Analyzing and Predicting Focus of Attention In Remote Collaborative Tasks

Jiazhi Ou, Lui Min Oh, Susan R. Fussell, Tal Blum, Jie Yang
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{jiazhiou, keithoh, sfussell, tblum, jie.yang}@cmu.edu

ABSTRACT

To overcome the limitations of current technologies for remote collaboration, we propose a system that changes a video feed based on task properties, people's actions, and message properties. First, we examined how participants manage different visual resources in a laboratory experiment using a collaborative task in which one partner (the *helper*) instructs another (the *worker*) how to assemble online puzzles. We analyzed helpers' eye gaze as a function of the aforementioned parameters. Helpers gazed at the set of alternative pieces more frequently when it was harder for workers to differentiate these pieces, and less frequently over repeated trials. The results further suggest that a helper's desired focus of attention can be predicted based on task properties, his/her partner's actions, and message properties. We propose a conditional Markov model classifier to explore the feasibility of predicting gaze based on these properties. The accuracy of the model ranged from 65.40% for puzzles with easy-to-name pieces to 74.25% for puzzles with more difficult to name pieces. The results suggest that we can use our model to automatically manipulate video feeds to show what helpers want to see when they want to see it.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Remote Collaborative Tasks, Eye Tracking, Focus of Attention, Keyword Spotting, Computer-Supported Cooperative Work

1. INTRODUCTION

Imagine that an engineer in the United States needs to instruct a technician in India on how to service a faulty machine. They are forced by circumstances to collaborate over geographical distance. The participants in this collaborative task must work on physical objects. Their participation can be differentiated into a *helper* role (the person offering the knowledge to guide the operations) and a *worker* role (the person who actually performs the physical actions). Given the dynamic nature of collaborative tasks, helpers and workers must coordinate their interaction, so that assistance is provided in a timely manner.

Video systems have emerged to help bridge the distance between remote collaborators on physical tasks by providing them a shared visual space [15] for conversational grounding ([4], [6]). We are interested in how a better understanding of task dynamics can improve existing video systems.

Video systems that provide the helper with some view of the worker's environment have shown to help improve task performance, compared to audio-only systems ([8], [9], [11], [15]). However, these systems are handicapped by the reality of expensive bandwidth consumption and by the limited view angle and mobility of the camera. These systems, at best, can provide only a subset of the visual information available in side-by-side conditions. Some systems have attempted to address these issues by having a pan/tilt/zoom camera remotely controlled by the helper ([17], [18]); however, the task of manipulating the camera interferes with smooth interpersonal communication [21]. Other systems offering multiple views simultaneously are bandwidth intensive, and have not been proven to be beneficial [9]. Systems that allow switching between multiple views (e.g., [10]) circumvent bandwidth limitations, but incur high equipment costs and hinder common understandings of what view of the environment is being shared.

We hypothesize that if we can show the remote helper the desired view of the worker's environment in any specific instance of the task automatically, it will free the helper from having to control the camera. At the same time, the helper will have the necessary visual information to communicate effectively and assist the worker in the collaborative task. Before we can design a system that shows the right view at the right time, however, we need a better understanding of how properties such as the nature of the physical task, partners' actions, and speech characteristics affect helpers' visual attention.

To examine this issue, we created a real-time collaborative online task in which a remote helper instructs a worker on how to build a puzzle. We first investigated whether the helper's focus of attention towards the different visual resources showed

ICMI'05, October 4–6, 2005, Trento, Italy.

regularities by analyzing the effects of task properties, partner actions, and message content on the helper's gaze. The results, obtained from previous work [23], suggest that the most beneficial view of the workspace can be predicted based on these parameters. In the current study, we build on previous findings by creating and training a classifier that attempts to predict a helper's desired focus of attention. Following the development phase, we evaluated the classifier's performance. The results demonstrate the feasibility of creating intelligent camera systems based on task properties, partner actions, and message content.

In the remainder of the paper, we first discuss relevant previous work. Then, we describe the experiment in detail. Next, we present the experimental results, our proposed gaze prediction algorithm, and its subsequent evaluation. We conclude with a discussion of some remaining issues and our plans for future work.

2. RELATED WORK

The successful performance of collaborative physical tasks requires tight coordination of conversation and action. People plan their utterances by monitoring others' activities and changes in task status to determine what steps should be taken next. Video conferencing tools that provide remote helpers with views of a workspace lead to more efficient task performance than audio-only communications tools (e.g., [9][22]). However, performance with video systems rarely equals that of face-to-face interaction. This may be because video cameras may not show the right part of a workspace at the right degree of resolution at the right time.

