Modeling the Impact of Shared Visual Information on Collaborative Reference

Darren Gergle[†], Carolyn P. Rosé[‡], Robert E. Kraut[‡]

[†]Center for Technology and Social Behavior
Northwestern University
2240 Campus Drive
Evanston, IL 60208
dgergle@northwestern.edu

[‡]Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
cprose@cs.cmu.edu, robert.kraut@cmu.edu

ABSTRACT

A number of recent studies have demonstrated that groups benefit considerably from access to shared visual information. This is due, in part, to the communicative efficiencies provided by the shared visual context. However, a large gap exists between our current theoretical understanding and our existing models. We address this gap by developing a computational model that integrates linguistic cues with visual cues in a way that effectively models reference during tightly-coupled, task-oriented interactions. The results demonstrate that an integrated model significantly outperforms existing language-only and visual-only models. The findings can be used to inform and augment the development of conversational agents, applications that dynamically track discourse and collaborative interactions, and dialogue managers for natural language interfaces.

Author Keywords

Shared visual information, multimodal interaction, language use, discourse, communication, and modeling.

ACM Classification Keywords

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

INTRODUCTION

In order to develop widely deployable and successful conversational agents that interact with humans during collaborative physical tasks [8], applications that dynamically restructure their environment to minimize ambiguity, or video-mediated communication systems that adapt their views on the basis of a predictive model of what a speaking partner needs to see [39], we need a richer computational description of the ways in which shared visual information influences collaborative reference. In this paper we present a computational model that precisely describes how visual cues are combined with linguistic cues to enable effective reference during tightly-coupled, task-oriented interactions. The results demonstrate that an integrated model significantly outperforms both language-only and visual-only models of reference resolution.

BACKGROUND

A number of behavioral studies have begun to uncover the relation between shared visual information and spoken language use. For example, conversational pairs are more likely to replace full noun phrase (NP) descriptions with deictic pronouns such as "*that*" when shared visual information is available [19]. Distributional patterns of proximity markers (e.g., *this/here* vs. *that/there*) change according to whether speakers perceive themselves to be physically co-present or remote from their partner [7, 19]. And people use shorter and more syntactically simple language [42], and produce different surface realizations [9], when gestures or actions accompany their speech. Together this work suggests that both the linguistic and visual context shared by a collaborative pair has an impact on patterns of reference. Yet, a major gap exists between these empirical findings and the current state of technologies that deal with collaborative reference.

Surprisingly, the vast majority of computational models of reference rely solely on linguistic information and disregard the surrounding visual context [5, 29, 47, 50]. Without a more complete computational account of reference, we run the risk of developing agents, systems and technologies that fail. For example, if the goal were to develop a conversational agent for everyday interaction, then the agent needs to keep up its end of the conversational bargain by speaking and behaving in a natural way. It needs to understand speech and behaviors generated by people in real-world environments, and conversely, it needs to generate speech and actions in line with natural human behaviors. Studies have shown that when this does not occur, people overcompensate and adapt their communication patterns in ways that are unnatural (e.g., by producing hyper-articulated speech or adjusting their rate of disfluency [40, 41]), and these adaptations often lead to difficulties for computing systems.

MOTIVATION

There are several reasons for developing a computational model of referring behavior in shared visual contexts. First, an integrated model provides a deeper theoretical understanding of how humans make use of various forms of shared visual information in their everyday communication. Second, an explicit computational model can be used to inform the development of a range of technologies to support distributed group collaboration. Finally, an integrated model can be used to increase the robustness of existing interactive agents and dialogue managers that converse with humans in real-world situated environments.

A number of behavioral studies have demonstrated the need for a more detailed theoretical understanding of referring behavior in the presence of shared visual information [52]. Although these studies have shown that shared visual information about the objects and workspace can influence collaboration and communication in task-oriented interactions [33, 38], an explicit theoretical description of how this is possible and the mechanisms by which it occurs are underspecified. Theories such as Clark and colleagues' Grounding Theory [12, 13] provide excellent conceptualizations of human communication as a joint activity, yet they often remain modest in the details provided about the mechanisms and processes underlying successful communication. A detailed computational description of these processes can expose implicit and possibly inadequate assumptions underlying our current understanding.

The development of a multimodal model of reference can also yield practical guidelines for the development of technologies to support collaboration. Video-mediated communication systems, shared media spaces, and collaborative virtual environments are all technologies developed to support joint activities between geographically distributed groups. Yet, without a clear understanding of how visual information impacts language use we may unintentionally disrupt the critical information required for successful communication [4, 22].

