Effects of Task Properties, Partner Actions, and Message Content on Eye Gaze Patterns in a Collaborative Task

Jiazhi Ou, Lui Min Oh, Jie Yang, Susan R. Fussell
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
[jiazhiou, keithoh, jie.yang, sfussell]@cmu.edu

ABSTRACT
Helpers providing guidance for collaborative physical tasks shift their gaze between the workspace, supply area, and instructions. Understanding when and why helpers gaze at each area is important both for a theoretical understanding of collaboration on physical tasks and for the design of automated video systems for remote collaboration. In a laboratory experiment using a collaborative puzzle task, we recorded helpers' gaze while manipulating task complexity and piece differentiability. Helpers gazed toward the pieces bay more frequently when pieces were difficult to differentiate and less frequently over repeated trials. Preliminary analyses of message content show that helpers tend to look at the pieces bay when describing the next piece and at the workspace when describing where it goes. The results are consistent with a grounding model of communication, in which helpers seek visual evidence of understanding unless they are confident that they have been understood. The results also suggest the feasibility of building automated video systems based on remote helpers' shifting visual requirements.

Categories & Subject Descriptors: H5.3. Information interfaces and presentation (e.g., HCI): Group and organizational interfaces – collaborative computing, computer-supported collaborative work

General Terms: Experimentation, Human Factors

Keywords: Eye-tracking, computer-supported collaborative work, video mediated communication, video conferencing, gesture, conversational analysis, empirical studies

INTRODUCTION
Collaborative physical tasks are tasks in which two or more people work together on 3D objects in the real world. For example, surgical teams collaborate to operate on patients, telephone repair technicians provide instructions on how to fix equipment, and architects collaborate on building layouts. Because the expertise required for a task may not always be present at the worksite, there is growing demand for technologies to support remote collaboration on physical tasks.

A growing body of research demonstrates that video systems that focus on the workspace can provide value for remote collaboration on physical tasks (e.g., [14], [16], [20], [27]). Specifically, scene cameras providing static but wider views of the workspace appear to be more beneficial to collaborators than head-mounted cameras that show narrow, dynamic views or to audio-only systems. Systems that further incorporate an ability to point and gesture in the workspace are even more valuable (e.g., [15]).

Despite their successes, however, scene cameras have important limitations. Perhaps most problematic is that they are typically immobile, making it impossible for remote participants to pan the scene or zoom in on an area of interest in the workspace. Several systems have addressed this problem by allowing remote camera control (e.g., [30], [31], [35], [37]); however, the task of manipulating the camera interferes with smooth interpersonal communication. Investigators have also tried providing remote participants with multiple camera views (e.g., [16], [18]), but providing multiple views is bandwidth-intensive and has not proven beneficial to participants. In the present work, we consider a third possibility—combining the above strategies to provide the right camera feed at the right time during a task. More specifically, we investigate whether remote instructors' gaze patterns toward different visual resources show regularities that would enable us to predict what view of the workspace would be most beneficial at specific points of time during the task. Our long-term goal is to provide guidelines for automated camera control systems that present the right visual information at the right time.

In addition, we investigate how task properties such as the ability to discriminate between task pieces affect eye gaze patterns. Our results suggest that the use of visual information is predictable based on several parameters, including the global characteristics of the task (e.g.,

CHI 2005, April 2 – 7, 2005, Portland, Oregon, USA.

difficulty), the progress of the task (e.g., number of trials completed), and the worker's actions.

We investigate these issues using a real-time collaborative online task in which one partner (the "helper") instructs another (the "worker") on how to build a puzzle. We record the helper's eye gaze, the collaborators' dialogue, and the worker's actions in real time. Using this data, we investigate the relationship between the remote helper's focus of attention and task parameters. The results suggest that percentage of gaze toward different visual targets in the workspace is highly predictable based on these parameters.

In the remainder of this paper, we first present the theoretical framework guiding our work. Then, we describe our experimental method and the results of our study. We conclude with a discussion of the implication of our findings for the design of automated video systems.

THEORETICAL BACKGROUND

Collaborative Physical Tasks

When collaborators work on a task together, they have some information on the visual elements of their work environment. These pieces of information include the positions of work-related objects and tools and the status of the task itself (e.g., [12], [21], [43]). The collaborators take this visual information into account as they speak and act. Through conversation, the collaborators identify target objects to one another, describe actions to be taken and confirm the outcome of those actions taken.

In this paper, we focus on *instructional* tasks in which the collaborators' participation can be differentiated into either a helper role or worker role. The *helper* guides the worker to complete certain operations. The *worker* performs the actual task actions. Such a helper-worker relationship is similar to a teacher guiding a student on a lab project or a head resident teaching new doctors how to treat a patient.

Given the dynamic nature of collaborative tasks, helpers and workers must carefully coordinate their activities. A helper needs to know when it is appropriate to interrupt and provide assistance. After giving the advice, the helper needs to know if it has been comprehended as intended.

Visual information plays two roles in coordinating helper and worker communication and actions. First, it provides *situation awareness*—an ongoing awareness of the work environment and the activities taking place within it [10]. For example, if a teacher sees that a student is performing an incorrect operation, he or she can intervene to correct the student's mistake. Second, visual information can facilitate *conversational grounding,* or the interactive process by which communicators come to a state of mutual understanding ([5], [6], [7]). As seen in the following excerpt, each contribution builds up the common ground between collaborators.