Clark's theory of conversational grounding (e.g., [4] [5]) suggests that helpers will look at targets that help them determine whether or not their messages have been understood as intended. Other research has indicated that gaze patterns of speakers and listeners are closely linked to the words spoken, and help in the timing and synchronization of utterances (e.g., [1]). Vertegaal et al. found that in multi-party conversations, speakers looked at the person they were talking to 77% of the time and listeners looked at the speaker 88% of the time. They also built a multi-agent conversational system that uses eye gaze input to determine to which agent the user is listening or speaking [28]. Stiefelhagen et al. developed a system to estimate participants' focus of attention from gaze directions and sound sources. They demonstrated that acoustic information provides 8% relative error reduction compared to only using one modality [27].

Recent studies demonstrate that people naturally look at objects or devices with which they are interacting. Campana et al. describe a system that uses eye movements to determine what a speaker is referring to [3]. Maglio et al. investigated how people use speech and gaze when interacting with an "office of the future," in which they could interact with office applications (e.g. Calendar, Map and Address Book) via speech recognition, and found that people nearly always looked at a device before making a request to it [19]. Similaly, Brumitt et al. [2] investigated speech, gesture, touch, and other nontraditional interfaces to control lights in their "Easy Living Lab", a mock up of a small living room. They reported that people typically looked at the lights they wanted to control.

Eye gaze has been used as an important modality for building new human computer interfaces. Earlier work includes eye-controlled interfaces for the disabled [12], and eye gaze word processors [7]. In those interfaces, users can either make use of intentional or natural eye movements. Sibert and Jacob compared eye gaze with mouse input [25]. They found that eye-gaze selecting technology was faster than selecting with a mouse. Stiefelhagen and Yang illustrated a multimodal interface using eye gaze and speech to drive a panoramic image viewer [26].

A major problem of gaze-based interfaces is the difficulty in interpreting eye movement patterns due to unconscious eye movements such as saccades and to gaze tracking failure. Jacob [13] approached the problem by predicting a series of fixations separated by saccades and fitting the raw data to this model. Salvucci used hidden Markov models to interpret gaze data and reported good interpretation results in an eye typing study [24].

Oh et. al built a gaze-aware interface, "Look-To-Talk," (LTT) that could direct the speech to a software agent in a multi-user collaborative environment [20]. They compared LTT to "Talk-To-Talk" (TTT), a spoken keyword-driven paradigm, and "Push-To-Talk" (PTT), a keyboard-driven paradigm. They concluded that LTT is a promising interface. In this research, we are interested in the relationship between spoken utterances and gaze. Our goal is to predict focus of attention from keywords extracted from the dialogue during a collative physical task, i.e., Talk-to-Look.

3. METHOD

3.1 Design

Our experiment used an online jigsaw puzzle task adapted from Kraut and colleagues [16], in which a helper and worker collaborated to construct a series of puzzles. The helper could gaze freely among three areas to obtain visual information as he or she provided instructions:

- The *pieces bay*, in which the puzzle pieces were stored. By monitoring the pieces bay, the helper could assess whether the worker had selected the correct piece from among the alternatives.
- The *workspace,* in which the worker was constructing the puzzle. By monitoring the workspace, the helper could assess whether the worker had positioned a piece correctly.
- The *target solution*, which showed how the puzzle should be constructed. This appeared only on the helper's screen.

We manipulated the differentiability of the puzzle pieces (solid colors vs. shaded) and the complexity of the puzzle (5, 10 or 15 pieces). Each participant completed three puzzles for each condition (piece differentiability x puzzle complexity), randomly presented in a single block. The design formed a 2 (piece differentiability) x 3 (puzzle complexity) x 9 (trial) factorial within-subjects study. The order of the puzzle blocks was counterbalanced across participants.

3.2 Materials

We created 18 target puzzles by randomly selecting color pieces and forming configurations of 5, 10 or 15 pieces (see Figure 1). There were 6 different puzzles for each level of complexity, three formed from a pieces pool with solid colors (easier to describe), and three formed from a pieces pool with shaded colors (harder to describe). In the former case, there were at most two shades of the same color in the pieces pool (e.g., two distinctly different greens, such as bright green and dark green); in the latter case, there were five shades of the same color.

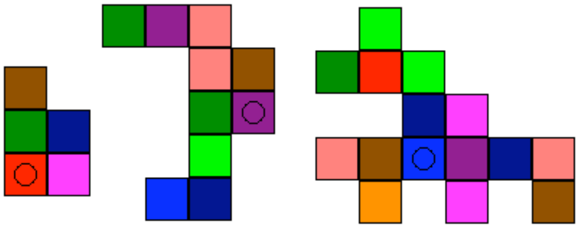

Figure 1. Examples of puzzle configurations with 5, 10, and 15 pieces, respectively.

The worker's screen was laid out so that the workspace and pieces bay were adjacent to each other. The helper's display was designed such that the 3 areas (*workspace*, *pieces bay*, and *target*) were in a triangular shape. The helper could shift his/her eye gaze from any area directly to any of the other two areas.

