Face-responsive interfaces: from direct manipulation to perceptive presence

Trevor Darrell, Konrad Tollmar, Frank Bentley, Neal Checka,
Loius-Phillipe Morency, Ali Rahimi and Alice Oh

MIT AI Lab
Cambridge MA 02139

Abstract. Systems for tracking faces using computer vision have recently become practical for human-computer interface applications. We are developing prototype systems for face-responsive interaction, exploring three different interface paradigms: direct manipulation, gaze-mediated agent dialog, and perceptually-driven remote presence. We consider the characteristics of these types of interactions, and assess the performance of our system on each application. We have found that face pose tracking is a potentially accurate means of cursor control and selection, is seen by users as a natural way to guide agent dialog interaction, and can be used to create perceptually-driven presence artefacts which convey real-time awareness of a remote space.

1 Introduction

A key component of proposed pervasive computing environments is the ability to use natural and intuitive actions to interact with computer systems. Faces are used continuously in interaction between people, and thus may be important channels of communication for future devices. People signal intent, interest, and direction with their faces; new, perceptually enabled interfaces can allow them to do so with computer systems as well.

Recent progress in computer vision for face processing has made it possible to detect, track, and recognize faces robustly and in real-time. To date, however, applications of this technology have largely been in the areas of surveillance and security (scene monitoring, access control, counterterrorism). In contrast, we are interested in the use of this "perceptive interface" technology for human computer interaction and computer mediated communication.

While computer vision systems for tracking faces typically have well defined outputs in terms of 3D position, orientation, and identity probability, how those signals should be used in interface tasks remains less understood. Even if we restrict our attention to a single aspect of face processing, e.g., face pose, it is apparent that there are a variety of interface paradigms that can be supported. There are many variables that distinguish possible paradigms, e.g., interaction can be direct or indirect, can mediate human communication or control an automated system, and can be part of an existing GUI paradigm or be placed within a physical media context.

In this paper we explore the use of face pose tracking technology for interface tasks. We consider the space of face interface paradigms, describe three specific instances in that space, and develop a face-responsive interface interface for each paradigm.

In the following section we analyze the characteristics of interaction paradigms for face pose interaction. We then review related work and our technology for robust, real-time face pose tracking. Following that we describe three prototype applications which adopt direct manipulation, gaze-mediated agent dialog, and perceptually mediated remote presence paradigms, respectively. We conclude with an assessment of the results so far, what improvements are needed, and future steps to make face pose interfaces usable by everyday users.

2 Face pose interaction paradigms

In contrast to traditional WIMP, command line, and push-button interfaces, perceptive interfaces offer the promise of non-invasive, untethered, natural interaction. However, they can also invade people's privacy, confuse unintentional acts with communicative acts, and may be more ambiguous and error-prone than conventional interfaces. Therefore, the particular design of a perceptive interface is very important to the overall success of the system.

Because the technology for perceptive interfaces is evolving rapidly, it is premature to propose a comprehensive design model at this stage. However, we believe there are some general principles which can expose the space of possible interface designs. We also believe that it is possible to build simple prototypes using current technology and evaluate whether they are effective interfaces.

The space of possible perceptive interfaces is quite broad. To analyze the range of designs, we have considered a taxonomy based on the following attributes that characterize a perceptive interface:

- Nature of the control signal. Is direct interaction or an abstract control supported?
- Object of communication. Does interaction take place with a device or with another human over a computer mediated communication channel?
- Time scale. Is the interaction instantaneous, or time-aggregated; is it real-time or time-shifted communication?

This is a non-exclusive list, but it captures the most important characteristics.

The perception of faces plays a key role in perceptual interfaces. Detection, identification, expression analysis, and motion tracking of faces are all important perceptual cues for active control or passive context for applications. In this paper we restrict our attention to the latter cue, face pose, and explore it's use in a variety of application contexts and interaction styles. We use a real-time face pose tracking method based on stereo motion techniques, described in more detail in the following section. We are constructing a series of simple, real-time prototypes which use this tracking system and explore different aspects of the characteristics listed above.

Our first prototype explores the use of head pose for direct manipulation of a cursor or pointer. With this prototype, a user could control the location of a cursor or select objects directly using the motion of his or her head as a control signal. Using the taxonomy implied by the above characteristics, it uses direct interface, device interaction, and real-time interaction. Our second prototype focuses on pose-mediated agent dialog interface: it also uses direct interface and is real-time, but interaction is with an agent character. The agent listens to users only when the user's face pose indicates he or she is attending to a graphical representation of the agent. A third prototype uses motion and pose detection for perceptive presence. It conveys whether activity is present in a remote space, and whether one user is gazing into a communication artifact. This prototype uses abstract control, human interaction, and is instantaneous.