H: "Ok, now take the salmon red piece"
W: "Er... which one? This one?"
H: "No. One shade darker."
W: "This?"
H: "Yeah."

Sometimes, contributions can be grounded immediately by an acknowledgment ("yeah", "uh huh"); other times, clarifications or corrections will be needed before the collaborators can reach common ground [25].

According to Clark and Marshall [5], there are three ways collaborators can establish common ground: through membership in a common group (*community co-membership*), by relying on previous communication (*linguistic co-presence*) and by sharing the same space (*physical co-presence*). In this research, we focus on physical co-presence and the visual resources it makes available for situation awareness and conversation grounding.

Shared Visual Space

We define *shared visual space* as the set of mutually visible entities, including participants' bodies and faces, people and objects in the work space, and the larger environment. When collaborators are co-present, they share substantial visual space, which they can rely upon when planning what to say or do next. Many previous studies (e.g., [14], [15], [16], [27], [36]) have found that pairs who worked side-by-side, where all sources of visual information are readily available, completed tasks more quickly and more accurately than pairs who worked apart and could only converse but not see each other's environment.

When people are co-present, they share several types of visual information, which vary in their importance for maintaining situation awareness and grounding conversation. For the purposes of the current study, it is useful to distinguish among three key resources: participants' heads and faces, participants' actions, and task objects [27]. As illustrated in Table 1, each of these visual resources has different benefits for collaboration depending upon the phase of the task.

Table 1. Visual resources in collaborative physical tasks (adapted from Kraut et al. [27]).

	Type of Visual Information		
Task Phase	*People's Heads/Faces*	*Actions*	*Objects*
Identify Objects	Gaze direction helps indicate objects	Use gestures to refer to task objects	Use gestures to refer to task objects
Procedural Instructions	N/A	Use gestures to demonstrate procedures	Create up-to-date descrip-tions of objects and locations
Monitor compre-hension	Observe facial expressions	Observe appropriate-ness of actions	Observe appropriateness of changes to task objects

Mediating Shared Visual Space in Remote Collaboration

While participants working side-by-side benefit from rich shared visual information, people working together at a distance must rely on some type of telecommunications, which limit the type and amount of visual information that can be shared. Most video systems can only provide a subset of the visual cues available in side-by-side collaboration. Benefits of video systems are task and situation dependent [47]. Studies show that task performance achieved by using such systems is either a middle ground between side-by-side and audio-only settings (e.g., [36], [16]), or not significantly better than audio-only (e.g., [14], [44]).

One improvement suggested by Gaver et al. [18] is to provide multiple video feeds that participants can switch between. Such an approach is problematic due to the high equipment requirements. Also, the ability to switch between video feeds made it difficult for users to understand what elements of the visual environment were shared.

An alternative strategy is to identify the key types of visual information used in side-by-side settings and to design or implement systems to provide the critical elements of shared visual space to remote collaborators. Fussell and colleagues [16], for example, assessed the value of two different video systems—a head-mounted video system with eye-tracking capability and a scene camera that provides a wider view of the work area—on a robot construction task. Each video system matches up to one or more sources of visual information provided by physical co-presence. Performance with the scene camera was faster than with audio-only, while the head-mounted camera with eye tracking capabilities provided little benefit. Moreover, the combination of head-mounted camera and scene camera led to longer performance times than the scene camera alone. The study highlights some of the difficulties of using video systems to deliver theoretically relevant visual cues.

A possible explanation for the marginal utility of the head-mounted camera is that the head-mounted camera shows where the worker is looking, not where the helper wants to look. Perhaps, when the worker is assembling a certain part of the robot, the helper wants to look around the workspace for the next piece. The head-mounted camera will only show the worker putting the piece together during this task phase and the helper is forced to be idle, as he or she cannot immediately access the desired visual information. In order to design effective video systems to support collaborative physical tasks, it is important to know what visual information the helper needs at each stage of the task. One way to investigate this issue is using eye-tracking methodology to track helpers' gaze patterns. We discuss this possibility in the next section.

Eye-Tracking as a Method for Understanding Visual Information in Collaborative Physical Tasks

Eye-tracking methods allow investigators to obtain a fine-grained understanding of people's use of visual information. A number of studies have used eye-tracking to investigate relationships between gaze and actions in non-collaborative physical tasks (e.g., [32], [33], [40]) and in complex athletic behaviors (e.g., [38], [46]).

Eye-tracking has also been used as a tool to understand interpersonal communication (e.g., [9], [26]). These studies have typically used a referential communication task in which one person describes a series of objects for another person, who must find the target in an array of alternatives.

Other studies have used eye-tracking to determine people's focus of attention in conversation. Vertegaal et al. [45] examined gaze at partners during a four-person conversation about current events and found that gaze strongly indicated participants' focus of attention. Stiefelhagen & Zhu [42] also studied gaze during four-party conversations with a focus on how head and eye movements were associated as cues of attention. Gullberg [22] studied gestures in conversational settings and found that consistent with previous research, (e.g., [1]) listeners do not always fixate speakers' gestures. In a slightly different vein, Dabbish and Kraut [8] used eye-tracking to investigate the effects of the degree of detail presented in online awareness notifications about a partner's status on the timing of electronic communications.

