Predicting Visual Focus of Attention From Intention in Remote Collaborative Tasks

Jiazhi Ou, Lui Min Oh, Susan R. Fussell, Tal Blum, and Jie Yang

***Abstract*—While shared visual space plays a very important role in remote collaboration on physical tasks, it is challenging and expensive to track users' focus of attention (FOA) during these tasks. In this paper, we propose to identify a user's FOA from his/her intention based on task properties, people's actions in the workspace, and conversational content. We employ a conditional Markov model to characterize a subject's FOA. We demonstrate the feasibility of the proposed method using a collaborative laboratory task in which one partner (the helper) instructs another (the worker) on how to assemble online puzzles. We model a helper's FOA using task properties, workers' actions, and conversational content. 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 proposed model can be used to predicate a user's FOA in a remote collaborative task without tracking the user's eye gaze.**

***Index Terms*—Computer-supported cooperative work, eye tracking, focus of attention, keyword spotting, remote collaborative tasks.**

I. Introduction

COLLABORATIVE physical tasks play an important role in many domains, including education, industry, and medicine. These are tasks in which two or more people interact with real objects in the 3-D world. In instructional collaborative tasks, the focus of the current paper, participants are assigned to either a *helper* role or a *worker* role. The helper offers the knowledge to guide the operations, while the worker provides the physical labor. Such a relationship is similar to a teacher instructing a student in a physics experiment or an engineer guiding a technician servicing a vehicle. As businesses gear towards globalization, it becomes more possible for collaborators on physical tasks to work together at a distance. For example, an engineer in the United States might guide machinery repairs in India. In this paper, our goal is to provide a better understanding of the dynamics of instructional collaborative tasks that can help improve existing multimedia systems such that they support remote collaboration more efficiently and effectively.

Existing multimedia systems usually provide audio and a video feed of the space of interest to remote collaborators. This shared visual space [22] facilitates conversational grounding [5] and provides situational awareness [10] of ongoing activities in the workspace. Research has shown that video systems that focus on the workspace improve task performance ([12], [13], [22]). Specifically, scene cameras that provide a static, wide view of the workspace appear to be more beneficial for remote collaboration on physical tasks than head-mounted cameras or audio-only systems. Systems that incorporate the ability to point and gesture in the workspace further improve performance (e.g., [14], [34]).

Although scene cameras improve performance on physical tasks over audio-only communication, they are still not as good as face-to-face interaction. A major limitation is that the visual information available to remote collaborators is confined by the viewing angle and mobility of the camera. One way to address this problem is to set up multiple cameras, each providing a different view of the workspace [13]. Such a system is bandwidth-intensive and has not proven beneficial to participants. Systems that allow switching between alternative views of the workspace (e.g., [15]) may circumvent bandwidth limitations but they incur high equipment costs and hinder common understandings among the collaborators about what view of the environment is being shared. Some systems offer a helper the ability to pan, tilt or zoom cameras remotely ([24], [25]). However, the task of manipulating the camera interferes with smooth interpersonal communication [33].

The overall objective of our research is to develop another solution—"intelligent" video systems that provide the right camera feed at the right time during a collaborative physical task. Specifically, we envision a system that can automatically show a remote helper the most beneficial view of the worker's environment at each point in time. This system frees the helper from having to control the camera manually yet provides him or her with the necessary visual information to assist the worker effectively. In order to implement such a system, we must be able to predict the helper's focus of attention (FOA).

In this paper we propose to model FOA from intention without tracking a person's gaze direction. We use the term "in-

Manuscript received August 31, 2006; revised February 21, 2008. First published October 3, 2008; current version published October 24, 2008. This work was supported by the National Science Foundation under Grants IIS-0208903, IIS-0325047, IIS-0329077, and CNS-0551554. 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. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. John R. Smith.

J. Ou is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: jiazhiou@cmu.edu).

L. M. Oh was with the Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA. He is now with the Defense Science and Technology Agency, Singapore 118230 (e-mail: keithoh@alumni.cmu.edu).

S. R. Fussell and J. Yang are with the Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: sfussell@cmu.edu; yang@cs.cmu.edu).

T. Blum was with the Language Technologies Institute at Carnegie Mellon University, Pittsburgh, PA 15213 USA. He is now with BBN Technologies, Boston, MA 02138 USA (e-mail: tblum@bbn.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.2001363

tention" to refer to where a person desires to look. We first infer the person's intention from different cues and then infer his or her FOA from that intention. The proposed method is based on the following hypothesis: In a remote collaboration task, a person's intention is directly related to task properties, partner actions, and message content. We investigate the problem in the context of remote collaborative tasks requiring tight coordination of conversation and action. Dialogue content in these tasks is tightly structured by task properties, specifically, the need to identify objects and place them in their correct locations. The manipulation of task objects also provides cues that can be used to model a person's intention. We employ a conditional Markov model to characterize helpers' FOA based on dialogue content and worker actions. We demonstrate the feasibility of the proposed method using a laboratory task in which a helper instructs a worker on how to assemble online puzzles. The accuracy of the model ranged from 65.40% for puzzles with easy to name pieces to 74.25% for puzzles with difficult to name pieces.

The feasibility of this approach was first demonstrated in our earlier report of people's gaze behaviors during collaborative physical tasks [36]. Here, we extend these findings by providing a broader theoretical framework, fine-grained technical details, and additional data analyses.

The remainder of the paper is organized as follows: In Section II we discuss related work; in Section III, we present our model; in Section IV, we present our experiment, the results, and a discussion of the findings; and in Section V we present our conclusions.

II. Related Work

In this research, we are interested in modeling people's FOA in remote collaborative tasks.

A. Collaborative Tasks

When collaborators work on a task together, they have some information about 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 (e.g., [45]). 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.

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 presents instructions; the worker performs the actual task actions. This 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. To collaborate successfully, helpers and workers must carefully coordinate their activities. A helper needs to know when it is appropriate to provide assistance. After giving 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].

B. 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-located, they share substantial visual space, which they can rely upon when planning what to say or do next. Many studies (e.g., [6], [12]–[14], [22], [34], [46]) have shown that co-located pairs, who have all sources of visual information available, complete tasks more rapidly and accurately than pairs working at a distance.

It is useful to distinguish among three types of visual information available in co-located settings: participants' heads and faces, participants' actions, and task objects [22]. Each of these visual cues has different benefits for collaboration depending upon the phase of the task. For example, when deciding advice is needed, a view of the current state of the task may be most critical, whereas when determining if instructions have been followed correctly, a view of a worker's actions may be more valuable.

C. Mediating Shared Visual Space in Remote Collaboration

Unlike people working side-by-side, who benefit from rich shared visual information, those working together at a distance must rely on technologies that limit the type and amount of visual information that is shared [49]. Task performance using such systems is typically only slightly if at all better than that using audio-only systems and rarely as good as that achieved in side-by-side collaboration (e.g., [12], [13], [34]).

One improvement suggested by Gaver and colleagues [15] is to provide multiple video feeds (e.g., of partners and of the workspace) and allow participants to switch among them. Such an approach is problematic due to high equipment requirements. In addition, the ability to switch between video feeds makes it difficult for users to understand what elements of the visual environment are or are not shared.

