Action as Language in a Shared Visual Space

Darren Gergle
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
dgergle+@cs.cmu.edu

Robert E. Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
robert.kraut@cmu.edu

Susan R. Fussell
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
susan.fussell@cmu.edu

ABSTRACT
A shared visual workspace allows multiple people to see similar views of objects and environments. Prior empirical literature demonstrates that visual information helps collaborators understand the current state of their task and enables them to communicate and ground their conversations efficiently. We present an empirical study that demonstrates how action replaces explicit verbal instruction in a shared visual workspace. Pairs performed a referential communication task with and without a shared visual space. A detailed sequential analysis of the communicative content reveals that pairs with a shared workspace were less likely to explicitly verify their actions with speech. Rather, they relied on visual information to provide the necessary communicative and coordinative cues.

Categories and Subject Descriptors
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – collaborative computing, computer-supported cooperative work.

General Terms
Design, Experimentation, Human Factors, Performance, Theory.

Keywords
Shared visual space, empirical studies, sequential analysis, language, and communication.

1. INTRODUCTION
A good portion of technology development for CSCW tacitly assumes that the primary goal is to support spoken language. For a large number of tasks, however, successful interaction does not rely solely on spoken language. Rather, communicative information can be provided in the form of linguistic utterances, visual feedback, gestures, acoustic signals, or a host of other sources; all of which play an important role in successful communication. Everyday communication requires conversants to integrate these elements in an extremely rapid, flexible, real-time and cooperative fashion. Speakers generate and monitor their own activities; however, they also monitor the *language* and *actions* of their partners and take *both* into account as they speak.

Consider a group of architects, consultants and lay clients working together to discuss architectural plans for the design of a new corporate headquarters. Communication in the group is not merely composed of a series of individual utterances produced one at a time and presented for others to hear. Rather, speakers and addressees take into account what one another can see [36], they notice where one another's attention is focused [1,6], point to objects in the space and say things like "that one" and "there" [4], make hand gestures, eye contact, facial expressions, and share knowledge about previously spoken discourse and behavioral actions [9]. Many observational studies have demonstrated this rich interplay between speech and action that takes place in collaborative interactions [5, 23, 37].

Previous research has demonstrated the value of shared views of a workspace for collaboration on physical tasks [20, 21, 25, 29, 30, 31]. These studies have uniformly found that participants in side-by-side settings, in which they share full views of one another and the workspace, perform better than participants using communications tools. Several recent studies [18, 20, 21, 30] have further shown that pairs perform better when they are using video tools that provide views of the workspace than when they are using audio or text-based communication alone.

Recently, there has been growing interest in the design of tools to allow collaborators to remotely perform tasks such as architectural planning. These activities, which we call collaborative physical tasks, involve intricate dependencies between verbal communication and physical actions. Telemedicine applications, remote repair systems, and collaborative design technologies are all examples of collaborative physical tasks. Any successful CSCW tool for remote collaboration on physical tasks will need to support the dependencies between speech and action found in these tasks.

To build tools that support collaborative physical tasks at a distance, we need a better understanding of the mechanisms through which the presence of a shared view of a workspace improves task performance. Although early research was satisfied in assessing whether the presence of a shared visual space affected the quality of task performance, recent research has begun to fill in the details (see [11, 30] for recent efforts in this direction). How, for example, does seeing a partner's gaze or actions alter a person's behavior? How does awareness that one is being watched influence one's own behavior? Understanding the mechanisms by which visual information affects communication is essential for designing systems to support remote collaboration on physical tasks. By identifying how visual information and speech can influence and substitute for one another, we can make informed decisions about when and how to provide this visual information in CSCW tools.

This paper tests the hypothesis that a shared view of a workspace allows a pair completing a physical task to substitute action for

language. In the process of viewing whether a worker has completed an instruction correctly, an instructor also receives as a side-effect, accurate information about whether the worker has understood his instructions.

To draw this conclusion, we conducted new analyses of the data collected in Kraut, Gergle and Fussell [30]. We use sequential analysis techniques [2, 3, 15, 22] to explore the role visual information plays in communication and demonstrate how it can be used in concert with, as a modifier of, or as a replacement for speech in a dyadic referential communication task. We briefly review prior performance findings and then detail the structural similarities and changes to communication that occur when language is complimented with visible actions. We provide some of the first quantitative demonstrations of the way in which actions and language interact and unfold over the duration of a communication episode, and how these sequences vary according to the presence of shared visual information. At a theoretical level, this work extends previous analyses of the effects of media on interpersonal communication [10] by providing a richer understanding of the way that physical actions and language are integrated to perform joint tasks and ground communication. At a more applied level, we use this knowledge to develop new design guidelines for technology to support distributed group work.

1.1 Action and Language in Communication

When people work together to solve a problem, they approach their task with different perspectives—different spatial viewpoints, different levels of background knowledge, and different roles. In order to coordinate their activities, they need a common set of goals and a shared language to discuss them. We use Herbert Clark's *Conversational Grounding* theory (e.g., [9, 12, 13]) as the framework for our investigations of the relationships between actions and speech.

In order to identify the critical elements of a shared visual space, we first need to understand how visual evidence is used for collaborative purposes. Clark observes that collaborative work occurs at multiple levels simultaneously. At one level, people collaborate to perform a joint task, such as co-constructing a puzzle. At a lower level, they use language (and other behaviors) to coordinate actions in order to perform the task. For example, they reach naming agreements for objects they can jointly see. Visual evidence can be helpful at both of these levels. At the higher level it can provide an up-to-date view of the state of the task. While at a lower level it may provide evidence about a partner's level of comprehension of the language being used.

1.1.1 Coordination and Grounding

Clark and his colleagues define conversational grounding as the collaborative process by which conversational partners work together to develop shared understanding. Common ground is comprised of the mutual knowledge, beliefs, attitudes, and expectations of the conversational partners [12], and the process of reaching common ground is referred to as grounding [13]. The difficulty of the grounding process, and thus the efficiency with which people communicate, is affected by a variety of factors including differences in spatial orientation [35], expertise [27], and socio-cultural background [16].