With the exception of the Dabbish and Kraut study, none of the studies described above looked at conversations in which participants had to manipulate objects or perform other physical activities while they were conversing. There are, however, some recent studies that provide strong evidence that people naturally look at objects or devices with which they are interacting. Campana and colleagues [3] describe a system that uses a speaker's eye movements to determine what he or she is referring to, and hence improve the performance of a dialogue system. Maglio and colleagues [34] investigated people's speech and gaze when interacting with an "office of the future" and found that subjects nearly always looked at the addressed device before making a request. Similar results are reported by Brumitt et al. [2].

In the most relevant study to the current investigation, Fussell et al. [17] used eye-tracking to determine the relative importance of different visual resources (e.g., partners' faces, partners' actions, task pieces) in a collaborative robot construction task. Results suggested that helpers look more often and for longer durations at the object being constructed, task pieces and tools, and the worker's hands than they look at the worker's face and other aspects of the work environment.

Although the Fussell et al. study is useful for understanding how helpers make use of visual information

during collaborative physical tasks, it suffered from some limitations that reduce its usefulness for system design. First, because the visual scene was constantly changing, gaze targets had to be hand-coded rather than automatically calculated from gaze coordinates within a set frame. This induces an undesirable level of error in the calculations. Second, there was no single unit of time that could be applied both to speech and to gaze coding, so it was impossible to investigate relationships between message content and visual requirements. Third, workers' actions were not taped or coded, so the investigators could not determine how these actions impacted helpers' gaze.

In the current study, we use a collaborative online jigsaw puzzle task adapted from Kraut and colleagues ([19], [20], [28]) to provide more detailed and accurate information about interrelationships among gaze, speech and actions. In this task, a remote helper verbally instructs a worker to arrange color blocks on a computer screen in a way that matches a target puzzle. The workers' work area can be yoked to the helper's screen, so that there is visual evidence of the worker's puzzle construction actions.

Previous studies using online puzzle tasks ([19], [20], [28]) suggest that it provides a useful analog of real-world 3D tasks in which variables of interest can be tightly controlled. In addition, the online puzzle task has two important benefits for the investigation of gaze: First, the helper's focus of attention can be automatically and robustly computed on a 2-D plane, allowing for a detailed understanding of where helpers are looking and minimizing errors induced by hand-coding gaze. Second, the time stamping of the automatically processed gaze coordinates can be lined up precisely with the time stamping of utterances and worker actions. This is generally not possible with 3D tasks.

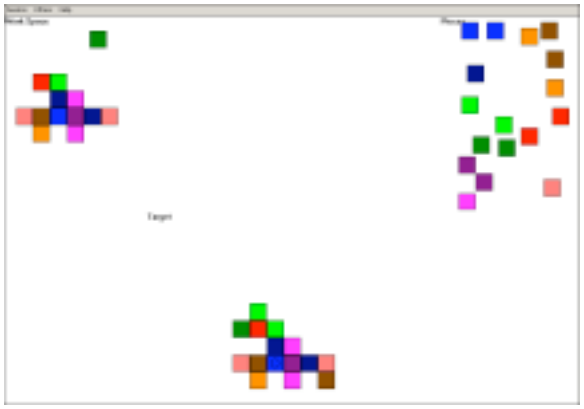

Figure 1. Helper's display in the online puzzle task. The solution to the puzzle (target) is shown in the center bottom. The pieces bay, from which pieces can be selected, is shown on the upper right. The workspace, where the worker is assembling the puzzle, is shown on the upper left.

THE CURRENT STUDY

In our version of the online jigsaw puzzle task, both the *workspace*, in which the worker is constructing the puzzle, and the *pieces bay*, in which the puzzle pieces are stored until use, are yoked to the helper's screen. Helpers can thus direct their gaze such that they obtain visual evidence of the piece the worker is selecting or visual evidence of the workspace where the worker is assembling the puzzle, but not both at the same time. In addition, the target puzzle solution is presented online as a third possible gaze target. (Figure 1).

In this study we analyze how helpers distribute their visual attention across the workspace, puzzle piece bay, and target puzzle. We give some examples of how the two sources of visual cues can facilitate different phases of the puzzle task in Table 2.

Table 2. Functions of the two shared visual sources in four sample task phases.

Task Phase	Visual Information	
	Workspace	Pieces Bay
Helper specifies a color piece	By looking at the current progress of the task the helper can decide which color piece the worker needs next	The helper needs to describe the color piece accurately
Helper specifies the location of the single color piece	The helper needs to know where the grabbed color piece should go	N/A
Worker grabs a color piece	The helper can plan how to describe the location for that piece	The helper can monitor whether the worker is grabbing the right piece
Worker positions a color piece in the workspace	The helper can confirm whether the worker positions the color piece correctly	N/A

Each puzzle, which is composed of a certain number of color blocks, is characterized by two parameters: complexity and color difficulty. We employed eye-tracking technology to compute and record the intersection of the helper's line of sight and the screen plane. We also tracked and recorded the worker's movements, both in the workspace and the pieces bay.