3.3 Equipment and Software

LCD monitors were used for displays and adjusted for color consistency. Sony wireless microphones were used to record the conversation between the subjects on separate channels.

An eye tracking system, consisting 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, was calibrated to each helper and recorded the intersection of the helper's line of sight with the screen plane at 60 Hz. The video feed of the scene, showing the coordinates of the helper's eye gaze and the worker's actions, was then recorded using a Panasonic DV-VCR.

The helper's focus of attention on any one of the 3 areas (*workspace*, *pieces bay*, and *target*) over time was derived from eye gaze coordinates. To overcome the unreliable metric posed by the zero error of the magnetic sensor and the pupil/corneal reflection tracker, we clustered all gaze coordinates in each session using K-Means vector quantization (VQ). We first chose 3 initial centers in the same triangular fashion as the 3 areas on the helper's display. Within 10 iterations, the algorithm converged and the outputs were 3 new centers. Subsequently, the helper's gaze coordinates were indexed based on their proximity to these 3 new centers. An example of the clustering result of the eye gaze coordinates for one of the sessions is shown in Figure 2.

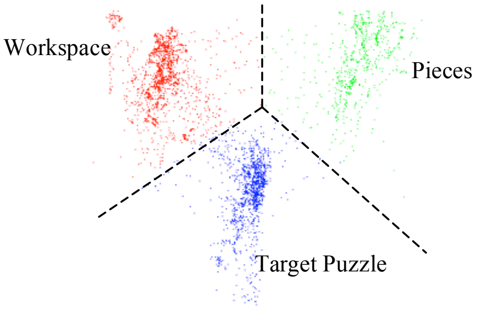

Figure 2. An example of eye gaze distribution from one section of the tasks. After running K-Means VQ algorithm we got 3 clusters and classified each point's focus of attention.

3.4 Participants and Procedure

Twenty-four college undergraduate and graduate students, all with normal color vision, participated in this study. Participants were randomly assigned to the helper and worker roles. They were seated in the same room at their respective computer terminals with a barrier between them so that they could hear but not see one another, simulating a remote collaboration environment.

The experimenter first calibrated the eye-tracker on the helper. After calibration, the helper gave verbal instructions to the worker on how to select puzzle pieces from the pieces bay and assemble them properly in the workspace to complete the target puzzle. The worker was allowed to converse freely with the helper and ask questions whenever necessary. The helper was able to see the worker's actions in the pieces bay and workspace. In order to prevent eye fatigue, participants were given a 5-minute break after half of the puzzles were completed. After the break, the experimenter recalibrated the eye-tracker. Sessions lasted 60 to 90 minutes.

4. STATISTICAL ANALYSIS

We employed statistical analysis to look at the relationships between the helper's gaze pattern and the following three factors: task properties, worker's actions and message content. Details of the analysis were presented in [23].

4.1 Gaze and Task Properties

To look at how task properties affect the gaze, we used 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 this analysis, 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.

As shown in Figure 3, gaze toward the pieces bay was significantly higher for shaded than for solid pieces (F [1, 182] = 255.98, p < .0001), and 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*).

We also found a significant effect of trial (F [1, 182] = 37.68, p < .0001), indicating that helpers spent less time monitoring the pieces bay over trials. However, as can be seen in Figure 4, 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).

4.2 Gaze and Worker's Actions

Worker actions in the workspace and pieces bay were automatically detected. As anticipated, gaze toward the workspace was higher when the worker was acting in that area, and vice versa when the worker was acting in the pieces bay (see Figure 5). However, the effect of worker actions on gaze was not significant for solid-color puzzles. We assume this is because the solid colors are easy to describe and distinguish so that the helper can be confident that the worker is grabbing the right piece without monitoring the pieces bay.

4.3 Gaze and Instructional Content

In order to analyze the relationships between the helper's eye gaze and the content of their instructions, two coders separated transcribed utterances into clauses and coded each of them as one of the categories shown in Table 1. In a subset of the data, the two coders agreed with each other 95% of the time.

Table 1. Coding of clauses

Code	Instructional Content
0	Description of color piece, e.g. "Take the green block"
1	Description of location, e.g. "And then put that to the right of the dark gray"
2	Correcting color piece, e.g. "A little lighter than that"
3	Correcting location, e.g. "It's on the very right"
4	Others

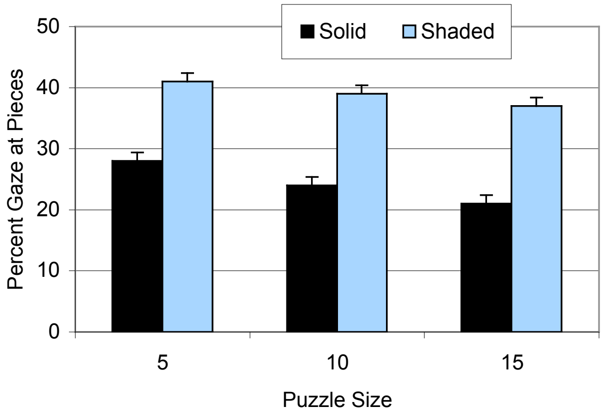

Figure 3. Percentage of gaze directed toward the pieces bay as a function of piece discriminability (shading) and puzzle size.