An alternative strategy is to identify the key types of visual information used in side-by-side settings and to implement systems to provide these critical visual cues to remote collaborators. Monk and Gale [31], for example, demonstrated that seeing where a partner was looking in a workspace was more valuable than seeing that partner's face alone. Fussell and colleagues [13] found that pairs were faster building a robot when the remote helper had access to a static but wide angle view of the workspace than when he/she had access to a head-mounted camera that tracked the worker's gaze. Moreover, providing helpers with both the wide angle and head mounted cameras led to worse performance than providing them with the wide angle camera alone. These studies highlight some of the difficulties of using video systems to deliver theoretically relevant visual cues.

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

Eye-tracking methods can provide a fine-grained understanding of people's use of visual information. Eye-tracking has been used to investigate relationships between gaze and actions in individual physical tasks (e.g., [21], [27], [39]) and in complex athletic behaviors (e.g., [38], [48]). It has also been used as a tool to understand interpersonal communication (e.g., [9], [20]), typically using 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 FOA in conversation. Vertegaal *et al.* [47] examined gaze at partners during a four-person conversation and found that gaze strongly indicated participants' FOA. Stiefelhagen *et al.* [44] also studied gaze in four-party conversations with a focus on how head and eye movements were associated as cues of attention. Gullberg [18] found that listeners do not always fixate speakers' gestures.

E. Modeling Focus of Attention

In this paper, focus of attention refers to a spatial location; a subject concentrates on some features of the environment to the (relative) exclusion of others. FOA is an important cue in many multimedia applications. For example, a dialogue system can provide better services if it can tell where the user is looking (e.g., [40], [42]). FOA data can also be used to identify who is being addressed in a multi-party dialogue [47]. If a remote video conferencing system can predict a speaker's FOA, it can provide that speaker with the most beneficial camera view of the remote site at any given moment [36].

Gaze coordinates output by eye trackers are a good indicator of FOA. However, it can be difficult to interpret eye movement patterns. Researchers have developed different methods for identifying FOA from raw gaze data [43]. For example, Jacob [19] investigated the use of eye movements as a computer input modality. Campbell *et al.* used a three-step algorithm to detect whether a user is reading or scanning text [4]. Their system first quantized the output data by averaging every three raw data points. The quantized stream was then tokenized into events and assigned a point indicating evidence supporting reading or scanning. Finally, evidence was obtained by summing the points of the pooled data. The system switched modes when the pooled evidence crossed a threshold. Salvucci used hidden Markov models (HMM) to interpret gaze data in an eye typing study [42]. A two-state (saccade/fixation) HMM was first applied to find a sequence of fixations. Then, a global HMM was constructed with submodels of target areas (e.g., a letter). The optimal path in the entire HMM was decoded with a Viterbi algorithm. In the current study, we obtained ground truth of FOA from gaze tracking data. Target areas were separated spatially in such a way that an unsupervised clustering algorithm could be applied effectively.

The relationship between gaze and actions has also been examined. Clark's theory of conversational grounding (e.g., [5], [7] [8]) suggests that helpers will look at targets that help them determine whether or not their instructions have been understood as intended. Prior work also suggests that people look at objects or devices with which they are interacting (e.g., [2], [3], [28], and [29]). Argyle and Cook indicated that gaze patterns of speakers and listeners are closely linked to the words spoken, and help in the timing and synchronization of utterances [1]. Griffin and Bock studied the time course of sentence formulation [17] and found a systematic temporal linkage between eye movements and spoken utterances.

Predicting FOA from modalities other than gaze from eye trackers is a relatively new area. Stiefelhagen *et al.* [44] used video from a panoramic camera and/or audio to predict FOA in a four-party meeting. They first estimated head poses using neural nets and then applied a Baysian model to compute the posterior probability of the FOA given head position. Accuracy was 74% using visual cues alone, 63% using audio alone, and 76% using a linear combination of video and audio. Otsuka *et al.* [32] examined FOA in a four-person conversation. They defined the current state of the conversation as a "regime" that can take values of "convergence," "dyad-link," and "divergence." The sequences of regimes and gaze patterns were modeled as hidden states and could be inferred by the utterances (who is speaking) and head directions using a Markov switching model.

The problem discussed in the current paper differs from previous work in two ways. First, our work is done in the context of remote collaboration on physical tasks involving objects that can themselves be foci of attention. Second, while speech versus silence might be sufficiently informative in multi-party conversation, in our task we need to understand message content in order to provide helpers with the most beneficial view at each point in time.

III. Modeling Focus of Attention Using Conditional Markov Models

In this paper, we propose to model FOA using conditional Markov models. Our problem is very similar to text annotation problems such as part of speech (POS) tagging and shallow parsing—given a sequence of uttered words, our goal is to output their labels. In text annotation problems, the labels are POS tags or syntactical phrases, whereas in our problem, the labels are the helpers' foci of attention ($g_t \in \mathcal{G}$).

Hidden Markov Modeling (HMM) is a powerful tool for processing sequential data that has been successfully applied to many natural language processing problems including speech recognition and information extraction (e.g., [11], [40]). HMM is a generative model that models the joint probability of the hidden states and the observation. However, to make the inference tractable, HMM assumes the conditional independence between the current observed feature set (e.g., the current word) and the other feature sets (other words) given the current state. This assumption is not true in many real world applications.

Maximum Entropy Markov Modeling (MEMM) [30] directly models the probability of the label sequence given the observation sequence. It does not need to assume conditional independencies among observations given a state. With MEMM, we can use features with richer representation of the observation sequence. For example, the label of the current state might depend

on the features extracted from the surrounding observations of the current position. The main problem of MEMM is the label bias problem, in which the probability mass of the current state will be all passed to low entropy next states regardless of the observation [26].

To overcome the problems of HMM and MEMM we employ a conditional Markov model. The conditional Markov model is an implementation of conditional random fields (CRFs), a probabilistic framework for labeling and segmenting structured data [26]. CRFs define a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. CRFs outperform MEMMs and HMMs on real-world tasks in many fields, including bioinformatics, computational linguistics and speech recognition. However, as an undirected graphical model, there is a global normalizer in the random field of CRF, which makes exact inference in the model intractable. Specifically, there is no analytical solution to the model parameters; instead, iterative techniques such as iterative scaling and gradient-based methods need to be used for the learning and inference of CRF, requiring much training data and computational expense.

Similar to the CRF, we factorize the dependencies between the current state and the next state, and between states and observations. Specifically, because the observations come from different modalities—workers' actions and the conversation—we use factors to represent the relationships between states and workers' actions, and between states and the conversation separately. To make the training fast and to decrease the number of parameters, we use probabilities of overlapping features as potential functions. When more data is available, we can train the model discriminately by using the maximum entropy framework and maximize the conditional likelihood. In the rest of this section we formally define the problem and discuss the model design and training/testing algorithms in detail.

