse estimates of activity, but our gaze tracker may be used to infer finer time intervals when a conversation is more desirable.

Cook's Collage

The Cook's Collage application explores how to support retrospective memory, such as keeping a record of transpired



Legend:
Wider colored sections indicate
more gaze attention per time interval

(a)

(b)

Figure 6: Family Intercom: (a) A son checks on his elderly parent through the picture frame, (b) Gaze activity at remote family location; the box represents the current time.

events, using a household cooking scenario [19]. The Collage is composed of an attentive environment that captures the cook's actions and shows them on a display. Figures 7 and 8 show the Collage's overall setup and display. We use three eye trackers to provide additional information about user behavior while cooking. Our system estimates user attention in three regions: the Collage's display, recipe, and cooking area.

Figure 7 shows the setup for this experiment with our eye tracking system. We use three cameras (green boxes) to track the user's head pose within the three regions (red boxes). Currently, we use the head pose data to evaluate the usability of the Collage's display. The Cook's Collage expects the user to refer to its display every time the cook returns to cooking after an interruption for additional cooking prompts. Ignoring the display after an interruption could suggest the need for additional user interface improvements or a better placement of displays.

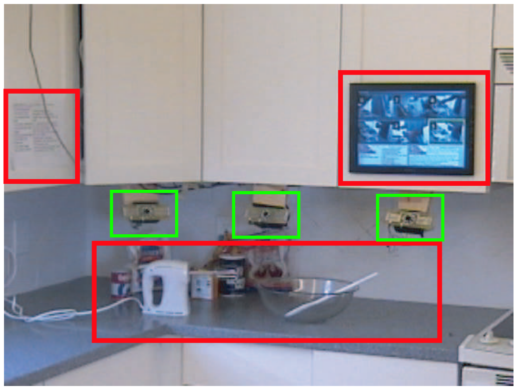


Figure 7: Cook's Collage setup. Red boxes from left to right: cooking recipe, cooking area, Collage display, green boxes: cameras.

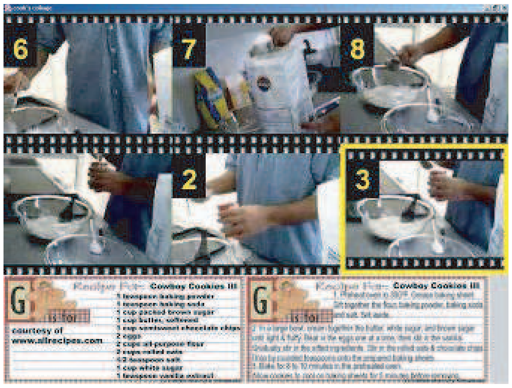


Figure 8: Cook's Collage display showing cooking steps.

5. RESULTS

Our quantitative results demonstrate the effectiveness of the head pose calculation subsystem. Since both applications use head pose data in the form of regions gazed by the user, we verify whether gaze estimates by a human are the same as those actually estimated by the system. The results are given in form of percentage of correct recognition by our tracking system.

Our subject pool consists of 4 subjects for the Family Intercom and 4 the for Cook's Collage. For the Family Intercom experiment, a subject performs a regular daily activity like reading or eating while sitting near a table containing the user's family pictures. A separate 15 minute sequence consisting of 225 frames is recorded throughout the experiment to capture ground truth for verification. The video is hand labeled to represent the ground truth of regions viewed by the user. These hand labeled frames are compared with the regions estimated by the head pose tracking system. The percentage of accuracy gives the fraction of frames estimated correctly by the system.

For the Cook's Collage experiment, subjects were asked to cook a recipe provided to them and a video was also recorded to capture ground truth. Since the length of this experiment is shorter than the Family Intercom experiment, a 10 minute sequence was used for comparison instead. This

Subject	**Correct estimate**
Subject 1	87%
Subject 2	88%
Subject 3	90%
Subject 4	84%
Average correct estimate	*87.25%*

Table 1: Statistics of estimated eye contact for the Family Intercom.

Subject	**Correct estimate**
Subject 1	81%
Subject 2	84%
Subject 3	78%
Subject 4	82%
Average correct estimate	*81.25%*

Table 2: Statistics of estimated eye contact for the Cook's Collage.

is also motivated by the fact that the user is more actively involved in the cooking experiment, which provides a larger amount of head poses for comparison purposes.

The Family Intercom experiment consists of one region where the user needs to look and in the Cook's Collage experiment there are three regions. The result set of the Cook's Collage experiment does not distinguish between the individual gazes at these three regions. Rather, the data shows the accuracy of head pose estimation for the entire experiment. Thus, one of the possible reasons for having a smaller accuracy for the Cook's Collage experiment might be due to having to track user head pose at more regions. The large accuracy of gaze recognition demonstrates the practicality of head pose estimation for attentive user interfaces.