We next review related work and describe our tracking technology, and then present the cursor control, agent dialog and perceptive presence prototypes. We conclude with a discussion and evaluation of these prototypes, and comments on future directions.

3 Previous work on face pose tracking

Several authors have recently proposed face tracking for pointer or scrolling control and have reported successful user studies [31, 19]. In contrast to eye gaze [37], users seem to be able to maintain fine motor control of head gaze at or below the level needed to make fine pointing gestures[1]. However, performance of the systems reported to date has been relatively coarse and many systems required users to manually initialize or reset tracking. They are generally unable to accurately track large rotations under rapid illumination variation (but see [20]), which are common in interactive environments (and airplane/automotive cockpits).

Many techniques have been proposed for tracking a user's head based on passive visual observation. To be useful for perceptive interfaces, tracking performance must be accurate enough to localize a desired region, robust enough to ignore illumination and scene variation, and fast enough to serve as an interactive controller. Examples of 2-D approaches to face tracking include color-based [36], template-based [19, 24], neural net [29] and eigenface-based [11] techniques. Integration of multiple strategies is advantageous in dynamic conditions; Crowley and Berard [9] demonstrated a real time tracker which could switch between detection and tracking as a function of tracking quality.

Techniques using 3-D models have greater potential for accurate tracking but require knowledge of the shape of the face. Early work presumed simple shape models (e.g., planar[3], cylindrical[20], or ellipsoidal[2]). Tracking can also be performed with a 3-D face texture mesh [28] or 3-D face feature mesh [35].

Very accurate shape models are possible using the active appearance model methodology [8], such as was applied to 3-D head data in [4]. However, tracking

[1] Involuntary microsaccades are known to limit the accuracy of eye-gaze based tracking[18].

3-D active appearance models with monocular intensity images is currently a time-consuming process, and requires that the trained model be general enough to include the class of tracked users.

We have recently developed a system for head pose tracking, described below, based on drift-reduced motion stereo techniques which are robust to strong illumination changes, automatically initialize without user intervention, and can re-initialize automatically if tracking is lost (which is rare). Our system does not suffer from significant drift as pose varies within a closed set since tracking is performed relative to multiple base frames and global consistency is maintained.

4 A motion stereo-based pose tracking system

Our system has four main components. Real-time stereo cameras (e.g., [10, 16]) are used to obtain real-time registered intensity and depth images of the user. A module for instantaneous depth and brightness gradient tracking [12] is combined with modules for initialization, and stabilization/error-correction. For initialization we use a fast face detection scheme to detect when a user is in a frontal pose, using the system reported in [33]. To minimize the accumulation of error when tracking in a closed environment, we rely on a scheme which can perform tracking relative to multiple base frames [26].

When it first comes online, the tracker scans the image for regions which it identifies as a face using the face detector of [33]. As soon a face has been consistently located near the same area for several frames, the tracker switches to tracking mode. The face detector is sensitive only to completely frontal heads, making it possible for the tracker to assume that the initial rotation of the head is aligned with the coordinate system. The face detector provides the tracker an initial region of interest, which is updated by the tracker as the subject moves around. Since depth information is readily available from the stereo camera, the initial pose parameters of the head can be fully determined by 2D region of the face with the depth from stereo processing.

When we observe erratic translations or rotations from the tracker, the tracker automatically reinitializes by reverting to face detection mode until a new target is found. This occurs when there is occlusion or rapid appearance changes.

4.1 Finding Pose Change Between Two Frames

Because synchronized range and intensity imagery is available from stereo cameras, our system can apply the traditional Brightness Change Constraint Equation (BCCE) [13] jointly with the Depth Change Constraint Equation (DCCE) of [12] to obtain more robust pose change estimates.

To recover the motion between two frames, the BCCE finds motion parameters which minimize the appearance difference between the two frames in a least-squares sense:

$$\delta^* = \arg\min_{\delta} \epsilon_{BCCE}(\delta)$$

$$\epsilon_{BCCE} = \sum_x \|I_t(x) - I_{t+1}(x + u(x;\delta)\|^2 \quad (1)$$

where $u(x;\delta)$ is the image flow at pixel x, parameterized by the details of a particular motion model. In the case of 3D rigid motion under a perspective camera, the image flow becomes:

$$\begin{bmatrix} u_x \\ u_y \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix} (\delta_\omega \times X + \delta_\Delta), \quad (2)$$

where X is the world coordinate of the image point x, δ_ω is the infinitesimal rotation of the object, δ_Δ is its infinitesimal translation, and f is the focal length of the camera[5].