Hypotheses

To analyze the influence of the task characteristics on the helper's focus of attention, we examine four hypotheses. These hypotheses are formulated in terms of percentage

gaze directed at the pieces bay vs. the other two gaze targets (workspace, target puzzle):

1. Less differentiability among pieces will lead to greater percentage gaze at the pieces bay. We hypothesized that the helper would look at the pieces bay more when the colors are more difficult to differentiate. With difficult-to-differentiate colors, there is less certainty that the worker can correctly identify the piece and therefore a greater need for visual evidence of the workers' understanding.

2. Greater puzzle complexity will be associated with greater percentage gaze at the pieces bay. We hypothesized that the helper would look at pieces bay more when there are more color blocks in the task, again because of a greater need for visual evidence of the workers' understanding.

3. Gaze at the pieces bay will decrease over trials. We hypothesized that the helper would spend less time looking at the pieces bay over trials, because the helper will establish grounding of the colors with the worker. That is, with repeated successes, the helper will become more confident that the worker can identify the piece from his/her description and therefore require less visual evidence of comprehension.

4. Helpers' gaze toward the workspace and pieces bay will be correlated with actual worker actions in those areas. Our assumption is that helpers look in these regions for visual information because the evidence of the worker's understanding they require is, in fact, in these areas.

In addition to testing these hypotheses, we perform a preliminary exploration of the relationship between the content of helpers' messages and their eye gaze.

METHOD

Design

The design formed a 2 (piece differentiability) by 3 (puzzle complexity) by 9 (trial) factorial within-subjects study. The factors manipulated were whether the pieces were solid or shaded, the number of color pieces in the target puzzle and the difficulty in differentiating the color blocks. The three puzzles for each level piece differentiability (solid vs. shaded) were presented in a single block. The order of presentation of puzzles was counterbalanced across participants.

Participants

Twenty-four college undergraduate and graduate students, all with normal color vision, participated in this study for $15 each.

Equipment and Software

The eye tracking system consisted of an ISCAN RK-426PC pupil/corneal reflection tracker, an ISCAN HMEIS head mounted eye-imaging system with head tracking sensor, a Polhemus InsideTRAK magnetic position sensor, and a stand-alone scene camera (see Figure 2).

Before the task, the helper's display's 3-D coordinates relative to the Polhemus magnetic sensor were registered into the software. The experimenter then calibrated the helper's eyes by asking him/her to look at five stationary points on the display. After calibration, the software could compute the helper's line-of-sight with his/her eye position and head position. The intersection of the line-of-sight and the defined plane, which was the helper's display, was also calculated and overlaid on the video of the scene camera in real-time. The software recorded the calculated coordinates at a frequency of 60HZ.

The video of the scene camera overlaid with the helper's eye gaze and the movement of the worker were recorded on a Panasonic DV-VCR. Wireless microphones were used to record the conversation between the subjects.

Figure 2. System setup on the helper's side. The pupil/corneal reflection tracker and the head tracking sensor are on the helper's helmet, the Polhemus magnetic position sensor is behind the helper, and by the side from behind there is a stand alone scene camera taking videos of the helper's display.

The helper's display was designed in the way that the 3 areas (*workspace*, *pieces bay*, and *target*) were positioned in a triangular shape (Figure 1 above). Therefore, the helper could shift his/her eye gaze from one area directly to one of the other two areas. To obtain the helper's focus of attention at a specific time we registered the index of one of the areas with the recorded eye gaze coordinates at that time. Nevertheless, due to the zero error of the magnetic sensor and the pupil/corneal reflection tracker, the absolute coordinate itself is not a reliable metric. To overcome this, we made use of the global information: for each section we clustered all eye gaze points by using K-Means vector quantization (VQ) method. We chose 3 initial centers located triangularly. Within 10 iterations the algorithm converged and the outputs were 3 new centers. Given an eye gaze coordinate, we classified the focus of attention with the index of its closest center. An example of eye gaze

coordinate distribution from one section of the tasks and the clustering result are shown in Figure 3.

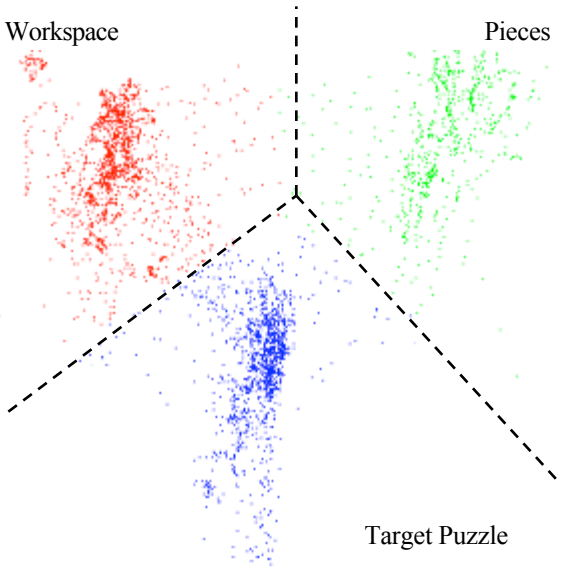

Figure 3. An example of eye gaze distribution from one section of the tasks. After running K-Means VQ algorithm we got 3 clusters and classified each point's focus of attention.

Materials

18 target color puzzles were created by randomly selecting color blocks and forming configurations of 5, 10 or 15 pieces (see Figure 4). There were 6 different puzzles for each level of complexity.