A. Problem Description

We define the problem as a classification problem. Let $\{w_1, w_2, \ldots, w_N\}$ be a sequence of uttered words, with corresponding starting times (st_i), and ending times (e_i), $\{st_1, e_1, st_2, e_2, \ldots, st_N, e_N\}$. The worker's action m_t at time t is obtained by an activity sensor, $m_t \in \mathcal{M}$, where $\mathcal{M}$ is the set of all possible worker actions. For example, in the online puzzle task used by [35] and described in more detail in Section IV, a helper instructs a worker on how to construct a jigsaw puzzle out of colored pieces. Based on these instructions, the worker selects a piece from the pieces bay and moves it to a target location in the workspace. In this task, $\mathcal{M}$ will be {Not-Moving, Moving-in-Workspace, Moving-in-Pieces-Bay} and m_t can be annotated by analyzing the mouse click/drag events at time t. In a 3-D collaborative task such as the circuit assembly task used in [37], $\mathcal{M}$ can be {Idle, Searching-Part, Assembling-Part}, and m_t can be annotated by detecting the worker's hand position in the video sequence.

Let g_t be the helper's gaze coded at time t, $g_t \in \mathcal{G}$, where $\mathcal{G}$ is the set of possible gaze targets that we want to predict.

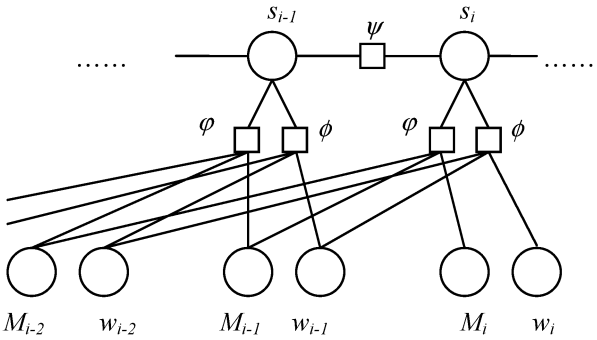


Fig. 1. A conditional Markov model. s_i, M_i, and w_i are state, action, and uttered word in time i; ψ, ϕ, and φ are factors.

In the jigsaw puzzle task, we want to know whether the helper is interested in looking at the workspace (where the worker is assembling the puzzle) or the pieces-bay (where the pieces are stored), and $\mathcal{G}$ is therefore {Workspace, Pieces-Bay}. At each time t we predict the helper's gaze as $\hat{g}_t$.

Because our objective is to control a camera automatically, the helper's FOA must be predicted in advance based on previous information from the dialogue and worker's actions. Using the predictive model, the camera shifts between different views (e.g., pieces bay versus workspace). We formulate online prediction as: At each sampling point t, given the previous words $(w_1, w_2, \ldots, w_i)$, and the previous worker's actions $(m_1, m_2, \ldots, m_t)$ as input, classify the next gaze code as $\hat{g}_{t+1} \in \mathcal{G}$.

B. A Conditional Markov Model Classifier

Under the assumption that a user's intention is directly related to task properties, workers' actions, and conversational content, we characterize this intention as the change of conversational topic. From the sequence of uttered words we can segment, label and code each clause and define $\mathcal{C}$ as the set of possible clause codes, which depend on the property of the task. For example, in the puzzle task, helpers need to describe the color pieces, describe where these pieces should be placed, and correct the workers if necessary. Thus $\mathcal{C}$ is {"Description of color piece," "Description of location," "Correcting color piece," "Correcting location"}.

When we take the stream of words and workers' actions as input, the clause boundaries (the end of a clause) and coding (the category of the clause) are not available. Therefore we predict the clause coding of each word. Pairing gaze and clause coding, we form a sequence of $|G| \times |\mathcal{C}|$ possible states. That is, a state s_i is the pair (g_i, c_i), where $g_i \in \mathcal{G}$ and $c_i \in \mathcal{C}$. Let $\vec{S}, \vec{W}$, and $\vec{M}$ be the state sequence, word sequence and action sequence, respectively. We align the word and action sequences if they are sampled at different rates. (An example of the alignment is discussed in Section IV.) The undirected graphical model with its factors is shown in Fig. 1. ψ captures the relationship between the current state and the previous state, while ϕ and φ characterize features $\vec{W}$ and $\vec{M}$ using a history window of n words. We choose $n = 3$ in the following discussion.

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:

$$\Pr\{\vec{S}|\vec{W},\vec{M}\} = \frac{1}{Z}\prod_i (\psi(s_{i-1},s_i) * \phi(w_i,w_{i-1},w_{i-2},s_i) * \varphi(M_i,M_{i-1},M_{i-2},s_i)) \quad (1)$$

where Z is the partition function:

$$Z = \sum_S \prod_i (\psi(s_{i-1},s_i) * \phi(w_i,w_{i-1},w_{i-2},s_i) * \varphi(M_i,M_{i-1},M_{i-2},s_i))$$

and we define

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

To decrease the number of parameters, we can cluster words into a small number of categories and classify each word as one of those categories. We then approximate ϕ by

$$\begin{aligned} \phi(w_i,w_{i-1},w_{i-2},s_i) &= \Pr\{w_i,w_{i-1},w_{i-2}|s_i\} \\ &\approx \Pr\{c_i,c_{i-1},c_{i-2}|s_i\} \end{aligned} \quad (3)$$

where c_i is the category of w_i.When training is supervised, we can employ a maximum likelihood method to learn the parameters and the Good-Tuning method [16] to estimate the parameters for the n-grams that have counts lower than T in the training data:

$$\Pr(s_i|s_{i-1}) = \frac{c(s_{i-1},s_i)}{c(s_{i-1})}, \quad (4)$$

$$\Pr(c_i,c_{i-1},c_{i-2}|s_i) = \begin{cases} \frac{n_c(j+1)}{n_c(j)} * \frac{j+1}{c(s_i)}, & if\ c(c_i,c_{i-1},c_{i-2},s_i) = j \le T, \\ \alpha_c * \frac{c(c_i,c_{i-1},c_{i-2},s_i)}{c(s_i)} & \end{cases} \quad (5)$$

$$\Pr(M_i,M_{i-1},M_{i-2}|s_i) = \begin{cases} \frac{n_M(j+1)}{n_M(j)} * \frac{j+1}{c(s_i)}, & if\ c(M_i,M_{i-1},M_{i-2},s_i) = j \le T, \\ \alpha_M * \frac{c(M_i,M_{i-1},M_{i-2},s_i)}{c(s_i)} & \end{cases} \quad (6)$$

where $c(*)$ denotes the number of times the combination * appears in the training corpus, $n_c(j)$ and $n_M(j)$ are the numbers of n-grams that have count j, and α_c and α_M are constants that guarantee (5) and (6) are appropriate distributions:

$$\alpha_c = \frac{\sum_{j>T+1} j * n_c(j)}{\sum_{j>T} j * n_c(j)}, \quad (7)$$

Training:

Input: uttered word sequence $\vec{W}$, worker's action coding sequence $\vec{M}$, gaze coding sequence $\vec{G}$, clause coding sequence $\vec{L}$, word-category mapping table.

Align $\vec{M}$, $\vec{G}$, $\vec{L}$ to $\vec{W}$ (Section IV).

Map each word into its category.

Generate state sequence $\vec{S}$ by pairing $\vec{G}$ and $\vec{L}$.

Count $c(s_i)$, $c(s_{i-1},s_i)$, $c(c_i,c_{i-1},c_{i-2},s_i)$, $c(M_i,M_{i-1},M_{i-2},s_i)$.