The DCCE of [12] uses the same functional form as equation (1) to constrain changes in depth. But since under rotation, depth is not preserved, the DCCE includes an adjustment term:

$$\epsilon_{DCCE} = \sum_x \|Z_t(x) - Z_{t+1}(x + u(x;\delta)) + V_z(x;\delta)\|^2,$$

where V_z is the flow towards the Z direction induced by δ. Note that the DCCE is robust to lighting changes since lighting does not affect the depth map. We combine the BCCE and DCCE into one function optimization function with a weighted sum:

$$\delta^* = \arg\min_\delta \epsilon_{BCCE}(\delta) + \lambda\epsilon_{DCCE}(\delta),$$

See [12] for a method for solving this system. In practice the depth gradient approach worked poorly for abrupt motion; see [22] for a forumlation stable to large translations which incorporates improved optimization criteria based on an range registration algorithm.

4.2 Reducing Drift

Given a routine for computing the pose difference δ_s^t between frames I_s and I_t, there are two common strategies for estimating the pose ξ_t of frame I_t relative to the pose of frame I_0. One approach is to maintain the pose difference between adjacent frames I_s and I_{s+1}, for $s = 0..t-1$, and to accumulate these measurements to obtain the pose difference between frames I_t and I_0. But since each pose change measurement is noisy, the accumulation of these measurements becomes noisier with time, resulting in unbounded drift. A common alternative is to compute the pose difference between I_t and I_0 directly. But this limits the allowable range of motion between two frames, since most tracking algorithms (including the one described in the previous section) assume that the motion between the two frames is very small.

To address the issue of drift in parametric tracking, we compute the pose change between I_t and several base frames. These measurements can then be combined to yield a more robust and drift-reduced pose measurement. When

the trajectory of the target crosses itself, pose differences can be computed with respect to early frames which have not been corrupted by drift. Trackers employing this technique do not suffer from the unbounded drift observed in other differential trackers.

In [26], a graphical model is used to represent the true poses ξ_t as hidden variables and the measured pose changes δ_s^t between frames I_s and I_t as observations. Unfortunately, the inference algorithm proposed is batch, requiring that pairwise pose changes be computed for the entire sequence before drift reduction can be applied.

We use a simple online algorithm to determine the pose of a frame I_t. Our algorithm first identifies the k frames from the past which most resemble I_t in appearance. The similarity measure we use is the sum of squared differences:

$$d_s^t = \sum_x \sum_y \|I_s(x,y) - I_t(x,y)\|^2. \tag{3}$$

Since the frames from the past have suffered less drift, the algorithm discounts the similarity measure of newer frames, biasing the choice of base frame toward the past.

Once the candidate base frames have been identified, the pose change between each base frame I_s to I_t is computed using the algorithm described in the previous section. The final pose assigned to frame I_t is the average pose of the two base frames, weighted by the similarity measure of equation (3):

$$\xi_t = \frac{\sum_i (\xi_{si} + \delta_{si}^t)/d_{si}^t}{\sum_i 1/d_{si}^t}.$$

As an alternative, see [25] for a related formulation using an explicit graphical model.

5 Cursor control prototype

Head pose or gaze is a potentially powerful and intuitive pointing cue if it can be obtained accurately and non-invasively. In interactive environments, like public kiosks or airplane cockpits, head pose estimation can be used for direct pointing when hands and/or feet are otherwise engaged or as complementary information when desired action has many input parameters. In addition, this technology can be important as a hands-free mouse substitute for users with disabilities or for control of gaming environments.

We implemented a prototype for head-pose driven cursor control using the tracking technology described above, and tested it in medium (screen/cockpit) and large (room) scale environments. The performance of our system was evaluated on direct manipulation tasks involving shape tracing and selection. We compared our tracker performance with published reports and side-by-side implementations of two other systems. We experimented with small and large head rotations, different levels of lighting variation, and also compared the performance of our tracker with that of a head-mounted inertial sensor.

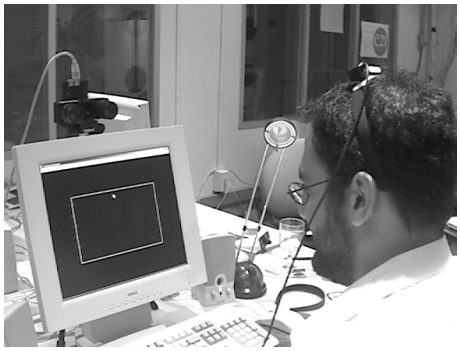


Fig. 1. A user during the desktop experiment. The SRI stereo camera is placed just over the screen and the user is wearing the Intertrax^2 device on his head.