6. SUMMARY & FUTURE WORK

In this paper, we demonstrate the ability of our system to track user head pose over multiple cameras in indoor settings. We are able to perform the tracking under varying lighting conditions for several users very robustly. In addition, we present a framework to seamlessly integrate our vision-based system with application prototypes to make higher-level inferences about user behavior. Our experiments with users in realistic applications shows the reliability of our system and its practicality in different kinds of application prototypes.

We plan on adding more training data to our recognition model to make it a true 'black box' to be used with different experimental applications in our laboratory by other researchers. We are also working with HCI and psychology researchers to design user studies to evaluate complete systems using the tracker. The feedback from these user studies could be used to modify the granularity of head pose data provided by the tracking system. We also plan to investigate how effective the gaze data has been in facilitating family communications and what new social implications arise from these kinds of perceptual systems. We would also like to conduct more experiments with several other application prototypes in our laboratory to explore new avenues for using perceptual interfaces based on vision-based eye tracking.

Acknowledgments

Several of our colleagues were extremely helpful with various aspects of this work especially David Minnen and Yifan Shi. We would like to thank Myron Flickner and David Koons from IBM Research Almaden for providing us with the Blue Eyes cameras and related basic software. Funding was provided by Aware Home Research Initiative, Georgia Tech Broadband Institute, Intel Undergraduate Research Awards, and Georgia Tech President's Undergraduate Research Awards.

7. REFERENCES

[1] A. Aaltonen, A. Hyrskykari, and K. Raiha. 101 spots on how do users read menus? In *Human Factors in Computing Systems: CHI 98*, pages 132–139, New York, 1998. ACM Press.

[2] Alan Allport. *Visual Attention.* MIT Press, 1993.

[3] M. Argyle. *Social Interaction.* Methuen & Co., London, England, 1969.

[4] M. Argyle and M. Cook. *Gaze and Mutual Gaze.* Cambridge University Press, Cambridge, UK, 1976.

[5] P. Barber and D. Legge. *Perception and Information.* Methuen & Co., London, England, 1976.

[6] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, July 1997.

[7] La Cascia, M. Sclaroff, and S. Athitso. Fast reliable head tracking under varying illumination: An approach based on robust registration of texture mapped 3-d models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[8] J.H. Goldberg and J.C. Schryver. *Eye-gaze determination of user intent at computer interface.* Elsevier Science Publishing, New York, New York, 1995.

[9] A. Haro, M. Flickner, and I. Essa. Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. In *IEEE Computer Vision and Pattern Recognition*, pages 163–168, 2000.

[10] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill. 3-d pose tracking with linear depth and brightness constraints. In *International Conference on Computer Vision*, 1999.

[11] B. Jabrain, J. Wu, R. Vertegaal, and L. Grigorov. Establishing remote conversations through eye contact with physical awareness proxies. In *Extended Abstracts of ACM CHI*, 2003.

[12] Y. Matsumoto, T. Ogasawara, and A. Zelinsky. Behavior recognition based on head pose and gaze direction measurement. In *IEEE International Conference on Intelligent Robots and Systems*, 2000.

[13] C.H. Morimoto, D. Koons, A. Amir, and M. Flickner. Pupil detection and tracking using multiple light sources. Technical report RJ-10117, IBM Almaden Research Center, 1998.

[14] E. Mynatt, J. Rowan, and A Jacobs. Digital family portraits: Providing peace of mind for extended family members. In *ACM CHI*, 2001.

[15] K. Nagel, C. Kidd, T. O'Connell, S. Patil, and

G. Abowd. The family intercom: Developing a context-aware audio communication system. In *Ubicomp*, 2001.

[16] R. Ruddarraju, A. Haro, and I. Essa. Fast multiple camera head pose tracking. In *International Conference on Vision Interfaces*, Halifax, Canada, 2003.

[17] A. Schoedl, A. Haro, and I. Essa. Head tracking using a textured polygonal model. In *Proceedings Workshop on Perceptual User Interfaces*, 1998.

[18] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *International Conference on Multi-Modal Interfaces*, 2002.

[19] Q. Tran and E. Mynatt. Cook's collage: Two exploratory designs. In *CHI 2002, Conference Proceedings*, 2002.

[20] A. L. Yarbus. *Eye Movements during Perception of Complex Objects*. Plenum Press, New York, New York, 1967.