Count $n_c(j)$ and $n_M(j)$, j=0, 1, 2, …

Calculate α_c and α_M by using Eq. (7) and Eq. (8).

Calculate $\psi(s_{i-1},s_i)$, $\phi(w_i,w_{i-1},w_{i-2},s_i)$, and $\varphi(M_i,M_{i-1},M_{i-2},s_i)$ by using Eq. (4), Eq. (5), and Eq. (6).

Output: $\psi(s_{i-1},s_i)$, $\phi(w_i,w_{i-1},w_{i-2},s_i)$, and $\varphi(M_i,M_{i-1},M_{i-2},s_i)$ for all combinations.

Testing:

Input: uttered word sequence $\vec{W}$, worker's action coding sequence $\vec{M}$; model parameters output by the training process.

Calculate the optimal state sequence recursively by using Eq. (10)

Map each state to gaze coding and generate gaze coding sequence $\hat{G}$.

Output: gaze coding sequence $\hat{G}$.

Fig. 2. Training and testing of the conditional Markov model.

$$\alpha_M = \frac{\sum_{j>T+1} j * n_M(j)}{\sum_{j>T} j * n_M(j)}. \quad (8)$$

When training is unsupervised, the EM algorithm can be applied.

Since Z in (1) does not depend on $\vec{S}$, we can decompose the factors and use a Viterbi algorithm to find the optimal path given the parameters and input sequences $\vec{W}$ and $\vec{M}$. At each word i we define the Viterbi probability of each state q as

$$\delta_i(q) = \max_{s_1,s_2,\ldots,s_{i-1}} \Pr(s_1,s_2,\ldots,s_{i-1},s_i=q| w_1,\ldots,w_i,M_1,\ldots,M_i). \quad (9)$$

And $\delta_i(q)$ can be calculated recursively:

$$\begin{aligned} \delta_3(q) &= \phi(w_1,w_2,w_3,q) * \varphi(M_1,M_2,M_3,q), \\ \delta_i(q) &= \max_{q'} (\delta_{i-1}(q') * \psi(q',q)) * \phi(w_i,w_{i-1},w_{i-2},q) \\ &\quad * \varphi(M_i,M_{i-1},M_{i-2},q). \end{aligned} \quad (10)$$

We choose the state that has the highest Viterbi probability as the predicted state:

$$\hat{q}_i = \arg\max_q \delta_i(q) \quad (11)$$

The decoded state sequence can be mapped to the corresponding gaze sequence and clause coding sequence.

Fig. 2 summarizes the training and testing of the conditional Markov model.

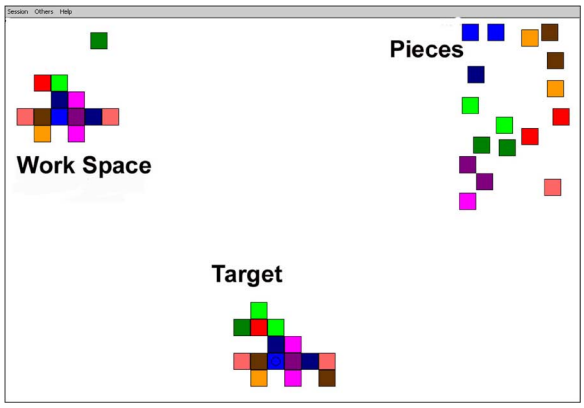


Fig. 3. Helper's display in the online puzzle task. The solution to the puzzle (target) is in the center bottom. The pieces bay, from which pieces can be selected, is on the upper right. The workspace, where the worker is putting the puzzle together, is on the upper left.

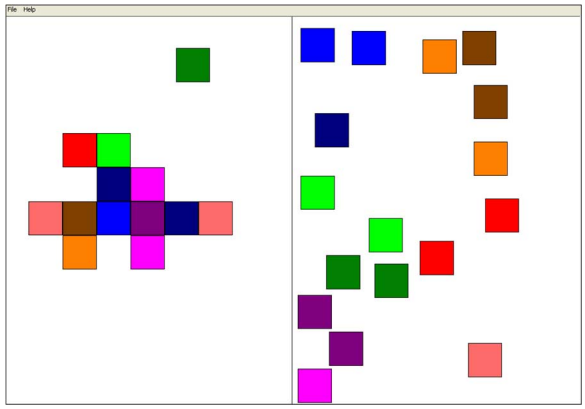


Fig. 4. Worker's display, with the pieces bay on the left and the workspace on the right. Workers' actions in these areas were transmitted to the Helper's display (shown in Fig. 3).

IV. Experiments

In order to verify the proposed model, we designed a remote collaborative task and performed several experiments.

A. Design

Our experiment used an online jigsaw puzzle task adapted from Kraut and colleagues [23], in which a helper and worker collaborated to construct a series of puzzles. The task is analogous to remote physical collaboration in that a helper offers instructions remotely and a worker manipulates the objects (puzzle pieces) online.

There were three areas among which the helper could gaze freely to obtain visual information (Fig. 3).

- 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 versus shaded colors) and the complexity of the puzzle (five, 10 or 15 pieces). Each participant completed three puzzles for each condition (piece differentiability $\times$ puzzle complexity), randomly presented in a single block. The design formed a $2\,(\text{piece differentiability}) \times 3\,(\text{puzzle complexity}) \times 9\,(\text{trial})$ factorial within-subjects study. The order of puzzle blocks was counterbalanced across participants.

The helper's display (Fig. 3) was designed such that the three areas (workspace, pieces bay, and target solution) were in a triangular shape. The helper could shift his/her eye gaze from any area directly to either of the other two areas with equal effort. The worker's screen (Fig. 4) was laid out so that the workspace and pieces bay were adjacent to each other.

We created 18 target puzzles by randomly forming configurations of five, 10 or 15 pieces (see Fig. 5). There were six different puzzles for each level of complexity, three formed from a pool of solid color pieces (easily distinguishable colors, making pieces easier to describe), and three formed from a pool of shaded colors pieces (five similar shades of the same color, making pieces harder to describe).

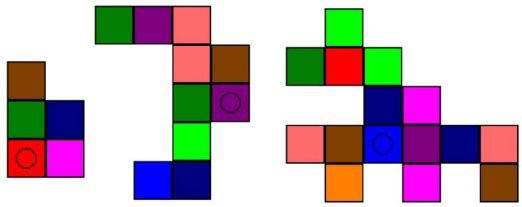


Fig. 5. Examples of puzzle configurations with five, 10, and 15 pieces.

Fig. 6. Helper's system. 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.

B. Data Collection

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.

Our eye tracking system (Fig. 6) 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. The system 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 helper's eye gaze coordinates and the worker's actions, was recorded using a Panasonic DV-VCR.

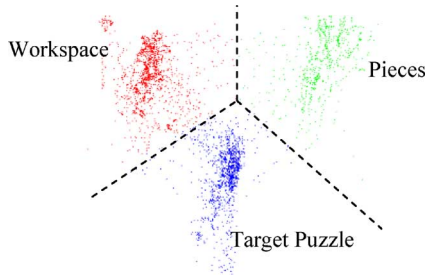


Fig. 7. Example of eye gaze distribution from one section of the tasks. After running K-Means VQ algorithm we obtained three clusters and classified each point in terms of the helper's FOA.