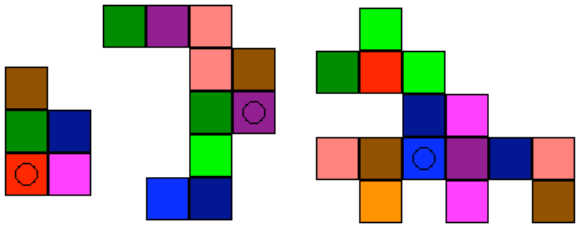

Figure 4. Examples of puzzle configurations with 5, 10, and 15 pieces, respectively.

For every puzzle complexity, three of the target color puzzles were formed from the pieces pool that was easily differentiated – there were at most two shades of the same color in the pieces pool. For example, there were only two different greens, bright green and dark green. The other three target color puzzles were formed from the pieces pool that was harder to differentiate—there were five shades of the same color in the pieces pool. For example, the RGB values of the five different greens were e6ffe6, b3ffb3, 80ff80, 4dff4d and 00dd00 (in hexadecimal). The orders of the 18 puzzles were counterbalanced by puzzle complexity and color difficulty.

LCD screen monitors were used as they displayed the colors more clearly that CRT monitors. Figure 5 shows the layout on the worker's screen.

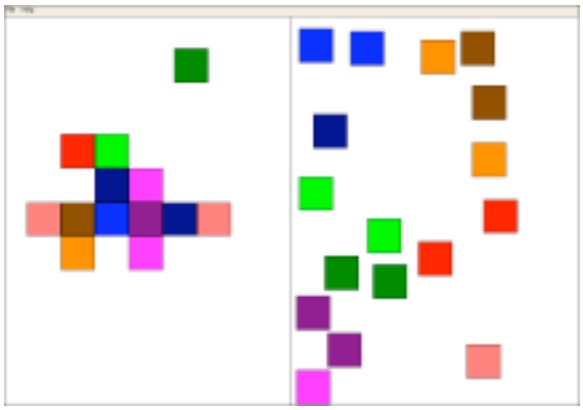

Figure 5. Worker's display, with the pieces bay shown on the right and the workspace shown on the left. Workers' actions in these areas were transmitted to the helper's display, as shown in Figure 1.

Procedure

Participants were randomly assigned to the helper and worker roles upon arrival for the study. They were then seated in the same room at their respective computer terminals with a barrier between them such that they could hear but not see one another.

The experimenter then calibrated the eye-tracker on the helper. After the calibration, the helper gave verbal instructions to the worker on how to move and arrange the color blocks from the pieces bay to the workspace such that configuration of the color blocks in the workspace matches that of the color blocks in the target area. The worker, who was allowed to converse freely with the helper and ask questions whenever necessary, would then drag-and-drop the color blocks based on the helper's instructions from the pieces bay to the workspace. The helper was able to see the worker's pieces bay and workspace and the worker's drag-and-drop movement as he or she gave the instructions.

When a puzzle was completed, the experimenter helped the helper select the next puzzle, such that the helper did not need to operate the computer at all. A 5-minute break was taken after 9 puzzles of the same color difficulty were completed to prevent eye-fatigue. After the break, the experimenter calibrated the eye-tracker on the helper again before resuming. Sessions lasted 60 to 90 minutes. During the tasks, the coordinates of the helper's eye gaze, the dialogue between the partners, and the movement of the worker are recorded.

RESULTS

Statistical Analysis

Data was analyzed using a mixed-model design in which subjects was a random factor and shading, puzzle complexity, trial, and block were fixed-subjects. This model takes individual differences in gaze into account while computing the fixed effects.

For the current analyses, we focus on percentage gaze directed at the pieces bay. However, because gaze toward the target (puzzle solution) remained relatively constant, gaze toward the pieces bay and gaze toward the workspace are inversely related (r = -.76). Consequently, the results for gaze toward the workspace show essentially the same pattern of significance but in the opposite direction. Overall, the fit of this model to the data was excellent (R Square = .69). A total of 18% of the variance was accounted for by the subject variable.

Puzzle Characteristics: Shading and Pieces

Gaze toward the pieces bay was significantly higher for shaded than for solid pieces ($F [1, 182] = 255.98$, $p < .0001$), supporting Hypothesis 1 (see Figure 6). Contrary to Hypothesis 2, however, gaze toward the pieces bay was significantly *lower* for puzzles with more pieces ($F [2, 182] = 11.28$, $p < .0001$). There was no interaction between shading and puzzle complexity ($F < 1$, *ns*).

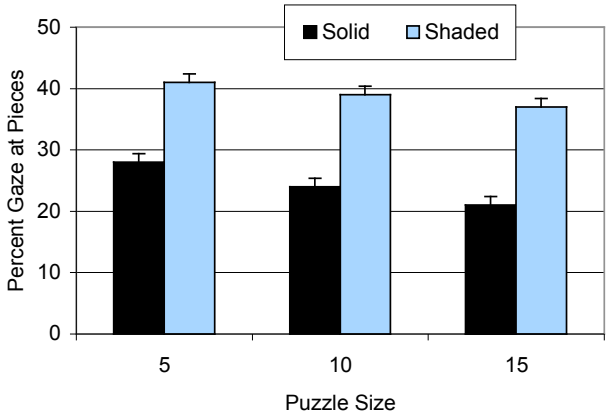

Figure 6. Percentage of gaze directed toward the pieces bay as a function of piece discriminability (shading) and puzzle size.