Finally, there are a number of educational applications of language technologies such as tutorial dialogue [35] and adaptive collaborative learning support [27], where text processing technologies may be used to process student explanations in the context of a running dialogue with a computer agent or with one or more human peer learners [17]. Interventions triggered by the resulting analysis may be in the form of simple prompts or full tutorial dialogue interactions. Thus, an additional motivation for this work is to improve the performance of state-of-the-art models of communication currently used to support conversational interactions involving intelligent agents [1, 16].

THE DIFFICULTIES OF TRACKING REFERENCE IN COLLABORATIVE DISCOURSE

Natural language provides a number of ways for someone to refer to things. For example, in the puzzle study paradigm we developed [34] (and shown in Figure 1), an entity described as "the bright blue block" by the Helper may subsequently be referenced using a variety of forms such as: *it*, *this*, *that*, *the piece*, *that bright blue one*, *the brightest blue piece*, etc. Each of these referring expressions contains clues about the status of a given object in a pair's current model of the task [26]. For example, it is unlikely that the Helper would use the pronoun "it" to refer to "the bright blue block" if she had since discussed several other pieces. Similarly, the Helper should use the phrase "the brightest blue piece," only if she knows that she shares visual access to three blocks of different shades of blue with her partner.

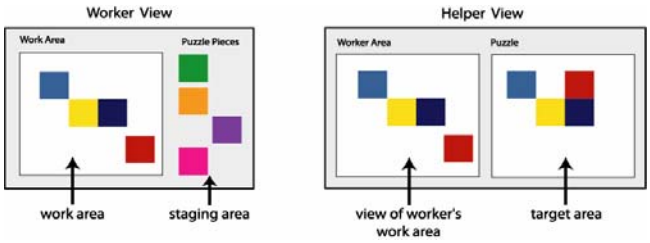

Figure 1. Puzzle study task. Worker's view (left) and Helper's view (right).

Linguistic Context in Support of Reference

In spoken dialogue, licensed referents[1] are often introduced through the prior linguistic context. This sets the stage for the later use of efficient referring expressions such as pronouns. Consider the following excerpt drawn from our previous work [20], whereby a Helper describes to a Worker how to construct an arrangement of colored blocks so they match a solution that only the Helper has visual access to:

(1) **Helper**: Take the dark red piece.
Helper: Overlap it over the orange halfway.

In excerpt (1), the first utterance uses the definite-NP "the dark red piece" to introduce a new discourse entity. Assuming the Worker has correctly heard the utterance, the Helper can now expect the entity to be the current focus of the discourse as established by the linguistic context [24]. This status provides license for the dark red piece to be subsequently referred to using a pronominal expression such as "it," in the second utterance.

Visual Context in Support of Reference

In contrast to the prior linguistic example, during task-oriented collaborations with physical objects, the visual context often plays a critical role in determining which objects are salient parts of a conversation. The following example demonstrates that it is not only the linguistic context that determines the potential antecedents for a pronominal expression, but also the shared visual context:

(2) **Helper**: All right, uh, take, um, the darkest orange block.
Worker: OK.
Worker: *[moved incorrect piece]*
Helper: Oh, that's not it.

[1] Licensed referents are those objects or entities that are syntactically available for future reference.

In excerpt (2), both the linguistic and visual information provide entities that could be potential targets of a referring expression. In this excerpt, the first pronoun "that," specifies the "*[…incorrect piece]*" that was physically moved into the shared visual workspace. While the second pronoun, "it," has as its antecedent the object specified by the definite-NP, "the darkest orange block."

Another problem when applying models based exclusively on linguistic properties to the puzzle study data is in the predicted use of a pronoun. In the following example, the visual information creates ambiguity for the pair that results in a full NP being repeated, while a model based solely on linguistic context would claim it is not needed.

(3) **Helper**: The bluish block goes in the upper right corner.
Worker: *[Blue block positioned in the workspace]*
Worker: *[Green block re-positioned in the workspace]*
Helper: The bluish block should be all the way in the corner.

In excerpt (3), if the model only accounted for the spoken contributions and disregarded the two visible moves, the repeated use of "The bluish block" in the last utterance would appear incoherent. Instead, the use of a pronominal phrase, "It should be all the way in the corner," would seem more coherent. This example demonstrates that the visual information introduces ambiguity regarding the most salient entity for the pair, and hence, which entity is the most likely referent of a pronominal expression.