Brennan [7] extended this model by proposing that speakers continually form and test hypotheses about what a conversational partner knows at any moment both to plan utterances and to revise them after they have been delivered.

1.1.2 The Role of Visual Information in Conversational Grounding

Clark and Brennan [10] have argued that the features of a media may change the costs of grounding. For example, the media may change the time speakers have to plan an utterance, the evidence from which speakers can infer a listener's state of understanding, or the listener's ability to provide feedback to show understanding or ask for clarifications. Recently, Kraut, Fussell, and colleagues have applied a similar "decompositional" approach to understanding the role of visual information provided by different media in the grounding process [17, 18, 21, 29, 30].

Assessing comprehension. One way visual information affects communication is by acting as a source of evidence for understanding. Visible workspaces can provide situational awareness [14] that gives evidence both about the current state of the task and the members' activity levels. In order for speech to be effective, it needs to occur at the right moment. Visual information provides a mechanism for preparing subsequent statements and task descriptions by providing awareness of where the task is in relation to its overall end goal. It can also provide information regarding the current activity levels and availability of others.

Visual information has been described as one of the strongest sources for verifying mutual knowledge [12]. By witnessing the actions of a conversational partner, one can more readily recognize when the partner is behaving incorrectly, when they are confused and do not understand a directive, or when they do not understand the general task [7]. Hesitations, lack of action, and incorrect actions are all visible indicators of a lack of understanding. Imagine a pair in which a guide is remotely instructing a traveler on how to navigate from one part of campus to another. If the guide is given access to the proper visual information and the traveler turns left when she should have turned right, they can intervene with new instructions right away. In addition, the situational awareness provided by the visual information serves as a mechanism by which the guide can plan the timing of additional utterances. Continuing with the navigation scenario, if there is a particularly tricky sequence of turns, the guide can precisely issue directives one at a time if he can see where the traveler is. Without any visual feedback the guide must continually query the traveler and rely on her to provide an accurate description of where she is and what she has done in order to successfully guide her across campus. Thus, visual information provides situational awareness that may change both the structure (e.g., who is speaking when) and the content (e.g., what is said when) of an interaction.

Clark and Krych [11] recently demonstrated that when shared visual information was available the amount of time spent checking for comprehension in a Lego construction task was reduced from 21% to 5%. Similarly, Kraut and Fussell demonstrated that experts were more likely to elaborate on prior instructions in a remote bicycle repair task because they could better monitor the novices' comprehension when they had a shared visual space [29].

Because shared visual information facilitates awareness of whether an utterance has been understood, it also serves as a basis

for pairs to coordinate the formulation of their shared language used to describe task objects and locations. For example, if the guide in the previous example tells the traveler to "go kitty-corner" from where she is, and the traveler simply stands there, her inaction may be interpreted to mean that "kitty-corner" is not part of their shared language. A reformulation of "go diagonally to the left" may quickly remedy the situation. When there is a shared visual space such a comprehension error can easily be detected. By seeing the actions of the partner, the speaker gets immediate feedback regarding whether or not the addressee understood the instruction.

Assessing task performance. Visual information also serves a role in allowing judgments of task performance to be formed. Even if the speaker were addressing a robot, with no need for grounding, it would be important to have a feedback loop to get verification both that an instruction had been heard and that it had the intended effects. This loop of action and feedback is more general than language and is a basic tenet of HCI design principles.

Synchronizing messages. Partners in conversations have to time their contributions in order to ensure orderly turn exchanges. Features of media have been demonstrated to alter how efficiently turns are exchanged. For example, visual information allows pairs to overlap signals. When the pairs must rely on speech to describe their situation, talking at the same time will likely lead to confusion and incomprehensible speech. However, when a shared visual space is available the pairs can overlap their signals by relying on multiple modes of communication. For example, while the speaker describes the task, the addressees can demonstrate their understanding using action—in effect parallelizing the modes of communication. Whereas when using spoken language to achieve this, addressees often have to wait for an opportunity to interject, leading to a less efficient exchange. However, simply because this can be done does not mean it is optimal. If attentional focus is not shared, then the communicative intent of the action may be missed and yield misconceptions about the degree to which information is mutually shared.

1.1.3 The Principle of Least Collaborative Effort

The principle of least collaborative effort [13] states that speakers and listeners will strive to use the least amount of *joint* effort required to achieve their conversational goals. Therefore, we predict that when a shared visual environment provides cues to others' comprehension, both speakers and listeners will make use of these cues to the extent possible to reduce their collaborative effort. If, for example, a speaker can *see* whether the addressee has understood an instruction, and the addressee is aware that the speaker can see this, both may rely on the visual evidence in lieu of producing spoken evidence of comprehension such as "ok" or "I got it". By drawing inferences about understanding from task performance, the addressee gives off indications of their comprehension simply as a side effect of the action itself. Therefore, a shared visual space not only makes language more efficient, it may also eliminate the need for some language. This visual signal is also more likely to be an accurate reflection of understanding than the addressees' subsequent verbal confirmation. The partner's meta-cognition ("I think I understood what the speaker said") may not always be accurate.

Similarly, a speaker can modify his or her utterance midway through when presented with visual evidence of comprehension or lack thereof, minimizing the effort of speech production [8].

In previous studies, workers listening to instructions have been observed to manipulate shared visual space in ways that are consistent with the principle of least collaborative effort. For example, they intentionally alter camera angles so that they can use deictic pronouns such as "this" and "here" rather than lengthier verbal equivalents [29], and they increase their dwell time on visual targets when a remote helper is monitoring their gaze [8]. However, the analysis techniques used by these earlier studies were not sufficiently refined to demonstrate empirically how collaborators make use of visual information in conversational grounding.