Ground truth of the helper's FOA (workspace, pieces bay, or target solution) at each point in time was derived from eye gaze coordinates. To interpret the raw gaze coordinates, we had to address issues related to unconscious eye movements such as jitter and the unreliable metric posed by the zero error of the magnetic sensor and the pupil/corneal reflection tracker ([4], [19], [42]). In addition, when a subject gazes at one area and looks back later, the fixation point can be different by some offset (drift). Fortunately, we do not need to compute the fixation over each small time segment for this application. Instead, taking advantage of the fact that the three areas are spatially laid out triangularly, we can take all data points from one session and cluster them using K-Means vector quantization (VQ). We first chose three initial centers in the same triangular fashion as the three areas on the helper's display. Within ten iterations, the algorithm converged and the outputs were three new centers. Subsequently, the helper's gaze coordinates were indexed based on their proximity to these three new centers (Fig. 7).

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 at computer terminals in the same room with a barrier between them so that they could hear but not see one another, simulating remote collaboration.

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 in the workspace to complete the target puzzle. The worker was allowed to converse with the helper and ask questions as necessary. The helper and worker were not restricted to a controlled vocabulary and could talk freely. 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 five-minute break after half of the puzzles were completed. After the break, the experimenter recalibrated the eye-tracker. Sessions lasted 60 to 90 minutes. In each session we recorded the dialogue, the worker's mouse activity, and the helper's gaze. An example of how the helper's dialogue, the helper's gaze, and the worker's actions are synchronized is shown in Fig. 8.

C. Clause Coding

Due to the nature of the instructional task, helpers' utterances dominated in the conversation (an average of 96.1% of the words spoken). We used a subset of the transcribed conversations to build a set of grammar. Two coders separated transcribed utterances (including filler words) into clauses and coded each clause according to the categories shown in Table I. Reliability was established by having both coders code a sample of 10% of the data. In this sample, the two coders agreed with each other 95% of the time.

D. Offline Prediction

Notice that in online prediction, the system does not have supervised knowledge of clause boundaries and clause coding. However, for comparison purposes, we want to know how good the system can be if this syntactic and semantic information is available. It is similar to asking a human classifier to review the task process offline and label the gaze of the sentence after the whole sentence is spoken. We define this as *offline prediction*

In offline prediction we assume that clause coding (Table II) at time t is known. Given the clause coding and the worker's action, the helper's gaze at that time can be predicted by using maximum likelihood estimation:

$$\hat{g}_i = \underset{j=0 \text{ or } 1}{\arg\max} \Pr\{j | clause_t, m_t\} \tag{12}$$

while $clause_t$ is the clause coding at time t. In both training and testing phases the clause coding and worker's actions are known. We estimated the conditional probabilities of each gaze target $Pr\{j|clause_t, m_t\}$ from training sample frequencies.

E. Winner-Takes-All Strategy

In online prediction we use a conditional Markov model to infer the clause boundaries and clause coding from the word sequence and action sequence. To circumvent the problem introduced by the high sampling rate (60 HZ), we employ a smoothing technique called the "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 do not expect to predict gaze well at every sampling point. Therefore we apply a winner-takes-all strategy and smooth the helper's gaze and worker's action data based on the boundaries of words. That is, the decision is only

$$M_i = \begin{cases} -1, & \text{if no movements between time } e_{i-1} \text{ and } e_i \\ \underset{j=0 \text{ or } 1}{\arg\max} \left|\{m_t | m_t = j, e_{i-1} < t \le e_i\}\right|, & \text{otherwise,} \end{cases} \tag{13}$$

$$G_i = \underset{j=0 \text{ or } 1}{\arg\max} \left|\{g_t | g_t = j, e_{i-1} < t \le e_i\}\right|. \tag{14}$$

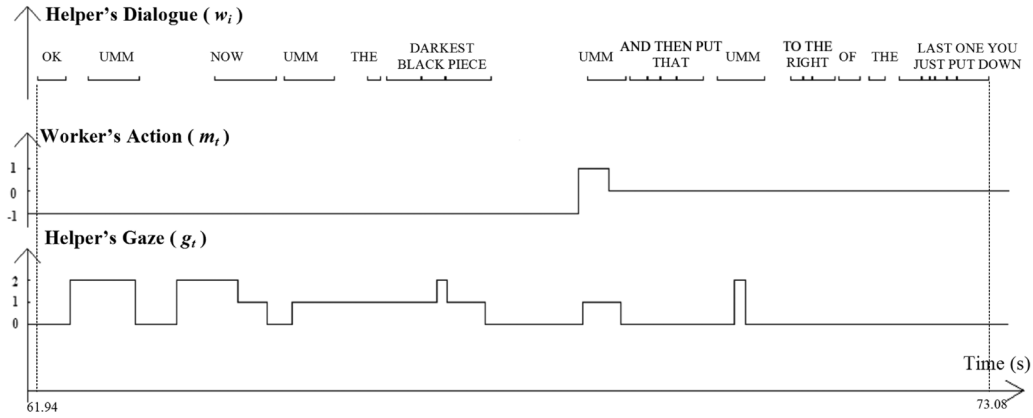


Fig. 8. Demonstration of the three sources of data in a 12-second period. The helper's was giving the instruction "OK UMM NOW UMM THE DARKEST BLACK PIECE UMM AND THEN PUT THAT UMM TO THE RIGHT OF THE LAST 'ONE YOU JUST PUT DOWN". Starting time and ending time of each word are aligned with the worker's 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).

TABLE I
CODING OF HELPER CLAUSES

Message Content	Instructional Content
Description of color piece	"Take the green block"
Description of location	"And then put that to the right of the dark gray"
Correcting color piece	"A little lighter than that"
Correcting location	"It's on the very right"

TABLE II
CONFUSION MATRIX FOR SOLID PUZZLES

Observed \ Actual	0 (Workspace)	1 (Pieces Bay)
0 (Workspace)	75.17%	24.83%
1 (Pieces Bay)	47.24%	52.76%

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 shown in the equation at bottom of previous page. 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 Fig. 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).