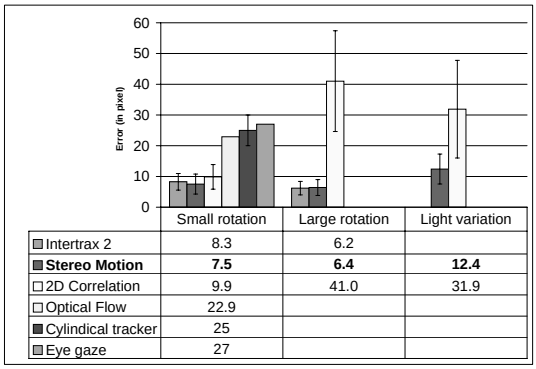
5.1 Desktop Experiment

As shown in figure 1, in the desktop experiment users sat about 50 cm away from a typical 17" screen, subtended a horizontal angle of about 30 degrees and a vertical angle of about 20 degrees. The screen displayed a black background and a white rectangular path drawn in the middle. The task was to use head pose to move a 2D pointer around the screen to trace the rectangular path as accurately as possible. Users were allowed to take as much time as they liked, as long as they were able to complete the path.

The desktop experiment involved eight experiments per subject. Each subject used the tracking system described above, as well as a 2-D normalized correlation tracker similar to that proposed in [19] and a wired inertial rotation sensor (InterSense's Intertrax^2 [14]). Each of the trackers was tested in small-screen and wide-screen mode. The former allows the user to trace the rectangle using small head motions. The latter simulates a larger screen which requires larger head rotations to navigate. In addition, the correlation tracker and the stereo motion tracker were tested in the small-screen mode under abruptly varying lighting conditions (see [23] for full details.)

The first three rows of figure 2 compares the accuracy of the stereo motion tracker with the 2D normalized cross-correlation tracker and the Intertrax^2 tracker. The histogram shows the average error and standard deviation of 4 subjects. The average error is computed as the average distance in pixels between every point on the cursor trajectory and the closest point on the given rectangular path. The three last rows of the same figure compares our results with some published systems: an optical flow tracker[15], cylindrical tracker[20], and an eye gaze tracker[37].

In a desktop environment, small rotations are sufficient to drive a cursor, since the angle subtended by the screen tends to be small. This situation serves as a baseline where all three trackers can be compared under moderate conditions. Under the small rotation scenario, all trackers showed similar deviation from the



	Small rotation	Large rotation	Light variation
Intertrax 2	8.3	6.2	
Stereo Motion	**7.5**	**6.4**	**12.4**
2D Correlation	9.9	41.0	31.9
Optical Flow	22.9		
Cylindical tracker	25		
Eye gaze	27		

Fig. 2. Comparison of average error on tracing task of the desktop experiment. The error bars in the histogram represent the standard deviation between user results.

given trajectory, with an average deviation of 7.5 pixels for the stereo motion tracker, 9.8 pixels for the normalized cross-correlation tracker, and 8.3 pixels for the inertial tracker.

Navigating a pointer on a wide screen (multiple monitors, projection screens, cockpits) requires larger head rotations. As expected, the correlation tracker loses track of the subject during rotations beyond 20 degrees, because the tracker is initialized on the appearance of the frontal face only. It incurred an average error of 41.0 pixels. The stereo motion tracker, however, successfully tracks the head as it undergoes large rotations, with an average error of 6.4 pixels. The Intertrax^2 tracker shows an average error of 6.2 pixels. Note that due to the accumulated drift of the inertial sensor, typical users had difficulty controlling the cursor in the last portion of the trajectory.

We observe that the inertial rotation sensor Intertrax^2 is accurate for a short period of time, but it accumulates noticeable drift. Approximately after 1 minute of use of the tracker, subjects were often forced to contort their bodies significantly in order to compensate for the drift. The normalized cross-correlation tracker appears to be suitable for situations involving small head rotations and minimal illumination changes. The stereo motion tracker is robust to lighting variations because it largely relies on depth information, which is unaffected by the illumination changes. In addition, it can track arbitrarily large transformations without suffering from drift due to the drift reduction algorithm described in section 4.2.

5.2 Interactive Room Experiment

As shown in figure 3, the second experiment was run in an interactive room with large projection screens. Users were sitting about 1.8 meters away from a 2.1m x 1.5m projection screen, subtended a horizontal angle of about 100 degrees and a vertical angle of about 80 degrees. Subjects were asked to perform two tasks:

Fig. 3. Setup for the room experiment. The SRI stereo camera is placed on the table.

	Average error (in pixel)	Standard deviation (in pixel)
Small rotation	6.3	0.4
Large rotation	6.1	0.6
Light variation	11.5	3.1

Table 1. Experimental results of the stereo-based tracker inside the interactive room.

the tracing task described above, and a selection task where the user must reach different colored squares without touching the red squares. A short interview was performed following the experiment to obtain feedback from the subject about the usability of these head trackers.

With more then 90 degrees of rotation to reach both sides of the screens, the limitations of the normalized cross-correlation tracker appeared clearly. Subjects could not use the tracker without unnaturally translating their heads over long distances to move the cursor correctly.

The stereo-based tracker was successful on both the tracing task and the selection task. Table 1 presents the average errors and standard deviation for the tracing task of 3 subjects.