1.2 The Current Study

In the current study, we use Clark's conversational grounding framework and the principle of least collaborative effort to investigate the relationships between visual information and speech in a collaborative puzzle task. Collaborative physical tasks can vary along a number of dimensions, including the number of participants, the degree of interdependency among actors, the number of objects being shared, the ease with which objects can be described, and the dynamics of the environment, to name but a few. The task on which we focus here, a jigsaw puzzle solving task, falls within a general class of "mentoring" collaborative physical tasks, in which one person directly manipulates objects with the guidance of one or more people, who frequently have greater expertise about the task. In collaborative physical tasks, people must maintain awareness of both the state of task objects and of one another's activities.

1.2.1 The Puzzle Task

To investigate relationships between vision and language in a controlled setting, we developed a collaborative online jigsaw puzzle task. In this task, one participant (the "Helper") instructs another participant (the "Worker") on how to complete a puzzle consisting of four blocks. The goal is for the Worker to arrange their pieces so they match a target solution the Helper is viewing.

The online implementation of the task allows us to manipulate with a high degree of specificity how much overlap exists between Helper and Worker views of the workspace and task properties such as the difficulty of labeling squares. Figure 1 shows a view of the screen from the Worker's side (left) and Helper's side (right). The Worker's screen consists of a staging area on the right in which the puzzle pieces are shown, and a work area on the left in which he/she constructs the puzzle. The Helper's screen shows the target solution on the right and a view (if any) of the Worker's work area on the left. This view into the Workers work area can be manipulated in a number of ways to investigate the effects of visual evidence on conversational grounding.

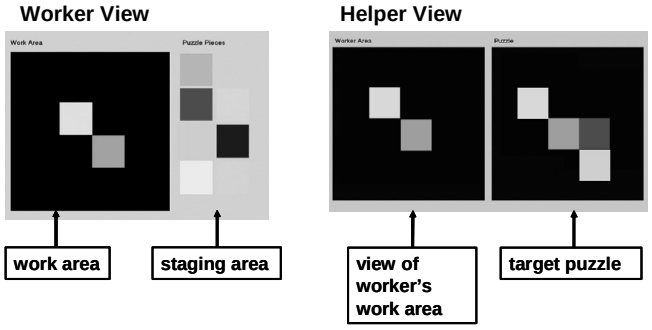

Figure 1. Worker (left) and Helper (right) displays.

The sequential nature of the puzzle task makes it ideal for investigating the interrelationships between speech and visible actions. In order to successfully add a piece to the puzzle, pairs had to first identify which was the correct piece and then guide that piece into the correct location. This identification-placement sequence had to be repeated four times to complete the puzzle, once for each piece. The basic task structure can be summarized as follows:

1. Identify the piece
2. Move the piece onto the workspace
3. Position the piece spatially within the larger work area
4. [Repeat steps 1 to 3 for subsequent pieces]
5. Jointly agree to be finished with the trial

Each of the basic steps above can be further decomposed into what Clark and Wilkes-Gibbs have called presentation-acceptance sequences [13]. For example, to conversationally ground step 1 (piece identification), the following sequence of events is required:

- Helper creates a referring expression for a puzzle piece
- Worker evidences understanding (or lack thereof) of this referring expression
- If understanding is evidenced, partners mutually agree that the piece has been identified
- If lack of understanding is evidenced, Helpers repair or replace the referring expression

Each component subtask can be realized via speech, action, or a combination of the two. Helpers can identify referents using verbal descriptions such as "the red piece" or by deictic expressions like "that one". Workers can evidence understanding by giving verbal acknowledgements (e.g., "ok"), by moving the correct piece into the workspace, or a combination of the two. Our hypothesis is that if a technology provides a shared view of the workspace, collaborators following Clark's principle of least collaborative effort will be more likely to use actions to ground each component of the task. Table 1 presents the type of evidence (spoken or visual) that is required at each substage.

Table 1. Type of information (spoken or visual) that can be used at various stages of the puzzle task.

Task Subgoals	Component subtasks	Shared Visual Space	No Shared Visual Space
Object Reference	Make reference to piece	Spoken	Spoken
	Verify referent	Spoken or Visual	Spoken
Object Placement	Make reference to piece	Spoken	Spoken
	Describe spatial positioning	Spoken	Spoken
	Verify spatial positioning	Spoken or Visual	Spoken

1.2.2 Previous Findings with the Puzzle Task

Earlier experiments using the jigsaw puzzle paradigm have found that a shared visual space improved performance more when the elements of the task were visually complex and lexically difficult to describe. A shared view of the work area was most beneficial when the pieces were difficult to label either because they were visually complex (tartan plaids) or a rapidly changing hue [21, 30].

Having a shared visual space also increased conversational efficiency. When Helpers could see what their partners were doing, significantly fewer words were required to perform the task. It appears that when there was no shared visual space, the Helper no longer had an up-to-date view of the task state and had to query the Worker to give an explicit description. The Worker in turn needed to respond with lengthier descriptions of the task state.

1.2.3 Using Sequential Analysis to Examine Grounding Sequences

As described above, the visual evidence provided by a technology appears to alter the way collaborators ground their utterances during each component of the puzzle task. However, although previous analyses suggest that communicators use visual evidence to facilitate grounding (e.g., [11, 17, 21, 29]), they have not used analytical techniques that can identify the precise ways that language and action are interrelated. In the current study, we build upon this prior work by using sequential analysis to determine if there is probable sequential structure, and if so does it vary by the availability of a shared view space.