Effects of Trial

Participants completed 9 trials per block of solid or shaded pieces, grouped into 3 puzzles of each puzzle size (counterbalanced across participants). We hypothesized that helpers would spend less time monitoring the pieces bay over trials. Consistent with this hypothesis, we found a significant effect of trial ($F [1, 182] = 37.68$, $p < .0001$). However, as can be seen in Figure 7, the trial effect only held for the easy to describe (solid) pieces; for shaded pieces, gaze toward the pieces bay remained high across all trials (for the interaction, $F [1, 182] = 27.49$, $p < .0001$).

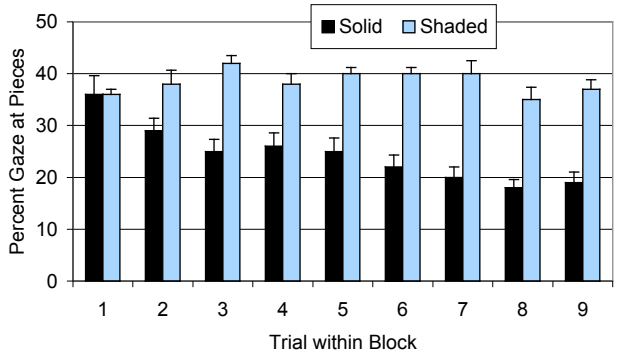

Figure 7. Percentage gaze directed at the pieces bay as a function of piece discriminability and trial.

Relationship between Gaze and Worker Actions

Worker actions were automatically detected and mapped onto the coordinates of the visual field to determine whether they occurred in the workspace or the pieces bay. Consistent with Hypothesis 4, gaze toward the workspace occurred 50% of the time that workers acted in this workspace. Gaze toward the pieces bay was higher when the worker was acting in that area, but just over 40% overall. This is probably due to the low levels of gaze toward the pieces bay for the solid colors in the later trials.

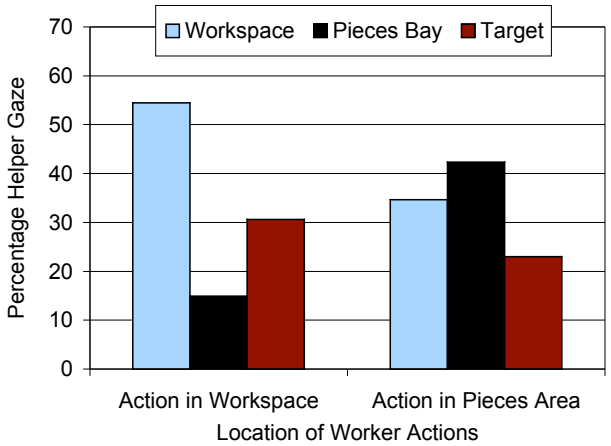

Figure 8. Helper gaze as a function of worker actions in the workspace and pieces bay

Relationships between Message Content and Gaze

In order to analyze the relationships between the content of helpers' instructions and their eye gaze, we devised an automated parsing program that separates each transcribed utterance into clauses describing the next piece of the puzzle and clauses describing where to place that piece in the puzzle. The start and end time of each clause were

labeled by an automatic speech recognition system, at accuracy of 10 milliseconds. We then computed eye gaze distributions in all clause segments. A preliminary evaluation of the parser showed that it correctly classified 94.5% of the clauses in 238 test sentences from five participants.

In Figure 9, we show the data from a sample of messages describing shaded puzzles from the first five participants in this study. This preliminary data clearly shows that gaze pattern varies as a function of the phase of instruction (description of the next piece vs. description of its location within the puzzle). When describing a piece, helpers overwhelmingly look at the pieces bay, whereas when they are describing a location, they are much more likely to look at the workspace.

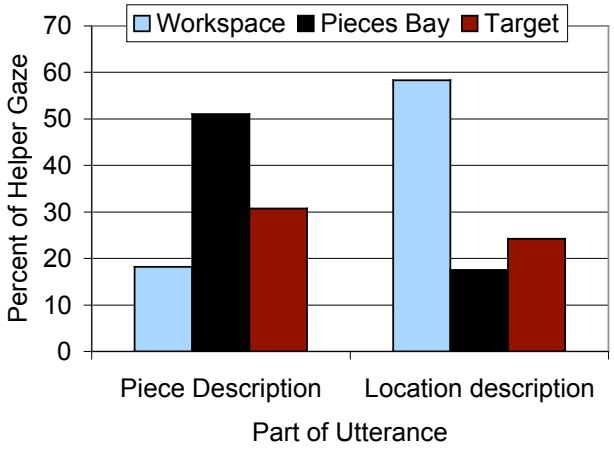

Figure 9. Relationship between helper message content and gaze toward workspace, pieces bay and target for shaded targets

DISCUSSION

The goals of this study were to understand how task properties such as the ability to distinguish among task pieces and repeated trials affect helpers' gaze during a collaborative task. Hypothesis 1 stated that when pieces were harder to discriminate (shades in the same color family), helpers would spend more time gazing at the pieces bay, to ensure that workers selected the correct piece on the basis of their instructions. This hypothesis was strongly confirmed.