The interviews after the second experiment showed that users doesn't like a linear mapping between the head pose and the cursor position. For slow movement of the head, the ratio cursor distance by head movement should be smaller to give more precision on small selections. For fast movement of the head, the ratio should larger to give more speed on large displacement. These observations corroborate Kjeldson results[19].

5.3 Discussion

For direct manipulation tasks such as driving cursors and selecting objects, the stereo head tracking system presented above is accurate to within a half degree of accuracy. Informally, we observed that this was approximately equal to the accuracy of some conventional input devices, for example novice (or non-dominant hand) trackball use. We believe this type of system will be an important module in designing perceptual interfaces for screen interaction and cockpit applications, and for disabled users who are not able to use traditional interfaces but need direct manipulation control. We next turn our attention to a more abstract use of pose, that of signaling intent to communicate.

6 Agent dialog prototype

As we move beyond traditional desktop computing and explore pervasive computing environments, we naturally come across settings where multiple users interact with one another and with a multitude of devices and/or software agents. In such a collaborative setting, interaction using conversational dialog is an appealing paradigm. Automatic speech recognition systems are becoming robust enough for use in these environments, at least with a single speaker and a close microphone. However, when there are multiple speakers and potential listeners knowing who is speaking to whom is an important and difficult question that cannot always be answered with speech alone.

Pose or gaze tracking has been identified as an effective cue to help disambiguate the addressee of a spoken utterance. In a study of eye gaze patterns in multi-party (more than two people) conversations, Vertegaal, et al. [32] showed that people are much more likely to look at the people they are talking to, than any other people in the room. Also, in another study, Maglio, et al. [21] found that users in a room with multiple devices almost always look at the devices before talking to them. Stiefelhagen et al. [30] showed that the focus of attention can be predicted from the head position 74during a meeting scenario.

Hence, it is natural to believe that using pose as an interface to activate automatic speech recognition (ASR) will enable natural human-computer interaction (HCI) in a collaborative environment. In conversational agents, the importance of nonverbal gestures has already been recognized [6]. We evaluated whether face pose could replace conventional means of signaling communication with an interactive agent. We implemented three paradigms for speaking with an agent: "look-to-talk" (LTT), a gaze-driven paradigm, "talk-to-talk" (TTT), a spoken keyword-driven paradigm, and "push-to-talk" (PTT), where the user pushes a button to activate ASR. We present and discuss a user evaluation of our prototype system as well as a Wizard of Oz (WOz) setup.

To compare the usability of LTT with the other modes, we ran two experiments in the MIT AI Lab's Intelligent Room [7](from here on "the I-Room"). We ran the first experiment with a real vision- and speech-based system, and the second experiment with a WOz setup where gaze tracking and ASR were

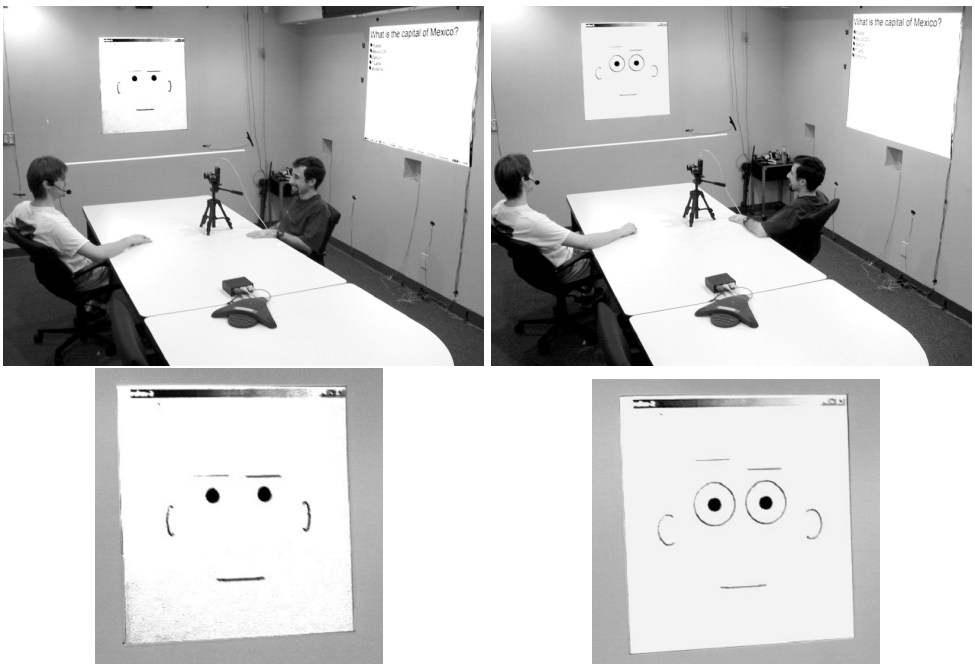


Fig. 4. Interaction with a conversational agent character using face pose. On the left the user is interacting with a colleague, and the agent is not listening to the user's speech commands. On the right the user is facing the agent, and the agent is listening to the user. The bottom row shows close up of the agent expression icons used to indicate not-listening and listening status.

simulated by an experimenter behind the scenes. Each subject was asked to use all three modes to activate ASR and then to evaluate each mode.