Toward an Integrated Model

A number of existing computational models of reference will accurately resolve the pronoun in excerpt (1) but fail to do so in excerpts like (2). Similarly, the same models would have difficulty describing the use of the repeated NP in excerpt (3). Together these examples demonstrate a number of ways that *both* the linguistic and visual context serve a central role in the ability of pairs to make use of efficient communication tactics such as pronominal reference.

Recently, a handful of systems have begun to integrate visual and linguistic information for reference resolution (e.g., [10, 31, 32]). Typically the expressions available in these systems are part of a command language bound to particular functions known by both the user and system (e.g., "open" a "folder," or "tell me about" a "*[pointer hovering over a position on a map]*"). While these systems have made significant progress in implementing working systems (e.g., [2]), their goals differ somewhat in that they typically aim to support specific interaction techniques. Our work aims to develop a richer theoretical understanding of human-to-human communication in the presence of various forms of shared visual information and to use this understanding of the interplay between linguistic context and shared visual information to develop a more general account of situated interpersonal communication. These findings can then be used to inform further refinement of existing multimodal systems.

Our approach is most closely related to a recent investigation by Byron and colleagues that explored the role of shared visual information in a task-oriented, human-to-human collaborative virtual environment [6]. They compared the results of a language-only model with a visual-only model, and developed a visual salience algorithm to rank objects according to recency, exposure time, and visual uniqueness. In a hand-processed evaluation, they found that a visual-only model accounted for 31.3% of the referring expressions, and that adding semantic restrictions (e.g., "open that" could only match objects that could be opened, such as a door) increased performance to 52.2%. This model differs from the work reported here in that it does not make simultaneous use of both visual and linguistic salience information. So, for example, referring expressions cannot be resolved to entities that have been mentioned but which are not visible. Furthermore, it could fail to resolve references that the linguistic context determines are highly salient and the visual context does not. Therefore, in addition to language-only and visual-only models, we develop an integrated model that uses a balance of linguistic and visual salience to support resolution.

THE MODELING FRAMEWORK

Our modeling framework augments a rule-based model of spoken discourse to account for the reference patterns found in various visual conditions. The approach adopts the ideas of Centering Theory originally developed by Grosz and colleagues [23, 24]. Centering Theory is a dynamic model developed to describe the mutual attentional state of discourse participants. It has been used to explore such linguistic concepts as common ground and discourse object salience [3, 30], and it provides a salience-based, real-time, dynamic method for describing discourse focus.

Overview of the Modeling Architecture

The major components of the modeling architecture are a Running Discourse History, a Transient Knowledge Base, a World Knowledge component, and a set of proposed ranking strategies for ordering the entities contained in the Transient Knowledge Base.

Running Discourse History

The Running Discourse History captures the utterances, actions and objects that can serve as potential referents in future utterances. From these various streams of data we parse and extract the major units needed for inclusion in the models. The visual and linguistic information from both the Helper and Worker are captured independently and synchronized on the basis of a common timestamp.

Transient Knowledge Base

At the heart of the model is a dynamically updated ranked-list of entities that contains the constituent entities ordered by their relative salience. The highest-ranked entity in the Transient Knowledge Base is considered the most likely candidate for a subsequent referring expression. In this way, the Transient Knowledge Base captures the current focus of the discourse, whether it is a recently mentioned object or a

highly prominent visible object or action. A number of algorithms describe how to rank this list in spoken discourse [5, 47, 49], yet, little work has been done to explore the role of visual salience and how it influences the ranking of entities in a shared model of discourse.

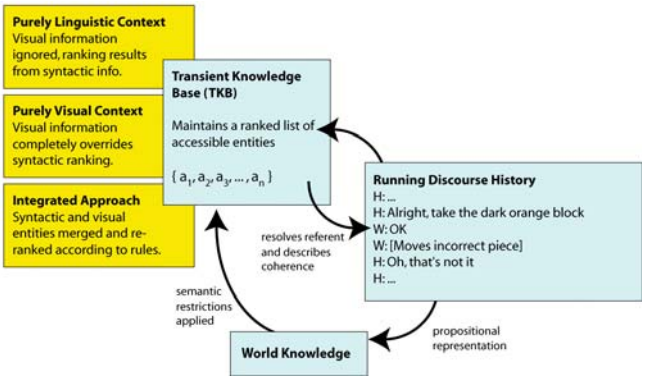

Figure 2. Modeling framework. Basic components (blue) and hypothesized ranking strategies (yellow).