F. Experimental Results

In our evaluation, we only consider gaze to the workspace and pieces bay (which are parts of the remote work area that might be transmitted via a video system), ignoring gaze towards the target (which is part of the helper's local environment and thus would not need to be transmitted). Let g_t, $t = 1 \ldots T$, be the actual gaze codes collected from the experiment and processed by the VQ algorithm. The classification error at time t is defined as

$$err(\hat{g}_t, g_t) = \begin{cases} 1, & \text{if } g_t = 0 \text{ or } 1, \text{ and } \hat{g}_t \neq g_t \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

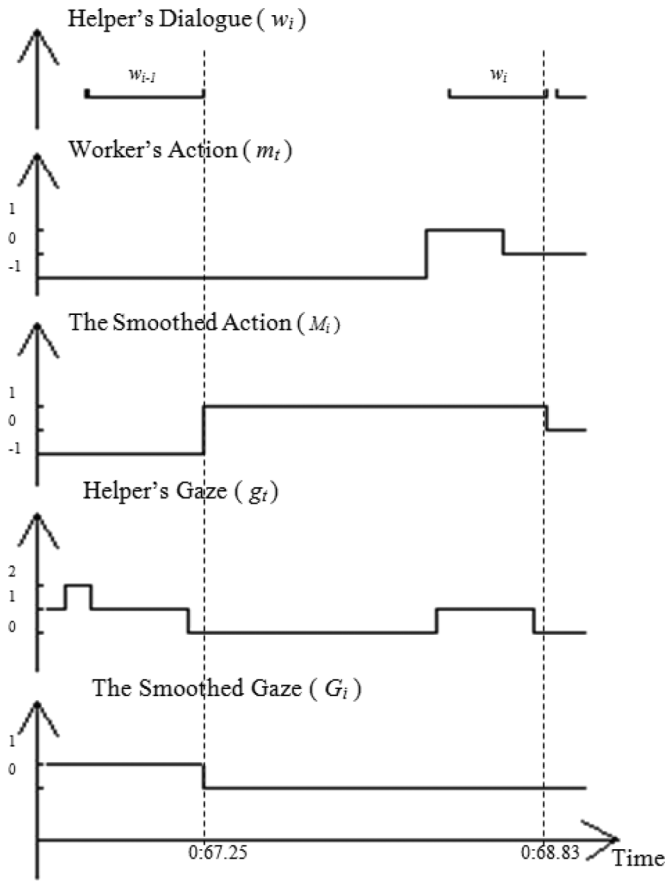


Fig. 9. Smoothed action and gaze data based on the Winner-Takes-All strategy [(13) and (14)].

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

$$Acc = 1 - \frac{\sum_{t=1}^{T} err(\hat{g}_t, g_t)}{|\{g_i | g_i = 0 \text{ or } 1\}|} \quad (16)$$

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

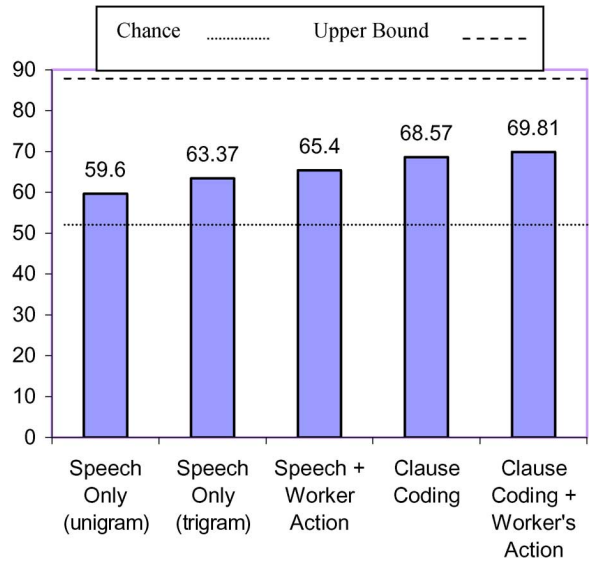


Fig. 10. Classification accuracies for solid-color tasks.

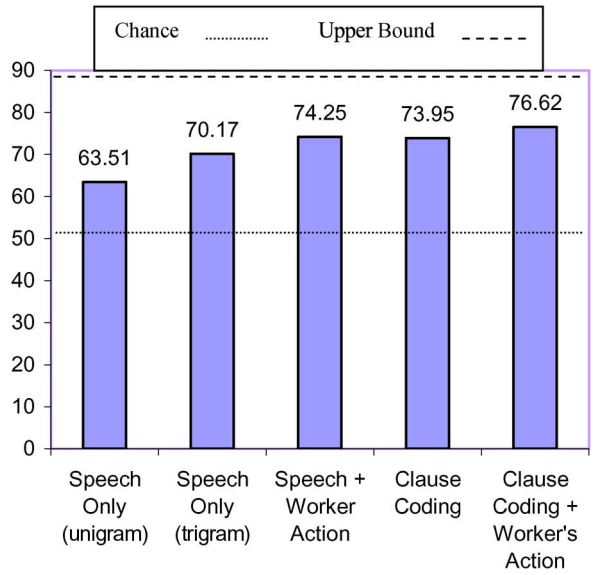


Fig. 11. Classification accuracies for shaded-color tasks.

Helpers behave differently when giving instructions for puzzles with solid versus shaded pieces. Solid color pieces are easier and faster to name than shaded pieces. Consequently, we trained and tested classifiers for solid and shaded puzzles separately. Data were collected with 12 users and a total of 216 puzzles (half solid color, half shaded). In each condition, we evaluated accuracy [(16)] by using a two-fold cross validation. We compared the performances among classifiers with different features: unimodal (speech only), multimodal (speech and action), and offline classification. Moreover, because we expect that the conditional Markov classifiers can classify clause coding, we evaluated the accuracies of this coding by comparing it with the human labels.

The accuracies of the different classifiers for solid and shaded puzzles are shown in Figs. 10 and 11. We should point out that because we only output gaze prediction at word boundaries, while gaze does change *within* word boundaries (as shown in Fig. 8), we already lose some accuracy after applying (13) and (14). The upper bounds for solid puzzles and shaded puzzles after smoothing are 87.40% and 88.46%, respectively.

TABLE III
CONFUSION MATRIX FOR SHADED PUZZLES

Observed \ Actual	0 (Workspace)	1 (Pieces Bay)
0 (Workspace)	57.56%	42.44%
1 (Pieces Bay)	15.42%	85.58%

Unigram versus Trigram: We first looked at the effect of including a history of three words. Unigram and Trigram are commonly used language models. In our unimodal analysis, including the history of words (Speech-Only-Trigram) outperforms using the current word only (Speech-Only-Unigram)

Unimodal versus Multimodal: Moreover, multimodal classifiers achieve better performance than unimodal classifiers. 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. The improvements are more obvious in shaded than in solid puzzles because the worker's action data are more discriminative in predicting the helper's FOA in shaded-color tasks. In contrast, in solid-color tasks, when the worker was acting in the pieces-bay, the chances that the helper was looking at the workspace and pieces-bay are almost equal. 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. In summary, Speech+Action is better than Speech-Only; Clause-Coding+Action is better than Clause-Coding (Figs. 10 and 11). These conclusions are also justified by statistical analysis [35].

Online versus Offline: The third observation is that if we know the sentence coding, an MLE classifier can do better than our proposed conditional Markov classifier. This indicates that knowing the user's intentionality is essential for predicting FOA. As described in [35], when describing a piece, helpers overwhelmingly look at the pieces bay, whereas when describing a location, they are much more likely to look at the workspace. When the worker asks the helper to clarify a color piece/location, the helper looks more in the piece bay/workspace. However, sentence codings require high level semantic knowledge and are not available in real time; they have to be inferred by other information (e.g., the ongoing dialogue.)

The confusion matrixes of Speech+Action classifiers for solid and shaded puzzles are shown in Tables II and III. It can be seen that the accuracies for individual classes (Workspace and Pieces-bay) are shifted by their prior probability. For solid puzzles, the algorithm classifies the Workspace better; whereas for shaded puzzles, it classifies the Pieces-bay better.