By examining the patterns of communication using sequential data analysis techniques we can begin to develop a deeper understanding of both the role that visible action plays in communication and how it interacts with task structure. Consider, for example, the following examples of conversational strategies for achieving the same subgoal of positioning a piece in the puzzle:

- Helper states piece position → Worker positions the piece → Helper states correctness
- Helper states piece position → Worker states understanding → Worker positions the piece → Helper states correctness
- Helper states piece position → Worker states understanding → Worker positions the piece → Worker states correctness → Helper restates piece position → Worker restates correctness

These are all different strategies for attempting to achieve the same component subtask of telling a partner where to put a piece and ensuring that it occurs. Some of these strategies may be more or less efficient; this depends on the mediated form of communication available to the pairs. For example, the sequence of "Helper states piece position → Worker positions the piece → Helper states correctness" may be extremely efficient when the Helper can see what the Worker is doing. However, in the event that the pairs do not share a visual space, this strategy may be extremely ineffective, both in the errors produced and the added time it takes to repair misunderstandings. The sequential analysis described in the method allows us to examine how these sequences differ across various conditions of shared visual space.

1.3 Hypotheses

When a shared view of the workspace is present, Helpers will use Workers' actions as evidence of comprehension. They will be more likely to follow their own statements with another statement, without waiting for a Worker's verbal response, than when a shared view of the workspace is not present.

When a shared view of the workspace is present, Workers will be more likely to let their actions speak for themselves as evidence of their comprehension. They will be less likely to offer verbal acknowledgements of understanding when they know the Helper can see their actions than when they know the Helper cannot see these actions.

2. METHOD

Participant pairs played the role of Helper and Worker in a referential communication task. The Helper described a target jigsaw puzzle and instructed the Worker on how to complete it. The goal was for the Worker to arrange their pieces so that they matched the target that the Helper was viewing.

The experimental displays for the Worker and Helper were written as communicating Visual Basic programs. By constructing the displays computationally, we were able to manipulate the visual space that participants shared and the visual nature of their task.

The main manipulation of interest for this paper is the extent to which participants viewed the same work area. In any trial, the Helper could either see the Worker's work area with no delay or could not see the work area at all. These are the Immediate and No Shared Visual Space conditions. The availability of the shared visual space was manipulated within the pairs. Each pair participated in six blocks of four trials each. Both the Helper and the Worker were informed between blocks of what one another could see.

2.1 Participants

Participants consisted of 12 pairs of Carnegie Mellon University undergraduate students. The participants received $10.00 and were randomly assigned to the role of Helper or Worker.

2.2 Apparatus

The Helper and Worker were each seated in front of separate desktop computers with 21-inch monitors. They communicated over a high-quality, full-duplex audio link with no delay. The general structure of the Worker's and Helper's displays was shown in Figure 1 above. The worker's display contained a staging area where pieces for the puzzle were stored and a work area where the Worker constructed a four-piece puzzle. The Helper's display contained the puzzle target (the goal state) on the right and a duplicate of the Worker's work area on the left. This view either remained blank (in the no shared visual space condition) or showed an immediate replication of the Worker's work area (in the immediate shared visual space condition).

2.3 Measures

To investigate the relationship between the shared visual space and dialogue structure we developed a theoretically derived coding scheme to capture the primary purpose of each utterance and action. Since our principal interest was in determining under what circumstances action could replace spoken language, we transcribed separate streams for utterances and actions. Since the Worker could speak at the same time as the Helper, we devised three overlapping streams to accurately represent the communication between the pairs.

The final set of codes used in this study was represented by four major categories: Helper utterances, Worker utterances, Worker actions, and jointly occurring Worker utterances and actions. They are presented in Table 2.

Table 2. Utterance and behavior codes used.

Helper Utterances	
H_UTT$_{REFERENT}$	Helper makes reference to a specific piece (e.g., "Take the red one")
H_UTT$_{POSITION}$	Helper describes the position of a single piece (e.g., "Put that in the upper-left")
H_UTT$_{ACK_BEHAVIOR}$	Helper acknowledges a behavior (e.g., "Yes, that's perfect")
H_UTT$_{CONTEXT}$	Helper discusses contextual information about the task or process
Worker Utterances	
W_UTT$_{REF_OR_POS}$	Worker makes an utterance about a referent or a positional statement (e.g., "it's black and green?")
W_UTT$_{ACK_BEHAVIOR}$	Worker acknowledges a behavior (e.g., "I've done it")
W_UTT$_{ACK_UNDERSTAND}$	Worker acknowledges understanding (e.g., back-channels such as "mmm-hmm")
W_UTT$_{CONTEXT}$	Worker discusses contextual information about the task or process
Worker Utterances & Actions	
W_UTT+ACT$_{ACK_UND+MOV}$	Worker acknowledges and moves a piece close in time (e.g., "mmm-hmm" [Worker moves piece into the workspace])
W_UTT+ACT$_{ACK_BEH+POS}$	Worker acknowledges a behavior and positions a piece close in time (e.g., [Worker positions piece next to center square] "Done")
Worker Actions	
W_ACT$_{MOVE}$	Worker moves a piece into the workspace
W_ACT$_{REMOVE}$	Worker removes a piece from the workspace
W_ACT$_{POSITION}$	Worker positions a piece within the workspace or existing puzzle

The original data set contained onset and offset times capturing the entire duration of the utterance or action in the multi-stream event format previously described. This initial arrangement allowed us to look on the data with various lenses.

Figure 2 visualizes a small portion of the coded behaviors from the original data. Besides illustrating the raw data used for the sequential analyses, there are several points of interest in this small excerpt. For example, the first two bars on the graph represent a typical presentation-acceptance pair. The Helper begins at 3:15 by issuing a positional statement that tells the Worker where to put the puzzle piece. The Helper accepts this proposal by positioning the piece in the workspace. Notice that the Worker does not comment on whether or not she understood the position nor does she linguistically assess the quality of the move. Rather, the visual residue of her actions being visible implies her understanding. At around 3:19:50, the Helper treats this move as an acceptance and continues on with the next presentation of an instruction. It was this richness of exchanges that our coding scheme allowed us to capture.