6.1 "Look-to-talk" experiment

We set up the experiment to simulate a collaboration activity among two subjects and a software agent. The first subject (subject A) sits facing the front wall displays, and a second "helper" subject (subject B) sits across from subject A. The task is displayed on the wall facing subject A. The camera is on the table in front of subject A, and Sam, an animated character, is displayed on the side wall (Figure 4). Subject A wears a wireless microphone and communicates with Sam via IBM ViaVoice. Subject B discusses the task with subject A and acts as a collaborator. Subject B's words and pose are not detected by the environment.

Sam represents the software agent with which Subject A communicates. Sam is built from simple shapes forming a face, which animate to continually reflect the state of the software agent that it represents. During this experiment, Sam read quiz questions through a text-to-speech synthesizer, and was constrained to two facial expressions: non-listening and listening.

There were 13 subjects, 6 for the first experiment and 7 for the WOz setup. They were students in computer science, some of whom had prior experience with TTT in an intelligent environment.

Each pair of subjects was posed three sets of six trivia questions, each set using a different mode of interaction in counterbalanced order. In the WOz setup, we ran a fourth set in which all three modes were available, and the subjects were told to use any one of them for each question. Table 2 illustrates how users activate and deactivate ASR using the three modes, and what feedback the system provides for each mode.

After the experiment, the subjects rated each of the three modes on a scale of one to five on three dimensions: ease-of-use, naturalness, and future use. We also asked the subjects to tell us which mode they liked best and why

6.2 Discussions

For the first experiment, there was no significant difference (using anova at $a = 0.05$) between the three modes for any of the surveyed dimensions. However, most users preferred TTT to the other two. They reported that TTT seemed more accurate than LTT and more convenient than PTT.

For the WOz experiment, there was a significant difference in the naturalness rating between PTT and the other two ($p = 0.01$). This shows that, with better perception technologies, both LTT and TTT will be better choices for natural HCI. Between LTT and TTT, there was no significant difference on any of the dimensions. Five out of the seven subjects reported, however, that they liked TTT best, compared to two subjects who preferred LTT. One reason for preferring TTT to LTT was that there seemed to be a shorter latency in TTT than LTT. Also, a few subjects remarked that Sam seemed disconnected from the task, and thus it felt awkward to look at Sam.

Despite the subjects' survey answers, for the fourth set, 19 out of 30 questions were answered using LTT, compared with 9 using TTT (we have this data for five out of the seven subjects; the other two chose a mode before beginning fourth set to use for the entire set, and they each picked LTT and TTT). When asked why he chose to use LTT even though he liked TTT better, one subject answered "I just turned my head to answer and noticed that the Room was already in listening mode." This confirms the findings in [21] that users naturally look at agents before talking to them.

Under ideal conditions (i.e., WOz), users preferred perceptual interfaces to push-to-talk. In addition, they used look-to-talk more often for interacting with agents in the environment. This has led us to believe that look-to-talk is a promising interface. However, it is clear that having all three modalities available for users provides convenience and efficiency for different contexts and user preferences. We are currently working to incorporate look-to-talk with the other modalities. We are also investigating ways to improve gaze tracking accuracy and speed. As the prototype tracker performance approaches that of the WOz system, we expect the look-to-talk user experience to improve significantly.

Mode	Activate command	Active feedback	Deactivate command	Deactivate feedback
PTT	Switch the microphone to "on"	Physical status of the switch	Switch the microphone to "mute"	Physical status of the switch
LTT	Turn head toward Sam	Sam shows listening expression	Turn head away from Sam	Sam shows normal expression
TTT	Say "computer"	Special beep	Automatic (after 5 sec)	None

Table 2. How to activate and deactivate the speech interface for each of the three modes: Push-to-talk (PTT), Look-to-talk (LTT), and Talk-to-talk (TTT).

7 Perceptive Presence Prototype

In our "Perceptive Presence" project we are investigating the use of ambient media cues to indicate presence. In particular we are investigating perceptually grounded information for conveying the presence and activity of an individual between remote places. Our approach is to use motion detection and face-based tracking techniques presented above to become aware of a user's presence and focus of attention.

Face-based sensing has the advantage that the explicit signaling performed by the user that is similar to real-life communication. For example, a user can signal his presence by simply directing his gaze towards a specific picture or device, much as he would turn to a person in order to speak to them.