Linguistic entities and their salience ranking. The linguistic entities used to populate the Transient Knowledge Base are extracted by parsing, chunking and tagging the utterances contained in the Running Discourse History. Each linguistic object has a number of features that determines its availability as a potential referent and its ranking within the list. Grammatical function is the ranking mechanism used in this paper, and agreement constraints such as those based on gender, plurality (i.e., number), and binding constraints[2] [11, 29] are enforced when resolving a referential expression.

Visual entities and their salience ranking. In addition to the linguistic entities, the Transient Knowledge Base can be populated with visual entities. In the puzzle paradigm these elements consist of the blocks and their associated properties. Obviously, there are a great number of visual features that can impact the visual salience of a particular entity in a particular environment [45]. However, one particular attribute that is highly salient is object motion. For this reason, we use the recency of object motion as the primary visual feature in this paper. If visual information, as measured by this rather coarse attribute of salience, influences referring behaviors then a more complete future investigation of visual salience is warranted.

Integrating the elements of the linguistic and visual salience rankings. Together, the linguistic entity list and the visual entity list are intended to capture all the entities that could potentially be referenced. The evaluation presented in this paper examines the balance between visual and linguistic salience of the objects contained in the Transient Knowledge Base, and the hypothesized ranking strategies used to model the salience of the elements in a multimodal, task-oriented environment.

World Knowledge

The World Knowledge component is used to capture any previously existing shared knowledge the pairs may have and also serves to enforce semantic restrictions. The models in this paper match the evaluations of earlier pronoun resolution evaluations which assume no world knowledge, and rely instead on syntactic agreement criteria and binding constraints [48, 49]. However, this component is included in the framework in order to support future modeling endeavors.

THE PUZZLE CORPUS

The data for the evaluation were randomly selected trials from two major manipulations of our previously collected puzzle study data [20, 34]. As Table 1 demonstrates, the data consisted of 14 dialogues from the No Shared Visual Information condition where the Helper could not see the Worker's workspace at all. In this condition, the pairs needed to successfully complete the task using only linguistic information. Another 22 dialogues were selected from the Shared Visual Information condition, where the Helper received immediate visual feedback about the state of the Worker's work area. Each dialogue was collected from a unique participant pair.

Task Condition	**Corpus Statistics**		
	Dialogues	*Utterances*	*Pronouns*
No Shared Visual Information	14	336	76
Shared Visual Information	22	327	217
	36	663	293

Table 1. Overview of the evaluation corpus.

MODEL EVALUATION

Three models were developed in order to address the question of whether or not an integrated model of reference resolution could be more successful than a language-only model or a visual-only model. The following sections present a detailed description of the models and their development, an empirical evaluation of their performance, and a reflection on the findings and future avenues for modeling. The evaluation in this paper is a hand-processed evaluation on data that were automatically extracted.

Hypothesized Ranking Strategies

Three ranking strategies are examined, each of which corresponds to a hypothesized method for ranking possible referents in the Transient Knowledge Base. The ranking strategies are represented in yellow in Figure 2, and are described here:

Purely linguistic context. One hypothesis is that the visual information is completely disregarded and the entities are salient purely on the basis of linguistic information. While prior empirical work suggests this should not be the case, several existing computational models function at this level.

[2] Chomsky's Binding Theory (1982) describes whether a pronoun needs to be locally bound. For example, a reflexive pronoun such as "himself" needs to be locally bound, as in "John painted himself" versus "him" that cannot be locally bound, as in "John painted him."

Purely visual context. A second possibility is that the visual information completely overrides the linguistic salience. Thus, visual information dominates the discourse structure when it is available and relegates linguistic information to a subordinate role. This too should be unlikely given the fact that not all discourse deals with external elements from the surrounding world.

A balance of syntactic and visual context. A third hypothesis is that both linguistic and visual entities are required in order to accurately and perspicuously account for patterns of observed referring behavior. Salient discourse entities result from some balance of linguistic salience and visual salience.

Data Pre-Processing

Several challenges exist in preparing a multimodal corpus for use with models of reference, and a number of preparatory steps need to be taken in order to prepare the elements of the linguistic and visual context.

Dialogue transcription, segmentation, and alignment. To transcribe and segment the dialogue, we followed guidelines established by Heeman and Allen [28] for segmenting unconstrained, multiparty dialogue.

POS-tagging, noun phrase extraction, and subject/object tagging. To generate the appropriate features and entities, part-of-speech (POS) tagging, chunking (e.g., NP chunking), and subject/object detection was performed on the corpus. Each contribution was parsed using a memory-based shallow parser that was trained on the Penn Treebank II Wall Street Journal Corpus. The Tilburg Memory-Based Learner (TiMBL) v5.1 software package [14, 15] was used to extract the entities and tags needed for the language features of the models.