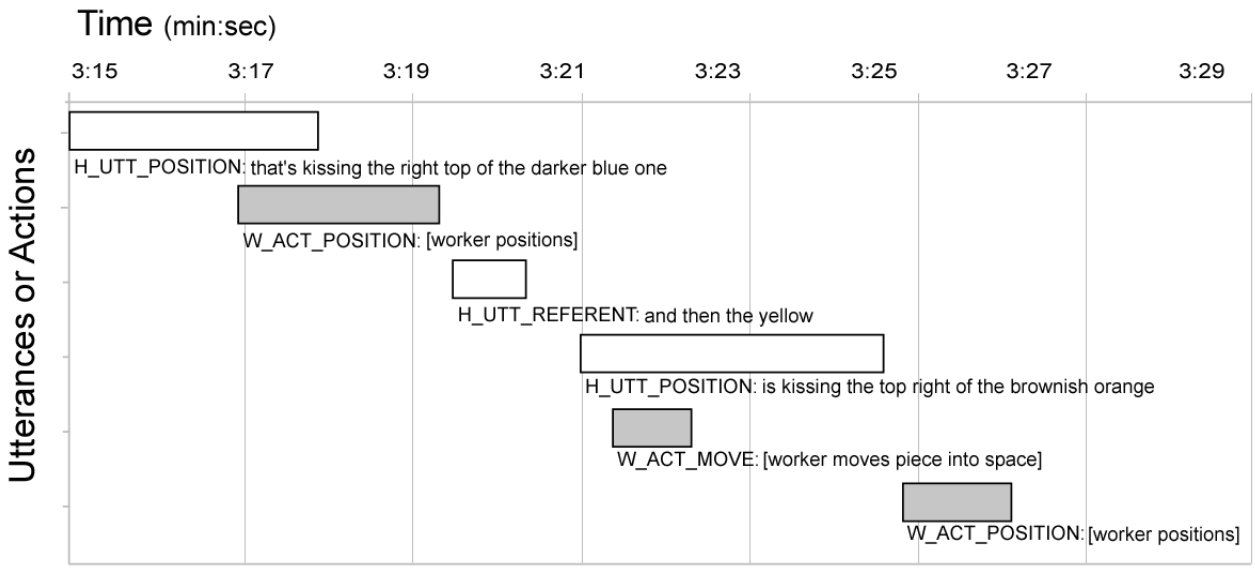

Figure 2. Demonstration of coded data (white = utterances; gray = actions).

Two independent coders classified a sample of utterances until they reached 90% agreement. They then each coded different transcripts, periodically coding a common transcript to ensure that the categories they used did not drift during the duration of the coding. Agreement remained high throughout.

2.3.1 Statistical Analysis

Our interest in this paper is on the impact of the fidelity of a shared visual space on conversational structure and tactics. We used log-linear modeling, lag-sequential analysis, and Chi-square techniques to examine the sequential nature of the data [2, 15, 22]. Using these techniques for analysis of group interactions is not a novel idea for studying group processes in HCI and CSCW [32, 33, 38], however, it is often times an under utilized technique due to the heavy time investment required.

Log-linear modeling is a general technique for analyzing multi-way contingency tables. This is a useful way to assess the global nature of the sequential structure by comparing the degree to which the data are sequentially structured versus being randomly distributed. Multivariate investigations allow you to explore how the sequential nature changes across experimental conditions. The lag-sequential method was used in this study as a confirmatory technique to look for theoretically driven sequential patterns that occur more often than expected by chance.

After using these two techniques to determine whether or not there are sequential differences across conditions, we use theoretically driven one degree of freedom Chi-square tests to examine particular areas of interest and determine exactly where the differences in sequence occur.

3. RESULTS

3.1 Event-Based Sequential Data Analysis

In the first portion of these analyses we model the sequences of the data by reducing the original multi-stream timed event sequential data into individual states—or event-sequential data. Basically, each temporal encoding was reduced to a single state with the overall order being determined by the onset time of the coded behavior. The original table consisted of 13 categories and 1413 cases.

We begin each model by first establishing that there was sequential structure to the data. If there were no sequential order, then we would expect one category to follow another at random, dependent only on the frequency of occurrence. Cell scores would simply represent the joint probabilities of the target and given categories. This initial test can be construed as similar to an omnibus test that provides license to continue more detailed testing regarding the nature of the sequential relations.

3.1.1 References to a Piece

If the process of making reference to a puzzle piece and confirming its correctness can be done either using spoken language or action (as suggested in our hypotheses), then we would expect vast differences in how the pairs communicated when they had a shared visual space versus when they did not. In order to explore if this was the case, we took one of the most structured aspects of the task—the component subtask of identifying and making successful reference to a puzzle piece—and explored its sequential event structure.

The process of successful reference begins with the Helper issuing a statement regarding which puzzle piece should be selected. For example, "It's kinda like a mauve color" would be the starting point of such a piece reference. We therefore constructed a 2 (**No SVS**; **Immediate**) × 2 (**H_UTT$_{REFERENT}$**; **~H_UTT$_{REFERENT}$**) × 13 (**All Categories**) matrix to represent the sequential frequencies between categories. The first dimension represents whether or not the pairs had a shared visual space and is referred to as the "SVS" dimension. The second dimension is referred to as the "Given" dimension and differentiates the cases when the initial expression occurred (**H_UTT$_{REFERENT}$**) versus those when it did not (**~H_UTT$_{REFERENT}$**). The third dimension is the "Target" dimension and differentiates among the utterances and actions that the Worker or Helper could perform following the initial expression. The resulting three-dimensional matrix contains cells with the

frequency of the transitions between the Target and Given events nested within the appropriate SVS condition.

An initial test of the model of independence revealed significant structure in the SVS × Given × Target matrix ($G^2(37)=564.8$, $p<0.001$). This indicated that it was highly unlikely that the observed cell frequencies were simply the result of random transitions. In other words, there was significant dependence between the dimensions of the table. We then proceeded to investigate the details of where this structure exists.