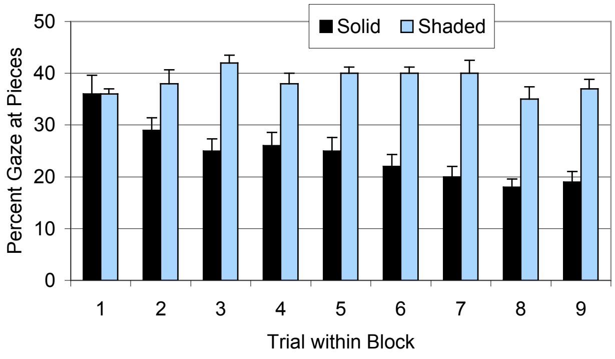

Figure 4. Percentage gaze directed at the pieces bay as a function of piece discriminability and trial.

We then computed eye gaze distributions for all clause segments as a function of clause coding. Distributions were computed to 1/60 second (Figure 6). The results illustrate that gaze pattern varies as a function of clause coding (description/correction of the next piece vs. description/correction 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.

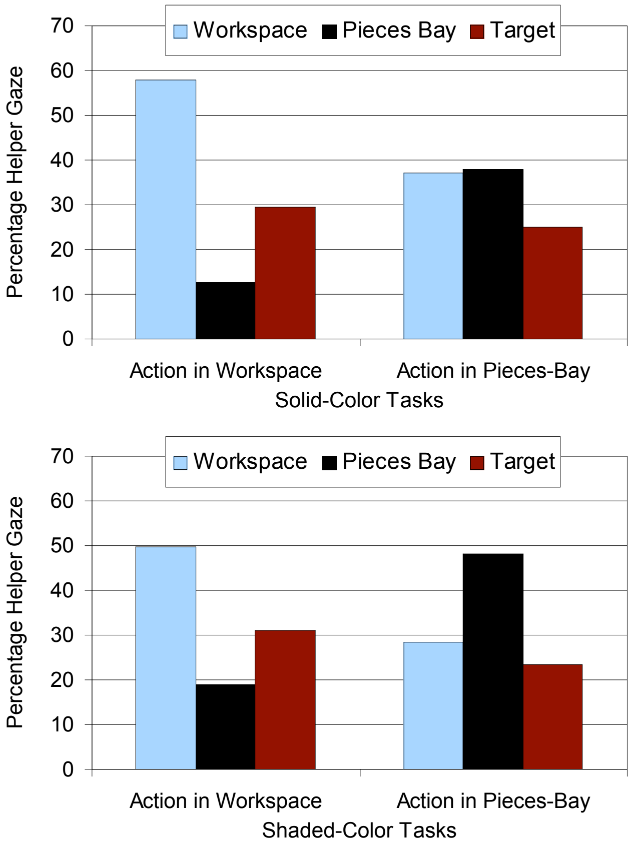

Figure 5. Helper gaze as a function of worker actions in the workspace and pieces bay

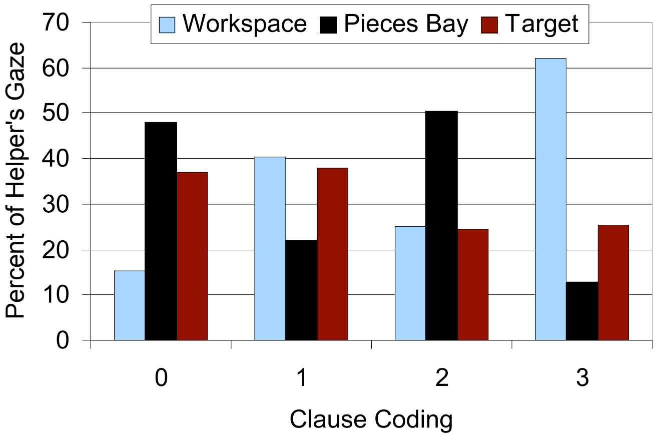

Figure 6. Relationship between helper message content and gaze toward workspace, pieces bay

5. PREDICTING FOCUS OF ATTENTION

The statistical analysis shows that a helper's gaze is highly related to the dialogue content and worker's actions. We implemented a conditional Markov model to predict the helper's gaze and explore the possibility of building an automated camera system to support remote physical collaboration.