The POS-tags are used to identify pronouns of various types. The output from the chunker identifies NPs that are the essential entities required to populate the Transient Knowledge Base. These constitute both the pronouns that need to be resolved as well as the entities that make up the coreference chains and may specify the referents of various pronominal expressions. Finally, the subject/object detection provides syntactic information for ranking the entities by grammatical function.

Extracting the visual data. In order to work with the visual information from the shared visual workspace, the actions and visible elements were extracted from detailed interaction logs provided by the puzzle study software. These logs contained information that could be pruned to develop the relevant data structures for the models.

The linguistic data was then aligned with the visual data using a common timestamp. Each contribution has a start and finish time, and the visual state of the shared workspace can be resolved whenever it is needed by the model.

Model Details

As previously mentioned, the models in this evaluation are based on Centering Theory [24, 25]. However, one area where the original formulation of Centering Theory and its related algorithms [5] are deficient is in their ability to describe reference in an online and real-time fashion at a finer-grained level of resolution than a complete sentence. This poses a problem for extending the model to account for visual information, since the stream of visual information is continuous and not easily partitioned into discrete bins in the same way as utterances or sentences. This was solved in part, by a solution proposed by Tetreault and his Left-Right Centering (LRC) algorithm [49]. The LRC algorithm makes provisions for incremental resolution by maintaining a partially-ordered list of potential entities that are available at any point during the construction of an utterance. This dynamic, real-time list of entities allows one to capture the attentional state of a discourse at a finer level of granularity than previous algorithms and makes it a natural candidate for extension to the visual domain.

The Language-Only Model

The LRC algorithm is used as the base model and algorithm for the language-only model. It uses grammatical function as a central mechanism for resolving references. It resolves references by first searching within the current utterance for possible antecedents, and makes co-specification links when it finds an antecedent that adheres to syntactic agreement and binding constraints. If a match is not found the algorithm then searches the lists of possible antecedents in prior utterances in a similar fashion. The primary structure employed in the language-only model is a ranked entity list sorted by linguistic salience. In this evaluation, the output of the subject/object detector was used to generate syntactic labels that would allow a given NP to be ranked in the entity list according to grammatical function. The grammatical function ranking was determined by the following precedence ranking: *Subject* > *Direct Object* > *Indirect Object* > *Other*. Any remaining ties (e.g., an utterance with two direct objects) were resolved according to a left-to-right breadth-first traversal of the parse tree.

The Visual-Only Model

The visual-only model captured the visible actions and utilized an approach based on visual salience. This method captured the relevant visual objects in the puzzle task and ranked them according to the level of recency with which they were active. Given the highly controlled visual environment that was used in the puzzle studies, timing information is available about when the pieces become visible, are moved, or are removed from the shared workspace. In the visual-only model, an ordered list of entities that comprise the shared visual space was maintained. The entities are included in the list if they were visible to both the Helper and Worker, and then they were ranked according to the recency of their activation.

The Integrated Model

The integrated model took advantage of the salience list generated from the language-only model and integrated it with that of the visual-only model. The method of integrating the list was informed by general perceptual

psychology principles stating that highly active visual objects attract attentional processes [45]. The visual objects were added to the top of the linguistic-salience list which essentially rendered them the focus of the joint activity. However, people's attention to static objects tends to fade over time. Following prior work that demonstrated the utility of a visual decay function [6, 31], a three-second threshold existed on the lifespan of a visual entity. From the time since the object was last active, it remained on the list for three seconds. After the time expired, the object was removed and the list returned to its prior state. This mechanism was intended to capture the notion that active objects are at the center of shared attention in a collaborative task for a short period of time, after which the speakers revert to their recent linguistic history for the context of an interaction.

RESULTS

Measures

The basic success measure used in this experiment is Mitkov's [36] measure of the total number of pronouns correctly resolved over the total number of pronouns attempted. Before model performance can be assessed, the actual antecedents of the pronouns need to be marked. Two expert coders marked the antecedents for each pronoun in the corpus. Each coder went through the segmented transcripts line by line and when they identified a pronoun they scored its antecedent, whether it was a noun phrase, another pronoun, or a visual entity or action. For the evaluation set examined in this study, the coders independently rated each of the 293 pronouns in the corpus. Scores were counted correct if both of the coders identified the pronoun and tagged the same antecedent. However, if only one of the coders identified a pronoun, or if the antecedents were different, their coding was scored as incorrect. Overall, the coders reached a reliability of 88% overall agreement. The remaining anomalies were resolved by discussion.