Since we were primarily interested in investigating the sequential differences due to whether or not the pairs had a shared visual space (i.e., whether the interaction of Given and Target categories varied across the experimental conditions), the proper model to test should include all main effects and two-way interactions. The results of such a model implied that the three-way interaction was indeed significant ($G^2(12)=33.412$, $p<0.001$). This suggests that there is sequential structure in the data, and that it varies across the experimental conditions.

In order to understand specifically where the sequential differences of interest occurred we went back to the initial independence model (i.e., the main effects model). Figure 3 shows the conditional probabilities and *z*-scores of the transitions between the code (**H_UTT$_{REFERENT}$**) and several subsequent categories of interest. Note that these diagrams do not represent all of the transitions. For instructional purposes we keep the number of nodes graphed to those that are significant and of theoretical interest.

A glance at the figure reveals where the conditional transitions vary and where large signed adjusted residuals exist (suggesting significant directional structure at greater or less than chance levels). For example, Figure 3 shows that following the Helper's description of a puzzle piece (**H_UTT$_{REFERENT}$**), the Worker moved the piece into the workspace 36% of the time when a shared visual space was available. However, when they did not have a shared visual space, this only occurred 19.6% of the time. Instead, the Helper issued an acknowledgement along with their movement 21.2% of the time. The *z*-scores in these figures serve to indicate the relative strength of the transitions while taking into account the overall frequencies of each of the categories.

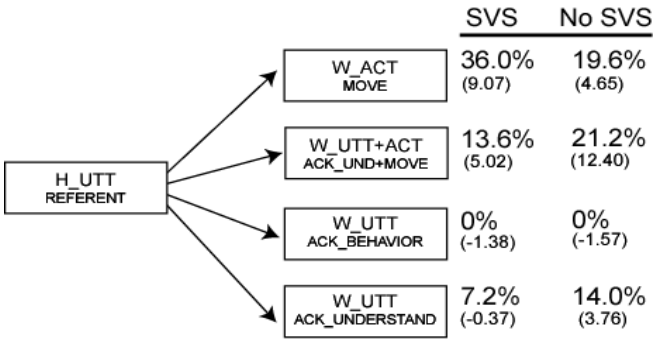

Figure 3. Conditional probabilities (percentages) and z-scores (in parenthesis) for models of piece referents.

If the pairs were indeed performing according to the principle of least collaborative effort, we should expect to find the transition between the Helper referent and the Worker movement more often than when there was no shared visual space. Similarly, when they had no shared space to rely on for grounding, we should expect the pairs to more often acknowledge the referent or move the piece while at the same time issuing an acknowledgement.

Table 3. Excerpts of pairs making object references with and without a shared visual space.

Shared Visual Space	No Shared Visual Space
H: OK, and the orange	**H:** Um, and then there's an orange brownish one
W: [Moved correct piece]	**W:** [Moved correct piece]
H: Um, touching the right corner, right top corner of the dark blue.	**W:** Yeah.
	H: That's touching the right top of the blue one

We found that when the pairs had a shared visual space they were much more likely to simply move the piece than to either move the piece and acknowledge that they had done so or simply acknowledge the statement (for the contrast, $\chi^2(1, N=169)=12.641$, $p<0.001$). When the pairs had a shared visual space, the Worker typically responded to the referent by simply moving the piece (as seen in the example in the left hand side of Table 3 and also illustrated in Figure 2). However, when there was no shared visual space, the Worker typically moved the piece and provided evidence using spoken language (as seen in the right side of Table 3).

3.1.2 Positioning a Piece

Similar to the prior model, when the Helper gave directives on where to position the piece, the Worker could respond in several ways depending on whether or not they were taking the media into account. In order to explore if this was the case, we took another commonly structured aspect of the task—the component subtask of successfully positioning a puzzle piece within the workspace and explored its sequential structure.

This process typically begins with the Helper issuing a statement regarding where a puzzle piece should be placed. For example, "You should put it in the upper-left corner". We then partitioned a similar table as described above but replaced the Given categories with the appropriate codes representing utterances about positional information (**H_UTT$_{POSITION}$;~H_UTT$_{POSITION}$**).

We again constructed a 2 (**No SVS**; **Immediate**) × 2 (**UTT$_{POSITION}$**; **~H_UTT$_{POSITION}$**) × 13 (**All Categories**) matrix. An initial test of the model of independence revealed significant structure in the SVS × Given × Target matrix ($G^2(37)=408.4$, $p<0.001$). We then proceeded to investigate the structure in a more detailed fashion.

Examining whether or not this structure varied across experimental conditions again requires us to test whether the interaction of Given and Target categories varied across the experimental conditions. The results suggest that the three-way interaction was indeed significant ($G^2(12)=21.2$, $p<0.05$). Once again, this suggests that there is sequential structure in the data, and that it varies across the experimental conditions.

In order to understand specifically where the sequential differences of interest occurred we again went back to the main effects model to investigate the significant sequential structure between the categories and how it differed across conditions of shared visual space. Figure 4 shows that following the Helper's description of piece placement (**H_UTT$_{POSITION}$**), the Worker moved the piece into the workspace 36.8% of the time when a

shared visual space was available and only used verbal acknowledgements of any sort in 12% of the cases (combining the three other categories displayed). However, when the pairs did not have a shared visual space, they simply positioned the piece only 17.0% of the time. Instead, the Helper issued an acknowledgement along with their positioning 13.2% and simply stated their understanding of where the piece should go 25.3% of the time (reserving the actual positioning of the piece an indeterminate number of turns later).

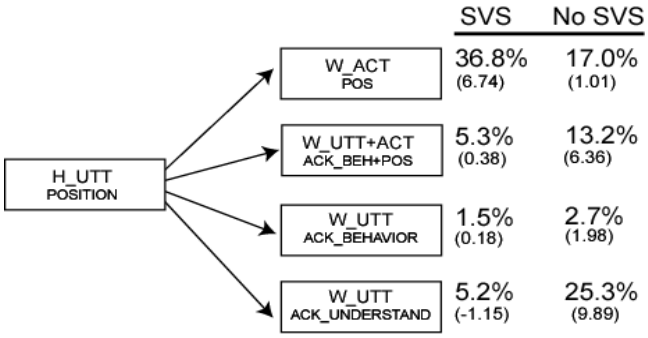

Figure 4. Conditional probabilities (percentages) and z-scores (in parenthesis) for models of piece position statements.