The conditional Markov model is predicting gaze coding and sentence coding simultaneously. To verify that, we compared the output state sequences in the test set with their true sentence codings. The accuracy for predicting sentence codings was 59% for solid puzzles and 48.37% for shaded puzzles. Instructional codings of solid puzzles were classified better because the language structure is simpler than that for shaded puzzles. The

TABLE IV
CONFUSION MATRIX OF SENTENCE CODING FOR SOLID-COLOR PUZZLES. (UNIT: NUMBER OF SENTENCES (%))

Actual \ Observed	Description of color piece	Description of location	Correcting color piece	Correcting locaction
Description of color piece	1363 (60.5%)	685 (30.4%)	91 (4.0%)	115 (5.1%)
Description of location	681 (23.9%)	1740 (61.1%)	81 (2.8%)	347 (12.2%)
Correcting color piece	46 (27.5%)	38 (22.8%)	22 (13.2%)	61 (36.5%)
Correcting locaction	32 (10.3%)	59 (19.0%)	23 (7.4%)	196 (63.2%)

confusion matrix of classifying sentence codings for solid-color tasks is shown in Table IV.

Comparison With Gaze Based Systems: Gaze (head pose) is a good indicator of FOA. Gaze based systems have been widely used in estimating FOA for meeting analysis. Stielfhagen *et al.* reported 54% accuracy in predicting the participants' focus of attention on three recorded meetings [44]. By using both head pose and sound, they achieved 76% accuracy in detecting the participants' focus of attention on the recorded meetings. More recently, Jovanovic *et al.* achieved accuracy of 66.45% for predicating FOA using solely gaze information in four-participant face-to-face meetings [20]. They have shown that combing utterance, contextual and speaker's gaze features, FOA can be predicted with an accuracy of 82.57%.

The goal of this research is to predict FOA without using gaze information. Although a gaze based system can achieve higher accuracy, it poses extra cost and technical constraints to the environment (e.g., lighting conditions) and the subject (e.g., freedom of movement). In addition, for our targeted application, remote collaboration, a human can correct system mistakes if needed, because humans are in-the-loop. Therefore, the proposed method provides a different solution to address the need of estimating FOA, though user studies are needed to address usability issues for different applications

V. CONCLUSION AND FUTURE WORK

With the goal of building an innovative multimedia system to support remote collaborative tasks, we have proposed to use conditional Markov modeling to predicate FOA from intention. Instead of tracking FOA via head position, we formulated the problem as a multimodal classification problem. We employed a conditional Markov model to classify FOA from the intention decoded from dialogue content and workers' actions.

We tested our method using data from an online puzzle task in which a remote helper's gaze shifted among the pieces bay, workspace, and target solution. The experimental results show that multimodal classifiers outperform unimodal classifiers. The overall accuracy was 65.40% for solid color puzzles and 74.25% for shaded puzzles.

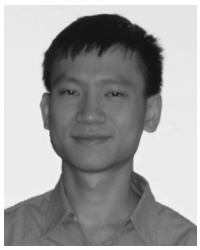
These results indicate the feasibility of predicting where a helper wants to look in real time from his/her intention during remote collaboration. The results can be used to allocate resources optimally and thus enhance remote collaboration. The results are consistent with a conversational grounding view of collaboration (cf. [5]). 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. The results are also consistent with our analysis of the relationships between the helper's FOA and three factors: task properties, workers' actions and message content [35].

Our future research will focus on two issues that are important for making our findings useful in practice. First, our experimental results suggest the importance of understanding workers' actions during the collaboration. While this information is easy to obtain in an online computer task, it is much more difficult when we move the setting to 3-D tasks. Event analysis on videos will be studied and combined with audio signals. Second, our accuracy rates were higher than random guessing but still far from perfect. We plan to conduct behavioral research to test the usability of the automatic view switching function. We will look at the impact of errors in prediction and how these side effects can be minimized.

REFERENCES

[1] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge, U.K.: Cambridge Univ. Press, 1976.
[2] B. Brumitt, J. Krumm, B. Meyers, and S. Shafer, "Let there be light: Comparing interfaces for homes of the future," *IEEE Pers. Commun.*, Aug. 2000, (2000).
[3] E. Campana, J. Baldridge, J. Dowding, B. A. Hockey, R. W. Remington, and L. S. Stone, "Using eye movements to determine referents in a spoken dialogue system," in *Proc. 2001 Workshop on Perceptive User Interfaces*, 2001, pp. 1–5.
[4] C. S. Campbell and P. P. Maglio, "A robust algorithm for reading detection," in *Proc. ACM Conference on Perceptive User Interfaces (PUI '01)*, 2001.
[5] H. H. Clark, *Using Language*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[6] H. H. Clark and M. A. Krych, "Speaking while monitoring addresses for understanding," *J. Memory Lang.*, vol. 50, pp. 62–81, 2004.
[7] H. H. Clark and C. E. Marshall, , A. K. Joshi, B. L. Webber, and I. A. Sag, Eds., "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*. Cambridge, U.K.: Cambridge Univ. Press, 1981, pp. 10–63.
[8] H. H. Clark and S. E. Brennan, *Grounding in Communication Perspectives on Socially Shared Cognition*, L. B. Resnick, R. M. Levine, and S. D. Teasley, Eds. Washington, DC: APA, 1991, pp. 127-149–.
[9] K. M. Eberhard, M. J. Spivey-Knowlton, J. C. Sedivy, and M. K. Tanenhaus, "Eye movements as a window into real-time spoken language processing in natural contexts," *J. Psycholinguistic Res.*, vol. 24, pp. 409–436, 1995.
[10] M. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, pp. 32–64, 1995.
[11] D. Freitag and A. K. McCallum, "Information extraction using hmms and shrinkage," in *Papers from the AAAI-99 Workshop on Machine Learning for Information Extraction*, Menlo Park, CA, 1999, pp. 31–36.
[12] S. R. Fussell, R. E. Kraut, and J. Siegel, "Coordination of communication: Effects of shared visual context on collaborative work," in *Proc. CSCW 2000*, New York, 2000, pp. 21–30.
[13] S. R. Fussell, L. D. Setlock, and R. E. Kraut, "Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks," in *Proc. CHI 2003*, New York, 2003, pp. 513–520.