Statistical Analysis

A number of analysis techniques were used to describe the performance of the models. A logistic regression was used to examine the overall performance of the models and to capture higher-order interactions of interest. The logistic model included Model Type (*Language*, *Visual*, *Integrated*), Lexical Complexity (*Solid* or *Plaid*), and Pronoun Type (*Personal*, *Demonstrative*, or *Demonstrative + NP*). Because the pronouns existed in a discourse, there was the possibility that observations within a trial were not independent of one another. Therefore, Trial was modeled as a random effect. In addition, all two-way interactions were included in the model. Three-way interactions were also investigated, but were not found to be significant, and were removed from the final analysis.

In order to directly compare the performance of the models on each pronoun encountered, a second analysis involved the creation of a confusion matrix. McNemar's test was used to test the agreement between the models and to help characterize differences in their performance. This approach examined each pronoun that had been resolved for each model, and provided an indication of whether or not a particular model faired better on the same piece of data, which in turn provided an aggregate statistical indication of model performance and also allowed a more detailed investigation of the patterns of failure that occurred. For example, examination of the data points in the off-diagonals of the confusion matrix could provide an indication of how one particular model outperformed another.

Model Performance Results

Table 2 presents the pronoun resolution rates of the three models according to whether the pairs shared visual information, and whether the puzzles included simple solid colors or more lexically complex plaid pieces.

Performance in the No Shared Visual Information condition

As can be seen in the "Total" columns of Table 2, the language-only model correctly resolved 67.1% of the referring expressions when applied to the set of dialogues where only language could be used to solve the task. However, when the language-only model was applied to the dialogues from the task conditions where shared visual information was available it performance diminished significantly. It only resolved 49.3% of the referring expressions correctly ($\chi^2_{(1, N=293)} = 7.17$, $p < .01$).

	No Shared Visual Information			**Shared Visual Information**		
	Solids	*Plaids*	***Total***	*Solids*	*Plaids*	***Total***
Language Model	*70.0% (21 / 30)*	*65.2% (30 / 46)*	***67.1% (51 / 76)***	*43.6% (17 / 39)*	*50.6% (90 / 178)*	***49.3% (107 / 217)***
Visual Model	*n/a*	*n/a*	***n/a***	*66.7% (26 / 39)*	*61.2% (109 / 178)*	***62.2% (135 / 217)***
Integrated Model	*70.0% (21 / 30)*	*65.2% (30 / 46)*	***67.1% (51 / 76)***	*69.2% (27 / 39)*	*73.0% (130 / 178)*	***72.4% (157 / 217)***

Table 2. Success rates for resolving pronominal reference.

The integrated model performed at the same level as the language-only model when there was no shared visual information available. The integrated model essentially reverts back to a language-only model, achieving the same 67.1% performance.

Performance in the Shared Visual Information condition

A direct comparison between the three models of reference can be made by exploring their performance on the data in the cases in which shared visual information was available. Model Type was a significant factor in the model, $G^2_{(2)} = 15.21$, $p < .001$, and contrasts between the different levels of Model Type revealed significant differences between the performance of each model (at $p < .05$ in all cases).

The language-only model correctly resolved 49.3% of the pronouns when applied to the trials performed in the presence of shared visual information. However, when the visual-only model was applied to the same data, it correctly

resolved 62.2% of the pronouns. The difference in performance between these two models was substantial, $\chi^2_{(1, N=217)} = 8.52$, $p < .01$, and indicated a major performance benefit for the visual model. The confusion matrix presented in Table 3 demonstrates that both the visual-only and language-only models correctly resolved pronouns missed by the other. An informal examination of the cases that the visual-only model correctly resolved and the language-only model failed (27.1% of the cases) revealed a few trends. A large proportion of these cases appeared to occur when an efficient referring expression was used to reference an entity that was not mentioned in the prior linguistic stream. For example, "Oh, that is one we need, so put it to the upper left". Another case was when contrastive statements were made regarding the current visible object and the targeted referent, for example, "…a darker color than that." There were also a small number of references that the language-only model mistook to refer to sub-features of a piece, while the visual-only model correctly suggested the whole block as an entity.

		Language	
		Incorrect	*Correct*
Visual	*Incorrect*	50 (23.0%)	32 (14.8%)
	Correct	59 (27.1%)	75 (34.6%)

Table 3. Confusion matrix between the Language Model and the Visual Model.