We tested these differences using a Chi-square analysis and determined that when the pairs had a shared visual space they were much more likely to simply move the piece than to either move the piece and acknowledge that they had done so or simply acknowledge the statement (for the contrast, $\chi^2(1, N=164)=34.427$, $p<0.001$).

Table 4. Excerpts of pairs making object references with and without a shared visual space.

Shared Visual Space	No Shared Visual Space
H: Put it corner to corner in the lower left.	**H:** And its bottom left corner touches the top right corner of the purple one.
W: [Positioned piece correctly]	**W:** [Positioned piece correctly]
H: Now take a light blue.	**W:** Mmm-kay. Got it.
	H: OK.

When the pairs had a shared visual space, the Worker typically responded to the positional information by positioning the piece (as seen in the left hand side of Table 4). However, when there was no shared visual space, the Worker typically positioned the piece and provided evidence of the action by using spoken language (as seen in the right side of Table 4).

4. DISCUSSION

The results demonstrate clearly that partners adapt their communication to the presence or absence of shared visual space. When a shared view of the workspace was available the Workers were more likely to let their actions "speak" and provide evidence of their comprehension. They were less likely to present verbal acknowledgements both when attempting to select the proper puzzle piece and when positioning a relevant piece within the workspace. The sequential analyses presented in this paper demonstrated that the Workers' actions replaced a typical utterance or action + utterance sequence when they knew that the Helper could see what they were doing. Similarly, the Helpers were more likely to use the Workers' actions as evidence of understanding by simply following the actions with their next description. By using actions to help ground their utterances, pairs in the shared visual space condition were able to communicate more efficiently.

It is important to note that in this experimental design the Worker's workspace always appeared the same regardless of whether the Helper could see what they were doing. Therefore, if the Worker were using a purely egocentric approach to communication it is not likely that the structure of the actions and utterances would change. Yet we clearly demonstrate that they adapt their communication in an effort to reduce collaborative effort when their partner can see their workspace. These results are consistent with Clark & Brennan's framework for analyzing the costs and benefits of different technologies. When media provide visual information about what the worker is doing, the ability of workers to ground utterances via actions sharply reduces the likelihood that they will provide verbal indicators of comprehension. Instead, they let their actions speak for themselves and demonstrate their understanding of the Helpers' utterances.

This work provides a major extension of previous research that demonstrated that pairs perform more quickly and accurately when they have a shared view of a common work area [11, 17, 18, 21, 30]. Yet, while having a shared visual space has been shown to improve both performance and conversational efficiency, this prior work did not describe precisely how pairs enhance their collaboration using visual information. We presented results demonstrating one way this occurs is by using action as evidence of comprehension when a shared view of the workspace is available.

4.1 Practical Design Implications

Knowledge of the mechanisms by which visual information can augment and change communication is crucial for designing systems to support remote collaboration, particularly in instances where support for collaborative physical tasks is the goal. By identifying the ways in which visual information and speech interoperate, we can begin to make informed design decisions regarding ways to support visual information in collaborative applications.

Our results highlight the importance of making it clear that people know precisely what remote collaborators can see in a shared workspace. It is not enough to simply allow others to see what is going on, but rather, mutual understanding of what is available to one another is needed. When confusion exists regarding what the Helpers can see, the pairs spend time trying to identify the mutually shared visual field. This reduces their overall efficiency since significant time is needed to identify what visual information is and is not shared [29].

In our puzzle task there are two levels with which the visual information seems particularly useful. At a higher level the pairs find it useful for task planning. For example, when planning subsequent directives the Helper often looks at surrounding contextual information. Previous work has suggested that the Helper often times looks to the instructions while formulating their description of the next step [19]. In this case, providing a wide-angle view of the workspace (a context-oriented view) is useful. However, when pairs are performing lower level

coordination of their language it is useful to have a focus-view of the workspace centered around the actions. Thus, for high-level task planning it may be useful to have a wider view of the work area, while for grounding communications it may be more useful to have focused views. A potential design avenue for simultaneously supporting these two levels might be through the creation of task specific focus + context designs. Initial design avenues in this area have been explored by Schafer and Bowman in exploring collaborative spatial navigation [34], and by Greenberg, Gutwin and Cockburn as general techniques in groupware applications [24]. Coupling these design explorations with detailed knowledge of how visual information serves the task may lead to a fruitful line of collaborative applications development for joint physical tasks.

We also demonstrated that when collaborators are aware of their partners' fields of view, asymmetric interfaces in which different parties have different modes of accessing the environment appear to be functional. Developing ways of providing awareness of others' views can enable efficient grounding and is crucial to the development of successful applications for remote collaboration on physical tasks.

In this study we demonstrated how awareness of others' views is critical. Actions provided a more efficient mechanism for establishing mutual understanding. The base rate ground truth was simply easier to establish when shared visual information was available. Rather than relying on imprecise conversation with another to determine if something had been done correctly, having it in view to verify mutual understanding was extremely useful, particularly in a tightly coordinated activity or one where the expertise is distributed.

These findings also suggest that using schematic representations in lieu of direct video feeds in low bandwidth conditions may be more useful to participants if they represent actions rather than the others' faces or bodies. For example, sensors might provide schematic feedback about what objects have been selected or moved. The value of schematic representations has been shown in similar settings by such tools as Gutwin and Penner's telepointer traces [26], which provide feedback about a partner's trajectory of cursor movements within a shared workspace.