[14] S. R. Fussell, L. D. Setlock, J. Yang, J. Ou, E. M. Mauer, and A. Kramer, "Gestures over video streams to support remote collaboration on physical tasks," *Human-Computer Interact.*, vol. 19, pp. 273–309, 2004.
[15] W. Gaver, A. Sellen, C. Heath, and P. Luff, "One is not enough: Multiple views in a media space," in *Proc. Interchi '93*, New York, 1993, pp. 335–341.
[16] I.J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 1953, pp. 237–264, 1953.
[17] Z. M. Griffin and K. Bock, "What the eyes say about speaking," *Psychol. Sci.*, vol. 11, pp. 274–279, 2000.
[18] M. Gullberg, "Eye movements and gestures in human face-to-face interaction," in *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements*, J. Hyona, R. Radach, and H. Deubel, Eds. Oxford, U.K.: Elsevier, 2003, pp. 685–703.
[19] R. J. K. Jacob, "Eye-movement-based human-computer interaction techniques: Toward non-command interfaces," in *Advances in Human-Computer Interaction*, H. R. Hartson and D. Hix, Eds. Norwood, NJ: Ablex, 1993, vol. 4, pp. 151–190.
[20] N. Jovanovic, R. op den Akker, and A. Nijholt, "Addressee identification in face-to-face meetings," in *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, 2006.
[21] B. Keysar, D. J. Barr, J. A. Balin, and J. S. Brauner, "Taking perspective in conversation: The role of mutual knowledge in comprehension," *Psychol. Science*, vol. 11, pp. 32–38, 2000.
[22] R. E. Kraut, S. R. Fussell, and J. Siegel, "Visual information as a conversational resource in collaborative physical tasks," *Human–Computer Interact.*, vol. 18, pp. 13–49, 2003.
[23] R. E. Kraut, D. Gergle, and S. R. Fussell, "The use of visual information in shared visual spaces: Informing the development of virtual co-presence," in *Proc. CSCW 2002*, New York, 2002, pp. 31–40.
[24] H. Kuzuoka, T. Kosuge, and K. Tanaka, "GestureCam: A video communication system for sympathetic remote collaboration," in *Proc. CSCW 1994*, New York, 1994, pp. 35–43, ACM.
[25] H. Kuzuoka, S. Oyama, K. Yamazaki, K. Suzuki, and M. Mitsuishi, "GestureMan: A mobile robot that embodies a remote instructor's actions," in *Proc. CSCW 2000*, New York, 2000, pp. 155–162, ACM Press.
[26] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Eighteenth Int. Conf. Machine Learning (ICML-2001)*, 2001, 2001.
[27] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, pp. 1307–1432, 1999.
[28] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[29] P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith, "Gaze and speech in attentive user interfaces," in *Proc. Int. Conf. Multimodal Interfaces*, 2000, vol. 1948, LNCS. Springer, 2000.
[30] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proc. ICML 2000*, Stanford, California, 2000, pp. 591–598.
[31] A. F. Monk and C. Gale, "A look is worth a thousand words: Full gaze awareness in video-mediated conversation," *Discourse Processes*, vol. 33, pp. 257–278, 2002.
[32] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances," in *Proc. Int. Conf. Multimodal Interfaces*, Trento, Italy, Oct. 4–6, 2005.
[33] J. Ou, (unpublished), "DOVE-2: Combining gesture with remote camera control," .
[34] J. Ou, S. R. Fussell, X. Chen, L. D. Setlock, and J. Yang, "Gestural communication over video stream: Supporting multimodal interaction for remote collaborative physical tasks," in *Proc. Int. Conf. Multimodal Interfaces*, Vancouver, BC, Canada, Nov. 5–7, 2003, 2003.
[35] J. Ou, L. M. Oh, J. Yang, and S. R. Fussell, "Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2005, pp. 231–240, ACM Press.
[36] J. Ou, L. M. Oh, S. R. Fussell, T. Blum, and J. Yang, "Analyzing and predicting focus of attention in remote collaborative tasks," in *Proc. ICMI'05*, Trento, Italy, 2005, 2005.
[37] J. Ou, Y. Shi, J. Wong, S. R. Fussell, and J. Yang, "Combining audio and video to predict helpers' focus of attention in multiparty remote collaboration on physical tasks," in *ICMI 2006*, 2006.
[38] R. R. D. Oudejans, C. F. Michaels, F. C. Bakker, and K. Davids, "Shedding some light on catching in the dark: Perceptual mechanisms for catching fly balls," *J. Exper. Psychol.: Human Percept. Perf.*, vol. 25, pp. 531–542, 1999.
[39] J. B. Pelz and R. Canosa, "Oculomotor behavior and perceptual strategies in complex tasks," *Vision Res.*, vol. 41, pp. 3587–3596, 2001.
[40] P. Qvarfordt and S. Zhai, "Conversing with the user based on eye-gaze patterns," in *Proc. of the CHI 2005 Conf. Human Factors in Computing Systems*, New York, 2005, pp. 221–230, ACM Press.
[41] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[42] D. Salvucci, "Inferring intent in eye-based interfaces: Tracing eye movements with process models," in *Human Factors in Computing Systems: CHI 99*, 1999, 1999.
[43] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Eye Tracking Research and Applications Symp.*, 2000, pp. 71–78.
[44] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. Neural Netw.*, vol. 13, pp. 928–938, 2002.
[45] J. C. Tang, "Findings from observational studies of collaborative work," *Int. J. Man-Mach. Stud.*, vol. 34, pp. 143–160, 1991.
[46] B. M. Velichkovsky, "Communicating attention: Gaze position transfer in cooperative problem solving," *Pragmatics Cogn.*, vol. 3, pp. 199–222, 1995.
[47] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes," in *Proc. CHI 2001*, New York, 2001, pp. 301–308, ACM Press.
[48] J. N. Vickers, "Visual control when aiming at a far target," *J. Exper. Psychol.: Human Percept. Perf.*, vol. 22, pp. 342–354, 1996.
[49] S. Whittaker and B. O'Conaill, "The role of vision in face-to-face and mediated communication," in *Video-Mediated Communication*, K. Finn, A. Sellen, and S. Wilbur, Eds. Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 23–49.



Jiazhi Ou received the B.S. degree in computer science from Sun Yat-sen University (a/k/a Zhongshan University), China, and the M.S. degree in computer science from Fudan University, China. He is pursuing the Ph.D. degree at the Language Technologies Institute in School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interests include multimodal integration, human computer interaction, and machine learning.

Lui Min Oh received the B.Sc.degree in electrical and computer engineering (additional major in human–computer interaction and a minor in psychology) in 2004 and the a M.Sc.degree in human–computer interaction in 2005, both from Carnegie Mellon University, Pittsburgh, PA.

He is currently based in Singapore, working on risk assessment and horizon scanning technologies (www.rahs.org.sg) in the Defence Science & Technology Agency. His research interests include sensemaking, social computing, information visualization and interaction design.

Susan R. Fussell received the Ph.D. degree in social and cognitive psychology in 1991 from Columbia University, New York, NY.

She is currently an Associate Research Professor in the Human Computer Interaction Institute at Carnegie Mellon University, Pittsburgh, PA. One line of her research focuses on understanding the role of different modalities in interpersonal communication, with the goal of informing the design of new systems that better support these communication processes in remote collaboration. Other projects investigate the role of culture in interpersonal communication, the benefits of visualization tools for remote collaboration, and the use of the Internet to foster environmentally conscious behavior.

Tal Blum holds a Masters degrees in computer science from the Hebrew University of Jerusalem, Israel and a Masters from the Language Technologies Institute at Carnegie Mellon University, Pittsburgh, PA.

He is currently doing research in information extraction at BBN Technologies, Boston, MA.

Jie Yang received the Ph.D. degree in electrical engineering from the University of Akron, Akron, OH, in 1994.

He is currently a Senior Systems Scientist with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. He has been leading research efforts to develop visual tracking, recognition, and analysis systems for multimodal human–computer interaction in both intelligent working spaces and mobile platforms, such as real-time face tracking systems, gaze-based interfaces, the lipreading system, the image-based multimodal translation agent, multimodal people ID, and automatic sign translation systems. His current research interests include multimodal interfaces, computer vision, and pattern recognition.