An informal examination of the cases that the language-only model correctly resolved and the visual-only model failed (14.8% of the cases) also revealed some interesting trends. First, there were a number of cases where the language-only model successfully resolved pronouns to linguistic entities where the last piece of visual information would have led to an incorrect referent. These included cases when the discourse included longer discussions regarding the details of a piece or a layout. There were also cases where the language-only model could successfully resolve references within a sentence. And finally, there were a small number of cases where an incorrect visual object was available and the pronoun instead referred to a previously introduced linguistic entity (e.g., "no, it is a different yellow piece").

Returning to the right-hand side of Table 2, when the integrated model was applied to the data from the cases when the pairs had access to the shared visual information, it correctly resolved 72.4% of the referring expressions. This was significantly better than the 49.3% exhibited by the language-only model ($\chi^2_{(1, N=217)} = 26.8$, $p < .01$). Similar to the last comparison, Table 4 reveals that both the integrated and language-only models correctly resolved pronouns that the other model did not. In this comparison, there appeared to be substantially more cases (33.9%) that the integrated model exclusively identified versus those that the language-only model did (10.6%). The differences between these two models were similar to those discussed above in comparing the performance of the visual-only model with the language-only model. However, in this case, the integrated model could resort to the linguistic-salience list when the shared workspace was inactive, and therefore benefit from the ranking of entities based on linguistic-salience.

		Language	
		Incorrect	*Correct*
Inte-grated	*Incorrect*	37 (17.0%)	23 (10.6%)
	Correct	74 (33.9%)	84 (38.5%)

Table 4. Confusion matrix between the Language Model and the Integrated Model.

Finally, the integrated model's 72.4% performance was significantly better than the visual-only model's 62.2% on the same data ($\chi^2_{(1, N=217)} = 17.29$, $p < .01$); indicating a major performance benefit to having an integrated model. It is interesting to note in Table 5 that the integrated model nearly dominates the visual-only model. There are only three instances where the visual-only model correctly resolves a referent that the integrated model did not. All three of these instances were cases where a longer visual decay parameter would have captured the proper referent. However, a longer decay could harm the performance of the integrated model by inhibiting a switch to the linguistic salience list.

		Visual	
		Incorrect	*Correct*
Inte-grated	*Incorrect*	57 (26.3%)	3 (1.4%)
	Correct	25 (11.5%)	132 (60.8%)

Table 5. Confusion matrix between the Visual Model and the Integrated Model.

Model Performance by Language Type

Finally, a detailed examination of the form of referring expressions successfully resolved differed across the model types. In other words, there was a significant Model Type × Pronoun Type interaction in the model, depicted in Figure 3 (for the interaction, $G^2_{(4)} = 17.43$, $p = .001$). An examination of this interaction reveals that the language-only model appears to perform best when resolving personal pronouns and decreases in success when resolving demonstrative pronouns, while the opposite trend is seen in both the visual-only and integrated models. This revelation reveals some interesting patterns regarding the appropriateness of the various models and suggests that future lines of work might explore strategic shifts in the use of the visual-salience or linguistic-salience lists triggered by the syntactic information in the utterance.

To summarize, the language-only model performed reasonably well on the dialogues in which the pairs had no access to shared visual information. However, when the same model was applied to the dialogues collected from task conditions where the pairs had access to shared visual information, the performance of the language-only model was significantly reduced. However, both the visual-only

model and the integrated model showed significantly increased performance over the language-only model; and the integrated model was the top performer overall.

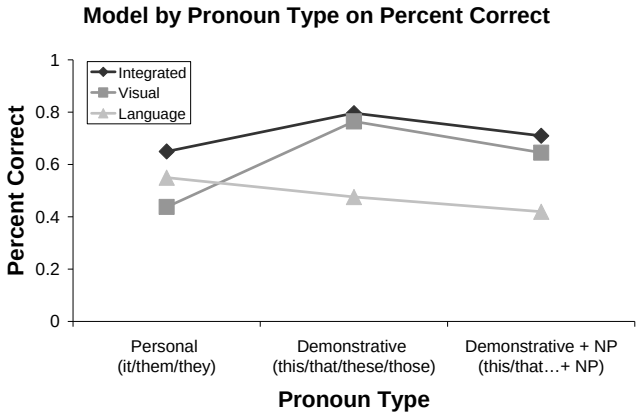

Figure 3. Effect of Model Type and Pronoun Type on successful pronoun resolution.