4.2 Theoretical Implications

The results provide a significant advance for the conversational grounding framework on interpersonal communication. In addition to replicating previous work demonstrating the importance of visual evidence for conversational efficiency, we have provided a detailed analysis of precisely *how* this visual information is used by participants as they ground their utterances during a collaborative task. In conjunction with other recent work on the role of gesture (e.g., [11, 20]), eye gaze (e.g., [8]) and other nonverbal behaviors in the grounding process, our findings help contribute to a global theory of conversational grounding.

The results also add to theoretical attempts to identify how specific features of specific technologies affect communication, collaboration, and performance (e.g., [10, 28]) by providing a decompositional analysis of the nonverbal behaviors affected by features of a technology. By combining sequential analysis techniques with detailed coding of speech and actions, we have been able to show in much greater depth how the availability of a shared view of the workspace affects interaction. We believe that the application of sequential analysis to interactions across a wide range of computer-mediated communications tools will lead to sizeable advances in CSCW theory.

4.3 Limitations and Future Directions

Using a stylized task such as the collaborative puzzle task in this experiment has both strengths and weaknesses. The strength of this paradigm is that it allows us to precisely manipulate dimensions of shared visual space and characteristics of collaborative tasks, and it permits precise measurement of the actions workers take in response to the instructions. We believe that this level of control of the experimental setting is essential to uncover the interdependencies between language and action in collaborative physical tasks.

A possible limitation of the paradigm is that the jigsaw puzzle task oversimplifies these interdependencies because of the limited range of instructional utterances and limited range of worker actions that are possible. However, it is important to note that many more complex tasks, such as building a toy robot or repairing a bicycle, are comprised of the same sorts of object identification-object positioning sequences we have studied here. Thus, we believe that our findings regarding the relationships among base level actions and language are likely to hold even when tasks involve a much more complex range of actions.

To further assess the generalizability of our findings, we are currently extending this work to more complex and realistic task domains. For example, a new study looks at how speech and action are interrelated in a collaborative virtual gaming system in which one person is instructing another on how to navigate in a virtual world. This is a first step in testing the generalizability of these findings to more realistic task environments.

5. ACKNOWLEDGEMENTS

This research was funded by National Science Foundation Grants #9980013 and #0208903, and the first author was supported by an IBM PhD Fellowship. We would also like to thank Darrin Filer, James Hanson, John Lee and Gregory Li for their work on the experimental apparatus.

6. REFERENCES

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

[2] Bakeman, R., & Gottman, J. M. (1997). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge: Cambridge University Press.

[3] Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. Cambridge: Cambridge University Press.

[4] Barnard, P., May, J., & Salber, D. (1996). Deixis and points of view in media spaces: An empirical gesture. *Behaviour and Information Technology, 15*(1), 37-50.

[5] Bekker, M. M., Olson, J. S., & Olson, G. M. (1995). Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. In *Proceedings of DIS 95*, 157-166. NY: ACM Press.

[6] Boyle, E. A., Anderson, A. H., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language & Speech, 37*(1), 1-20.

[7] Brennan, S.E. (2004). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *World Situated Language Use: Psycholinguistic, Linguistic and Computational Perspectives on Bridging the Product and Action Traditions*. Cambridge, MA: MIT Press.

[8] Brennan, S. E. & Lockridge, C. B. (Under review). *Monitoring an addressee's visual attention: Effects of visual co-presence on referring in conversation.*

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

[10] 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.

[11] Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory & Language, 50*(1), 62-81.

[12] Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In B. L. W. A. K. Joshi, I. A. Sag (Ed.), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

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

[14] Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors Special Issue: Situation Awareness, 37*(1), 32-64.

[15] Fienberg, S. E. (1978). *The analysis of cross-classified categorical data*. Cambridge, MA: MIT Press.

[16] Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*, 378-391.

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

[18] 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. In *Proceedings of CHI 2003*, 513-520. NY: ACM Press.

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

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

[21] Gergle, D., Millen, D. E., Kraut, R. E., & Fussell, S. R. (2004). Persistence matters: Making the most of chat in tightly-coupled work. In *Proceedings of CHI 2004*. NY: ACM Press.

[22] Goodman, L. A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis*. Cambridge, MA: Abt Books.

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

[24] Greenberg, S., Gutwin, C., & Cockburn, A. (1996). Awareness through fisheye views in relaxed-WYSIWIS groupware. In *Proceedings of Graphics Interface*, 28-38.

[25] Gutwin, C. & Greenberg, S. (2001) A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Journal of Computer-Supported Cooperative Work, 3-4*, 411-446.

[26] Gutwin, C., & Penner, R. (2002). Improving interpretation of remote gestures with telepointer traces. In *Proceedings of CSCW 2002*, 49-57. NY: ACM Press.

[27] Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116*, 26-37.

[28] Kraut, R. E., Fussell, S. R., Brennan, S. E., & Siegel, J. (2002). Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. In P. Hinds & S. Kiesler (Eds.) *Distributed work* (pp. 137-162). Cambridge, MA: MIT Press.

[29] Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction, 18*(1), 13-49.

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

[31] Nardi, B. A., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., & Sclabassi, R. (1993). Turning Away from Talking Heads: The Use of Video-as-Data in Neurosurgery. In *Proceedings of ACM INTERCHI'93*, 327-334.

[32] Olson, G. M., Herbsleb, J. D., & Rueter, H. H. (1994). Characterizing the Sequential Structure of Interactive Behaviors Through Statistical and Grammatical Techniques. *Human-Computer Interaction, 9*(3,4), 427-472.

[33] Sanderson, P. M., & Fisher, C. (1993). Exploratory Sequential Data Analysis in Practice. In *Proceedings of ACM INTERCHI'93—Adjunct Proceedings*, p. 221.

[34] Schafer, W., & Bowman, D. (2003). A comparison of traditional and fisheye radar view techniques for spatial collaboration. In *Proceedings of Graphics Interface*, 23-46.

[35] Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*, 1-24.

[36] Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

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

[38] Weingart, L. (1997). How did they do that? The ways and means of studying group processes. *Research in Organizational Behavior, 19*, 40-89.