Hypothesis 2 stated that when puzzles were more complex, as defined by the number of pieces involved, helpers would look more at the pieces bay. Instead, we found the opposite pattern of results. In retrospect, we believe that this is consistent with a view of gaze as a method of grounding utterances: As larger puzzles are constructed, there are fewer and fewer remaining pieces, making it less necessary for helpers to monitor the pieces bay for evidence of workers' comprehension.

Hypothesis 3 stated that over repeated trials, helpers would look less at the pieces bay because helpers and workers would have previously established common ground specifying what was meant by a given color description. This hypothesis was confirmed, but further analysis showed that the effect held only for the solid color, easy to distinguish pieces. For the more difficult shaded pieces, there was no effect of trial whatsoever. This suggests that even with prior evidence of comprehension, when task elements are particularly difficult to differentiate, helpers require visual evidence of workers' comprehension.

Hypothesis 4 stated that helpers' gaze would be correlated with actual worker actions. Although our results are consistent with this hypothesis, they are not conclusive. One problem is that the time frame within which these calculations are made may be too narrow. On average a worker drags/drops the mouse 3.22% of the time during the task. Moreover, taking a color piece from the pieces bay is much faster than positioning it in the workspace. The majority (75.80%) of the mouse movements happened in the workspace area.

A much more promising preliminary finding is that the part of the instruction currently being uttered is strongly associated with the primary focus of gaze. This suggests that automated real-time speech recognition and parsing techniques, once they are made robust, could be used to drive camera choices much as is done in automated auditoriums [11]. Although automated speech recognition remains a complicated technical challenge, we have found that for certain types of tasks, such as our puzzle or robot construction tasks (e.g., [15], [16]), the vocabulary appears to be sufficiently limited that automated speech recognition is feasible.

While the results suggest that automatic camera switching may be feasible, more work will be required before such a system can become a reality. Our current model can predict gaze toward the workspace or pieces bay at a significantly better than chance level, but it is far from 100% accurate. We are currently revising our models to improve predictive accuracy in two ways: First, we are performing full parsing of the transcripts to test whether the relationships we have found between part of utterance and gaze will hold across all subjects and trials. Second, we are using different windows of analysis to provide a better test of our hypothesis that helper gaze would be associated with worker actions. Finally, we are exploring new ways to do online prediction of sentence clause contents.

In future work, we will be conducting Wizard of Oz experiments to determine what level of accuracy in camera view selection people will find acceptable. We will also be broadening the scope of the work by applying our techniques to 3D collaborative physical tasks using an object marker system that will allow us to process gaze toward different targets as reliably in mobile tasks as we can in our current online puzzle simulation.

CONCLUSION

In summary, the results demonstrate that the amount of time helpers look at different targets can be reliably predicted by task characteristics and repeated trials. The results are consistent with a grounding view of communication, in which helpers seek visual evidence for workers' understanding when they lack confidence of that understanding either from a shared common vocabulary (color pieces) or previous interaction (trials). In addition, the preliminary results showing relationships between utterance parts and gaze suggest that with a robust real-time parser, camera shifts could be automated to show helpers what they need to see, at the time they need to see it.

ACKNOWLEDGEMENTS

This research was funded by National Science Foundation Grant #0208903. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Darren Gergle for his statistical advice, Leslie Setlock for her editorial assistance, and Xilin Chen for his help on setting up the Polhemus tracker. We also thank the five anonymous reviewers for their very helpful comments and suggestions.

REFERENCES

[1] Argyle, M., & Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge: Cambridge University Press.

[2] Brumitt B., Krumm J., Meyers B., & Shafer S. (2000). Let there be light: Comparing interfaces for homes of the future. *IEEE Personal Communications*, August 2000.

[3] Campana, E., Baldridge, J., Dowding, J., Hockey, B. A., Remington, R. W., & Stone, L. S. (2001). Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the 2001 workshop on Perceptive user interfaces* (pp. 1–5).

[4] Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47,* 30-49.

[5] Clark, H. H. & Marshall, C. E. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge: Cambridge University Press.

[6] Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* 1-39.

[7] Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

[8] Dabbish, L. & Kraut R. (2004). Controlling interruptions: Awareness displays and social motivation for coordination. *Proceedings of CSCW 2004* (pp. 182-191). NY: ACM Press.

[9] Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language processing in natural contexts. *Journal of Psycholinguistic Research, 24,* 409-436.

[10] Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37, 32*-64.

[11] Farid. M,, Murtagh., F., and Starck., J.L. (2002) Computer display control and interaction using eye-gaze, *Journal of the Society for Information Display, 10,* 289-293.

[12] Ford, C. E. (1999). Collaborative construction of task activity: Coordinating multiple resources in a high school physics lab. *Research on Language and Social Interaction, 32,* 369-408.

[13] Frey L. A., White, K. P. Jr., & Hutchinson T. E. (1990). Eye-gaze word processing. *IEEE Transactions on Systems, Man and Cybernetics, 20,* 944–950.

[14] Fussell, S. R., Kraut, R. E., & Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work. *Proceedings of CSCW 2000* (pp. 21-30). NY: ACM Press.

[15] Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E. M., & Kramer, A. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction, 19,* 273-309.

[16] Fussell, S. R., Setlock, L. D., & Kraut, R. E. (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. *Proceedings of CHI 2003* (pp. 513-520). NY: ACM Press.