5.1 The Problem

We define the problem as a classification problem. Given a section of the puzzle task with the sequence of transcribed words as $\{w_1, w_2, \ldots, w_N\}$, their corresponding start times derived from the wave signals by the speech recognizer as $\{s_1, e_1, s_2, e_2, \ldots, s_N, e_N\}$, then the worker's action m_t at time t is obtained by the analysis of the mouse click/move events. $m_t \in \{-1,0,1\}$, where -1, 0, 1 are the codes for Not-Moving, Moving-in-Workspace, and Moving-in-Pieces-Bay respectively. g_t is defined as the helper's gaze coded at time t,

$g_t \in \{0,1,2\}$, where 0, 1, 2 are the codes for workspace, pieces bay, and target respectively. m_t and g_t were processed at a sampling rate of 60HZ. An example of how a helper's dialogue and gaze, and the worker's action are synchronized is shown in Figure 7.

At each time t we predict the helper's gaze as $\hat{g}_t$. In our evaluation, we only consider gaze to the workspace and pieces bay, ignoring gazes towards the target. Let g_t, t = 1 .. T, be the actual gaze codes collected from the experiment and processed by VQ algorithm (Section 3.3). The classification error at time t is defined as:

$$err(\hat{g}_t, g_t) = \begin{cases} 1, & if \quad g_t = 0 \ \ or \ \ 1, \ \ and \ \ \hat{g}_t \neq g_t \\ 0, & otherwise \end{cases} \qquad (1)$$

and the performance of the classifier in one puzzle task, Acc is defined as:

$$Acc = 1 - \frac{\sum_{t=1}^{T} err(\hat{g}_t, g_t)}{|\{g_i \mid g_i = 0 \ \ or \ \ 1\}|}, \qquad (2)$$

where $|\{g_i \mid g_i = 0 \ \ or \ \ 1\}|$ is the number of gazes excluding those towards the target.

5.2 Offline Prediction

For comparison purposes, we first used human dialogue coding to perform the classification offline. The helper's gaze at time t is predicted using maximum likelihood estimation:

$$\hat{g}_t = \arg\max_{j=0 \ or \ 1} \Pr\{j \mid clause_t, m_t\}, \qquad (3)$$

while $clause_t \in \{0,1,2,3\}$ is the clause coding in Table 1 (Section 4.3) at time t. In both training and testing phases the clause coding and the worker's actions are known. We estimated the conditional probabilities of each gaze target from training sample frequencies.

5.3 Online Prediction

As our objective is to control the camera automatically in a video system (such as [22]) for remote collaborative physical tasks, the helper's focus of attention has to be predicted in advance based on previous information from the dialogue and worker's actions. Using the predictive model, the camera shifts between the workspace and the pieces bay. We do not need to predict camera shift to the target puzzle (solution), as it is always available at the helper's side. Moreover, in online prediction, the system does not have supervised knowledge of clause boundaries and coding.

We formulate online prediction as: at each sampling point t, given the previous words (w_1, w_2, …, w_i), and the previous worker's actions (m_1, m_2, …, m_t) as input, classify the next gaze code $\hat{g}_{t+1}$ as 0 (workspace) or 1 (pieces bay). The classification problem is illustrated in Figure 8.