New computing technologies offer greater bandwidth and the potential for persistent, always-on connections, such as instant messages and point-to-point video links (e.g., video-mediated communications). These technologies require most often that a user explicitly respond to each interaction through the normal devices, e.g. a mouse click in a dialog. However, increasing the volume and frequency of message traffic may not lead to greater connectedness, and may be a burden if users have to explicitly compose each message [34].

Work similar to ours has inspired and invoked a general interest of research in HCI in searching for communication with the purpose of expressing intention and awareness without having to interact with a keyboard and mouse. Brave and Dahley have, for instance, proposed to examine the potential of touch for use as a mood-induction technique for emotional communication [27]. Other visionary examples that stem from a mixture of art and human-computer interaction are proposed by Gaver & Martin [1] and Ishii & Ullmer [17].

Yet most of these projects have used technology which requires physical interaction. We are interested in passive, untethered interaction using a face-responsive interface, and have experimented with a pair of simple physical artifact that convey a user's presence and attention state in a remote location.

7.1 Presence lamp experiment

Our first experiment has been with "Perceptive Presence Lamps". These are a pair of lamps that convey remote presence through their illumination level. The light varies in intensity depending on the remote presence information received

from motion and face-based trackers, creating a living presence artifact. One lamp is placed in an office belonging to someone that a user would like to share their presence with, and the other lamp is placed in their office. In the current version we limit the concept to a pair of two lamps that are connected through the Internet. The lamp serves to answer questions such as "Is John present in the office?" or "Is John trying to get my attention?" The current lamp measures two level of presence. The first can be described as "physical presence." The lamp measures the amount of body movement close to the lamp in order to determine if a person is at their desk. If a person is present, the system signals a glowing light in the peer lamp.

The second level of presence information is "attention getting." If a user directs his focus on the lamp for a specific time period (a few seconds), the lamp interprets this as an attention-getting presence gesture and lights up the peer lamp to its brightest setting. When the face moves away or movement is no longer detected a message is passed to the peer lamp, which then dims appropriately.

The functional prototype that we created for this project integrates key components of vision-based face tracking, motion sensing, and conveys multiple levels of presence into a simple lamp design that easily fits on a desk. The lamp is small and relatively unobtrusive in an office setting. The dimming of the lamp is currently controlled with X10 commands sent over a powerline.

The prototype system (see Figure 5) was developed and initially tested over several weeks. Two peer colleagues whose offices were located in opposite sides of an office building used the lamps. Our preliminary results point to several findings. The users felt that action of looking at the lamp was a natural way of interacting. Despite the relatively crude resolution of the presence representation, it was perceived as supporting a connection to the remote space. However the context of the attention signal was often not clear to the particpants. We concluded that face-based tracking should be augmented with other clues that make is possible to extract other types of vision data that could support the interpretation of the interaction.

Additionally, the placement of the lamp (and hence the camera) seems to be crucial to correctly interpreting users intentions. Since the lamp also provide information about the other person a users must be able to look at lamp without that "gaze" become recognized as a "gaze" signal to send to the other lamp. Presently we use a audio cue and time-delay to resolve this issue, but we are experimenting with other approaches.

7.2 Discussion

We explored an untethered way to convey presence information in a given environment with a physical device. Our prototypes should be seen as experiments on how we can interact and communicate with our bodies, specifically faces, in order to express our attention and presence. Throughout the process care must be taken that face-tracking data is used in a sensible way, since the nature of human face and eye-movements is a combination of several voluntary and involuntary cognitive processes. Many issues remain to be investigated, e.g., to what

Fig. 5. The two upper images show the level of physical presence when both users are in their office but not explicitly looking at the lamp (which is dim). In the lower images, one of the users has noticed his lamp getting brighter and has returned that gaze.

detail we need to (and should) record and transmit perceptive information. The long-term idea is to provide a language of expressive activity and gesture that achieves intimate expression and yet is accessible by novice users. Many more studies will be needed with users in a variety of environments to fully characterize the types of expressive activity information that should be used.

8 Conclusion and Future work

We have explored the use of face pose tracking in three different human-computer interface paradigms: direct manipulation, conversational dialog, and remote presence. The stereo head tracking system we used requires no manual initialization, does not drift, and works for both screen and wall-scale interactions.

In experiments with direct manipulation cursor control tasks, we demonstrated the ability of users to trace outlines and select objects. Performance of this tracker was compared against that of a head-mounted inertial sensor and monocular vision techniques. Direct manipulation may be an important module in designing perceptual interfaces for intelligent environments, cockpit applica-

tions, and for disabled users who are not able to use traditional interfaces. We also constructed a prototype system for controlling conversational dialog interaction with an animated agent character. Users preferred perceptual modes of selection and felt "look-to-talk" was a natural paradigm. Finally, we explored perceptually driven remote presence through the use of lamps that conveyed the motion and face pose state from one room to another. Our results are very preliminary for this system, but our initial observations are that it is an interesting new mode of interaction and can create a sense of connectedness between remote collaborators or colleagues that is not possible through conventional communication channels. We plan to conduct more user studies with this prototype in the near future, and iterate our system design based on user feedback.