[17] Fussell, S. R., Setlock, L. D., & Parker, E. M. (2003). Where do helpers look? Gaze targets during collaborative physical tasks. *CHI 2003 Extended Abstracts* (pp. 768-769). NY: ACM Press.

[18] Gaver, W., Sellen, A., Heath, C., & Luff, P. (1993) One is not enough: Multiple views in a media space. *Proceedings of Interchi '93* (pp. 335-341). NY: ACM Press.

[19] Gergle, D., Kraut, R.E., & Fussell, S.R. (2004). Action as language in a shared visual space. *Proceedings of CSCW 2004* (pp. 487-496). NY: ACM Press.

[20] Gergle, D., Millan, D. R., Kraut, R. E., & Fussell, S. R. (2004). Persistence matters: Making the most of chat in tightly-coupled work. *CHI 2004* (pp. 431-438). NY: ACM Press

[21] Goodwin, C. (1996). Professional vision. *American Anthropologist, 96,* 606-633.

[22] Gullberg, M. (2003). Eye movements and gestures in human face-to-face interaction. In J. Hyona, R. Radach, & H. Deubel, (Eds.) *The Mind's Eyes:*

Cognitive and Applied Aspects of Eye Movements (pp. 685-703). Oxford: Elsevier Science.

[23] Hutchinson T. E., White, K. P. Jr., Martin, W. N., Reichert, K. C., & Frey L. A. (1989). Human-computer interaction using eye-gaze input. *IEEE Transaction on Systems, Man, and Cybernetics, 19,* 1527–1534.

[24] Jacob, R. J. K. (1993). Eye-movement-based human-computer interaction techniques. In H. R. Hartson & D. Hix (Eds.), *Advances in Human-Computer Interaction, Vol. 4* (pp. 151–190). Norwood, NJ: Ablex.

[25] Jefferson, G. (1972). Side sequences. In D. Sudnow (Ed.) *Studies in social interaction* (pp. 294-338). NY: Free Press.

[26] Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11,* 32-38.

[27] Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction, 18,* 13-49.

[28] Kraut, R.E., Gergle, D., & Fussell, S.R. (2002). The Use of visual information in shared visual spaces: Informing the development of virtual co-presence. In *Proceedings of CSCW 2002* (pp. 31-40). NY: ACM Press.

[29] Kraut, R. E., Miller, M. D., & Siegel, J. (1996) Collaboration in performance of physical tasks: Effects on outcomes and communication. *Proceedings of CSCW 1996* (pp. 57-66). NY ACM.

[30] Kuzuoka, H., Kosuge, T., & Tanaka, K.. (1994) GestureCam: A video communication system for sympathetic remote collaboration. *Proceedings of CSCW 1994* (pp. 35-43). NY: ACM.

[31] Kuzuoka, H., Oyama, S., Yamazaki, K., Suzuki, K., & Mitsuishi, M. (2000). GestureMan: A mobile robot that embodies a remote instructor's actions. *Proceedings of CSCW 2000* (pp. 155-162). NY: ACM Press.

[32] Land, M., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities. *Vision Research, 41,* 3559-3565.

[33] Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28,* 1307-1432.

[34] Maglio P., Matlock T., Campbell C. S., Zhai S., & Smith, B. A. (2000). Gaze and speech in attentive user interfaces. *Proceedings of the International Conference on Multimodal Interfaces*. Springer.

[35] Oh, K., Kramer, A. D. I., & Fussell, S. R. (in preparation). *Comparison of scene and laser pointing video systems for remote collaboration on physical tasks*.

[36] Ou, J., Fussell, S. R., Chen, X., Setlock, L. D., & Yang, J. (2003). Gestural communication over video stream: Supporting multimodal interaction for remote collaborative physical tasks. In *Proceedings of International Conference on Multimodal Interfaces*, Nov. 5-7, 2003, Vancouver, Canada.

[37] Ou, J. (unpublished). *DOVE-2: Combining gesture with remote camera control*.

[38] Oudejans, R. R. D., Michaels, C. F., Bakker, F. C., & Davids, K. (1999). Shedding some light on catching in the dark: Perceptual mechanisms for catching fly balls. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 531-542.

[39] Daly-Jones, O., Monk, A. & Watts, L. (1998). Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *International Journal of Human-Computer Studies, 49*, 21-58.

[40] Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41,* 3587-3596.

[41] Salvucci, D. (1999). Inferring intent in eye-based interfaces: Tracing eye movements with process models. *Proceedings of CHI 1999* (pp. 254-261). NY: ACM Press.

[42] Stiefelhagen, R., Yang, J., & Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks, 13,* 928-938.

[43] Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies, 34*, 143-160.

[44] Veinott, E., Olson, J., Olson, G., & Fu, X. (1999). Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. *Proceedings of CHI 1999* (pp. 302-309). NY: ACM Press.

[45] Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. *Proceedings of CHI 2001* (pp. 301-308). NY: ACM Press.

[46] Vickers, J. N. (1996). Visual control when aiming at a far target. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 342-354.

[47] Whittaker, S., & O'Conaill, B. (1997). The role of vision in face-to-face and mediated communication. In K. Finn, A.Sellen & S. Wilbur (Eds.) *Video-Mediated Communication (*pp. 23-49). Mahwah, NJ: Erlbaum.