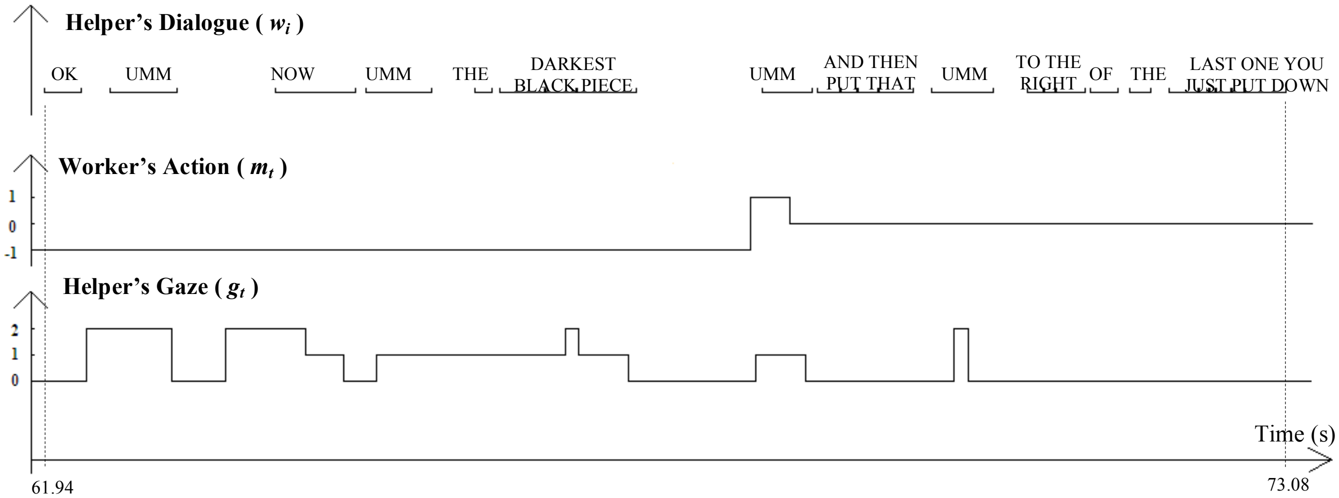

Figure 7. Demonstration of the three sources of data in a 12-second period. The helper's was giving the instruction "OK UMM NOW UMM THE DARKEST BLACK PIECE UMM AND THEN PUT THAT UMM TO THE RIGHT OF THE LAST `ONE YOU JUST PUT DOWN". Starting time and ending time of each word are aligned with the worker's action (-1: Not-Moving, 0: Moving-in-Workspace, 1: Moving-in-Pieces-Bay) and the helper's gaze (0: Workspace, 1: Pieces-Bay, 2: Target).

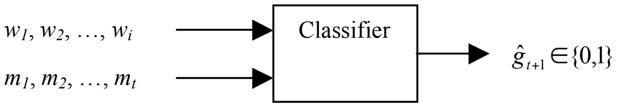

Figure 8. A multimodal classifier. $w_1, w_2, \ldots, w_i$ and $m_1, m_2, \ldots, m_t$ are dialogue and action data collected until point t, respectively; $\hat{g}_{t+1}$ is the classified gaze code.

Winner-Takes-All Strategy

Compared with the worker's action data (of which the majority is -1, or Not-Moving), the helper's utterances are a much richer source of information. Due to the difficulty in capturing the fluctuations of the helper's gaze within the start and end times of a single word, we don't expect to predict the gaze well at every sampling time. Therefore we apply a winner-takes-all strategy and smooth the helper's gaze and worker's action data based on the boundary of words. That is, the decision is only made at the end of each word. Let w_i, s_i, e_i be the ith word and its starting and ending time, and define the smoothed action M_i and gaze G_i as:

$$M_i = \begin{cases} -1, if \quad no \quad movements \quad between \quad time \quad e_{i-1} \quad and \quad e_i \\ \underset{j=0 \ or \ 1}{\arg\max} \, |\{m_t \mid m_t = j, e_{i-1} < t \le e_i\}|, \qquad otherwise \end{cases} \tag{4}$$

$$G_i = \underset{j=0 \ or \ 1}{\arg\max} \, |\{g_t \mid g_t = j, e_{i-1} < t \le e_i\}| \tag{5}$$

M_i and gaze G_i are interpreted as the majorities of action and gaze codes between time e_{i-1} and e_i (ignoring the target area). This process is graphically shown in Figure 9.

Now the problem becomes: given $w_1, w_2, \ldots, w_i$ and $M_1, M_2, \ldots, M_i$ as input features, output the prediction of G_{i+1} as 0 (workspace) or 1 (pieces bay).

A Conditional Markov Model Classifier

Since clause boundaries and coding have proven to be very useful for gaze prediction and are not available in online settings, we predict the clause coding of each word. To capture the dependencies between current word/action and previous word/action more directly we propose a conditional Markov model. Pairing gaze (G = 0 or 1) and clause coding (*Clause* = 0, 1, 2, or 3), we formed a sequence of 8 possible states. Let $\vec{W}$ and $\vec{M}$ be the word sequence and action sequence respectively. The probability of a state sequence $\vec{S}$ conditioned on the observation sequences $\vec{W}$ and $\vec{M}$ is inferred through factors ψ , ϕ, and φ:

$$\Pr\{\vec{S} \mid \vec{W}, \vec{M}\} \propto \prod_i (\psi(s_{i-1}, s_i) * \phi(w_i, w_{i-1}, w_{i-2}, s_i) * \varphi(M_i, M_{i-1}, M_{i-2}, s_i)) \tag{6}$$

and we define:

$$\begin{aligned} \psi(s_{i-1}, s_i) &= \Pr\{s_i \mid s_{i-1}\} \\ \phi(w_i, w_{i-1}, w_{i-2}, s_i) &= \Pr\{w_i, w_{i-1}, w_{i-2} \mid s_i\} \\ \varphi(M_i, M_{i-1}, M_{i-2}, s_i) &= \Pr\{M_i, M_{i-1}, M_{i-2} \mid s_i\} \end{aligned} \tag{7}$$

ψ captures the relationship between the current state and the previous state, while ϕ and φ characterize features $\vec{W}$ and $\vec{M}$ using trigram. Figure 10 shows the factor graph representation of the conditional Markov model.