We have argued that face tracking, and specifically information about face pose, allows a range of interesting new human computer interface methods. It will be most powerful in conjunction with other perceptual cues, including identity, spoken utterance, and articulated body tracking. Our group is working on these cues as well, and hopes to integrate them as part of future research.

References

1. Gaver B. and Martin H. Alternatives: Exploring information appliances through conceptual design proposals. In *Proc. of CHI'2000, Den Haag,*, 2000.
2. S. Basu, I.A. Essa, and A.P. Pentland. Motion regularization for model-based head tracking. In *ICPR96*, page C8A.3, 1996.
3. M.J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV95*, pages 374–381, 1995.
4. V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH99*, pages 187–194, 1999.
5. A.R. Bruss and B.K.P Horn. Passive navigation. In *Computer Graphics and Image Processing*, volume 21, pages 3–20, 1983.
6. J. Cassell. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In *Embodied Conversational Agents*, 2000.
7. M. Coen. Design principles for intelligent environments. In *Fifteenth National Conference on Artificial Intelligence.*, 1998.
8. T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *PAMI*, 23(6):681–684, June 2001.
9. J. L. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '97,San Juan, Puerto Rico*, 1997.
10. Videre Design. *MEGA-D stereo camera.* http://www.videredesign.com.
11. G.D. Hager and P.N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *PAMI*, 20(10):1025–1039, October 1998.
12. M. Harville, A. Rahimi, T. Darrell, G.G. Gordon, and J. Woodfill. 3d pose tracking with linear depth and brightness constraints. In *ICCV99*, pages 206–213, 1999.
13. B.K.P. Horn and B.G. Schunck. Determining optical flow. *AI*, 17:185–203, 1981.
14. InterSense Inc. *Intertrax 2.* http://www.intersense.com.
15. Mouse Vision Inc. *Visual Mouse.* http://www.mousevision.com.
16. Tyzx Inc. *Deepsea stereo system.* http://www.tyzx.com.

17. H. Ishii and B. Ullmer. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Proc. of CHI '97*, 1997.
18. R.J.K Jacob. *Eye tracking in advanced interface design*, pages 258–288. Oxford University Press, 1995.
19. R. Kjeldsen. Head gestures for computer control. In *Proc. Second International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pages 62–67, 2001.
20. M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of textured-mapped 3d models. *PAMI*, 22(4):322–336, April 2000.
21. Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, and Barton A. Smith. Gaze and speech in attentive user interfaces. In *ICMI*, pages 1–7, 2000.
22. Louis-Philippe Morency and Trevor Darrell. Stereo tracking using icp and normal flow. In *Proceedings Int. Conf. on Pattern Recognition*, 2002.
23. Louis-Philippe Morency, Ali Rahimi, Neal Checka, and Trevor Darrell. Fast stereo-based head tracking for interactive environment. In *Proceedings of the Int. Conference on Automatic Face and Gesture Recognition*, 2002.
24. Ravikanth Pappu and Paul Beardsley. A qualitative approach to classifying gaze direction. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan*, 1998.
25. A. Rahimi, L. Morency, and T. Darrell. Bayesian network for online global pose estimation. In *International Conference on Intelligent Robots and Systems (IROS)*, to appear (September 2002).
26. A. Rahimi, L.P. Morency, and T. Darrell. Reducing drift in parametric motion tracking. In *ICCV01*, volume 1, pages 315–322, 2001.
27. Brave S. and Dahley A. intouch: A medium for haptic interpersonal communication. In *Proceedings of CHI '97,*, 1997.
28. A. Schodl, A. Haro, and I. Essa. Head tracking using a textured polygonal model. In *PUI98*, 1998.
29. R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *Proceedings of Workshop on Perceptual User Interfaces: PUI 98, San Francisco, CA*, pages 25–30, 1998.
30. R. Stiefelhagen, J. Yang, and A. Waibel. Estimating focus of attention based on gaze and sound. In *Workshop on Perceptive User Interfaces (PUI 01).*, 2001.
31. K. Toyama. Look,ma - no hands!hands-free cursor control with real-time 3d face tracking. In *PUI98*, 1998.
32. R. Vertegaal, R. Slagter, G.C. Van der Veer, and A. Nijholtxs. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proc of ACM Conf. on Human Factors in Computing Systems*, 2001.
33. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
34. S. Whittaker, L. Terveen, and et al. The dynamics of mass interaction. In *Proceedings of CSCW 98, Seattle, ACM Press*, 1998.
35. L. Wiskott, J.M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 19(7):775–779, July 1997.
36. C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, July 1997.
37. S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (magic) pointing. In *CHI99*, pages 246–253, 1999.