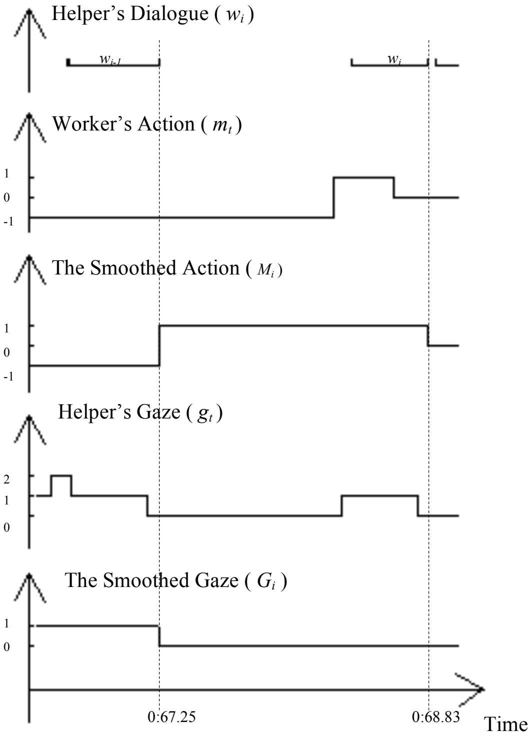

Figure 9. The smoothed action and gaze data based on the Winner-Takes-All strategy (Eq. (4) and Eq. (5)).

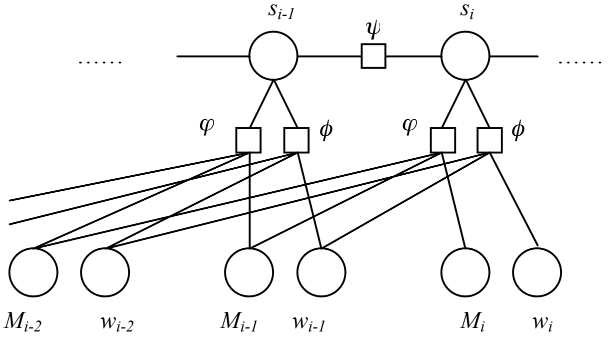

Figure 10. A conditional Markov model.

To decrease the number of parameters, we classify each word as one of the following 13 categories: *Color_Name* (e.g., "red"), *Preposition* (e.g., "above"), *Adjective* (e.g., "darkish"), *Verb* (e.g., "take"), *Linking_Verb* (e.g. "are"), *Noun* (e.g., "box"), *Pronoun* (e.g., "you"), *Positive_Feedback* (e.g., "yes"), *Negative_Feedback* (e.g., "wrong"), *Adverb* (e.g., "very"), *Conjunction* (e.g., "and"), *Non_Word_Utt* (e.g., "umm"), *Other*. ϕ is approximated by:

$$\begin{aligned} \phi(w_i, w_{i-1}, w_{i-2}, s_i) &= \Pr\{w_i, w_{i-1}, w_{i-2} \mid s_i\} \\ &\approx \Pr\{c_i, c_{i-1}, c_{i-2} \mid s_i\} \end{aligned} \tag{8}$$

where c_i is the category of w_i.

Since training is supervised, we employ a maximum likelihood method to learn the parameters, and Good-Tuning smoothing to estimate the unseen trigram in the training data. Moreover, as discussed in Section 4, the gaze distribution varies as a function of task characteristics. Therefore we estimate two sets of parameters, one for solid puzzles and the other for shaded puzzles. In testing we use Viterbi algorithm to find the optimal path given the parameters and observations. The performance of the classifier was evaluated according to Eq. (2).

5.4 Experimental Results and Discussion

To test our classifier, we used the data described earlier (i.e., 216 puzzle tasks from 12 pairs of subjects). We trained and tested with solid and shaded puzzles separately. Given puzzles of the same color differentiability, we used half of the data for training and the other half for testing. Then we switched the training set and testing set and performed the experiment again.

Offline and online gaze prediction accuracies for solid and shaded puzzles are listed in Table 2. In predicting gaze, the accuracies for shaded puzzles are significantly higher than those for solid puzzles. It can be explained by the statistical analysis in Section 4.2 (Figure 5), showing that data from workers' actions are more discriminative for shaded puzzles. Online prediction is not as good as offline prediction because it is a more challenging problem given that clause boundaries and coding are unknown.

Table 2. Test set accuracies in predicting helpers' focus of attention

Task Property	Offline Prediction Accuracy	Online Prediction Accuracy
Solid Puzzles	69.81%	65.40%
Shaded Puzzles	76.62%	74.25%

Tables 3 and 4 show the confusion matrixes. For solid puzzles, the system classified the workspace better, most likely because the prior probably of gaze toward the workspace is higher than that of gaze toward the pieces bay (see Section 4.3). In contrast, for shaded puzzles, the pieces bay had a higher prior probability and the classifier was better at classifying it.

Table 3. The confusion matrix for solid puzzles.

	0 (Workspace)	1 (Pieces Bay)
	Offline Prediction	
0 (Workspace)	79.56%	20.44%
1 (Pieces Bay)	43.13%	56.87%
	Online Prediction	
0 (Workspace)	75.17%	24.83%
1 (Pieces Bay)	47.24%	52.76%

Table 4. The confusion matrix for shaded puzzles.

	0 (Workspace)	1 (Pieces Bay)
	Offline Prediction	
0 (Workspace)	61.11%	38.89%
1 (Pieces Bay)	12.73%	87.27%
	Online Prediction	
0 (Workspace)	57.56%	42.44%
1 (Pieces Bay)	15.42%	85.58%

To examine the success of the conditional Markov classifier in predicting dialogue content, we define the instructional coding prediction accuracies as the percentage of correctly classified words. The accuracies for solid-color and shaded-color puzzles were 59.00% and 48.37%, respectively. Prediction of instructional coding for solid-color puzzles was much better because helpers used simpler language to describe the puzzle pieces.

6. CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated the feasibility of predicting focus of attention in remote collaborative tasks. The statistical analyses demonstrate that in remote collaboration, the percentage of time a helper looks at different targets is predictable from task properties, the worker's actions, and message content. The results are consistent with a conversational grounding view of collaboration (cf. [4]): When a helper lacks confidence that his/her instructions were understood in the context of previous interactions and a shared common vocabulary, he/she seeks additional visual evidence of understanding from the worker's environment.

Based on our analysis, we formulated the problem of predicting gaze in remote collaboration as a multimodal classification problem. We further employed a conditional Markov model to predict gaze as well as clause coding in real time. The experimental results show that overall accuracy is 65.40% for solid color puzzles and 74.25% for shaded puzzles. These results indicate the feasibility of developing intelligent video camera systems that predict where a helper wants to look in real time during remote collaboration. Such a system can optimally use network resources and enhance remote collaboration.

Our future research will follow up on three aspects of our findings. First, our classifier was more accurate when discriminative workers' action data was available. While it is easy to obtain this data during online collaborative tasks, it is much more difficult in 3D tasks. This suggests the need to monitor and interpret the worker's actions in physical collaborations. Second, our current model does not take workers' messages into account. Incorporating this information should enhance overall accuracies of prediction. Finally, our accuracy rates were higher than random guessing but still far from perfect. We plan to conduct behavioral research to determine how good an intelligent video camera system must be in order to be beneficial in practical use.

7. ACKNOWLEDGMENTS

This work 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, and Jeffrey Wong for the discussion sessions.

8. REFERENCES

[1] Argyle, M. & Cook, M. (1976). *Gaze and Mutual Gaze.* 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. *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, pp. 1–5.

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

[5] Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. B. Resnick, R. M. Levine, & S. D. Teasley (Eds.). *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA.

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

[7] Frey L. A., White K. P. Jr., Hutchinson T. E. (1990). Eye-gaze word processing. *IEEE Transactions on Systems, Man and Cybernetics*, 20(4), 944–950.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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, pp. 1527–1534.

[13] 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.

[14] Kraut, R. E., Fussell, S. R., Brennan, S., & Siegel, J. (2003). A framework for understanding effects of proximity on collaboration : Implications for technologies to support remote collaborative work. In P. Hinds & S. Kiesler (Eds.). *Technology and Distributed Work*.

[15] 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.

[16] 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. *Proceedings of CSCW 2002* (pp. 31-40). NY: ACM Press.

[17] 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.

[18] 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.

[19] Maglio, P. P., Matlock T., Campbell C. S., Zhai S., Smith B. A. (2000). Gaze and speech in attentive user interfaces. In *Proceedings of the International Conference on Multimodal Interfaces*, volume 1948 from LNCS. Springer, 2000.

[20] Oh, A., Fox, H., Kleek, M. V., Adler, A., Gajos, K., Morency, L., Darrell, T., Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment (2002), In *Proceedings of CHI '02 extended abstracts on Human factors in computing systems*, pp. 650 – 651.

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

[22] 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.

[23] Ou, J., Oh, L.M., Yang, J., & Fussell, S. R. (2005). Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 231 – 240. ACM Press.

[24] Salvucci Dario D. (1999). Inferring Intent in Eye-Based Interfaces: Tracing Eye Movements with Process Models. In *Human Factors in Computing Systems*: CHI 99, 1999.

[25] Sibert, L. E., and Jacob, R. J. (2000). Evaluation of Eye Gaze Interaction, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 281 – 288.

[26] Stiefelhagen, R., Yang, J. (1997). Gaze Tracking for Multimodal Human-Computer Interaction. In *Proceedings of International Conf. on Acoustics, Speech, and Signal Processing*, April 1997.

[27] 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.

[28] 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.