Error analysis

In order to inform further development of the model, a number of failure cases were examined in detail, particularly those in which all of the models failed. The first thing to note was that a number of the pronouns used by the pairs referred to larger visible structures in the workspace. An example of this was when the Worker would state, "like this?", and ask the Helper to comment on the overall configuration of the puzzle. In the current model, only the puzzle pieces are included as possible visual referents. One approach to alleviating this error is to integrate a richer notion of semantics with the additional visual entities in order to accurately model such situations (e.g., see [6]).

Another area where the models suffered performance problems was during references to higher-order referents such as general events or the state of the world. For example, "OK, this is going to be tough" where "this" specifies the general construction of the puzzle. Similarly, non-referential "its" as in "It is easy to make something work" posed a problem for the models. These are both common problems in reference resolution and may be addressed in the future by applying recent advances in these areas. Work by Müller [37] provides an automated method for filtering out non-referential "its," and this technique could be applied to refine the pronouns attempted by applying a filter earlier on in the processing pipeline.

In addition, there were several errors that resulted from chaining errors: When the initial referent was misidentified all subsequent chains of referents were incorrect. The approach used in this study to score the success of the resolved pronouns followed Walker's original description [51] where all referents are scored as incorrect if the original binding is incorrect. This makes sense from a systems perspective where incorrect inferences could be made if the initial referent is incorrect. However, recent evaluations have used a more lenient formulation whereby a "location-based" evaluation procedure is used [49]. These studies only look one step back and do not penalize for longer "error chains".

Finally, the visual-only model and the integrated model had a tendency to suffer from timing issues. For instance, the pairs occasionally introduced a new visual entity with, "this one?" However, the piece did not appear in the workspace until a short time after the utterance was made. In such cases, the object was not available as a referent on the object list. The implementation presented here followed the notion that actions typically precede the associated keywords or language [43]. Future work could include a richer model of gestures and spoken language alignment in order to successfully account for such issues (e.g., [18]).

DISCUSSION

The results of this experiment find that the language-only model performs in the range of previous studies of pronoun resolution on spoken discourse by successfully resolving approximately 67% of the pronouns encountered[3]. This apparent success is due, in part, to the fact that the approach captures the many well-known syntactic and psycholinguistic factors that contribute to entity salience. However, when the language-only model is applied to the portions of the corpus in which the pairs had access to shared visual information, its performance suffers. In fact, the application of the language-only model to the trials undertaken with shared visual information performs below 50%. One reason for this is that when shared visual information is available, action and language use can become interchangeable [21]; this is highlighted by the fact that the visual-only model performs at 62% and is better in many instances.

Overall, the integrated model is the best performer in this evaluation. Its performance is equivalent to the language-only model during trials without shared visual information available, since it falls back to using linguistic salience as a source for resolution. However, when applied to the cases where shared visual information is available, the integrated model performs significantly better than the language-only model. This is due, in part, to the fact that it captures linguistic references to physical actions.

A comparison of the integrated model to the visual-only model yields interesting results. The first is that the integrated model resolves reference when no shared visual information is available, while the visual-only model does not. Second, when shared visual information is available, the integrated model outperforms the visual-only model, and this difference exists regardless of the lexical complexity of the puzzle pieces. The integrated model captures elements of the discourse that are neglected by the visual-only model, particularly when there is a prolonged discussion about the

[3] Prior literature finds resolution rates of approximately 65% for similar evaluations using task-oriented spoken dialogues.

features of a given object. As a result, the decay parameter allows the model to shift its focus from active visual events to the conversation currently taking place. In a sense, this mimics the shift in attention that occurs between participants as they fluidly move between referring to objects and actions in the environment to those discourse entities produced in the spoken dialogue stream. Together these findings provide strong support for the need to have an integrated model of reference. Indeed, both linguistic entities and visual entities are central to accurate and perspicuous accounting of referring behaviors.

Throughout this paper, we focus on developing a computational understanding that can be applied to systems supporting collaborative physical tasks (e.g., [39, 44]) and collocated physical interactions (e.g., [46]). While at first glance these environments may appear limited, they are often rife with cross-modal references and complex linguistic behaviors that current models do not capture. To further our understanding of these patterns, we perform an evaluation using a limited amount of world knowledge. We do this for two major reasons. The first is to maintain a direct comparison to prior models of reference resolution that perform evaluations without a world knowledge component [48, 50]. The second is to control for the potentially conflating influence world knowledge could have in different visual environments. For example, if a rich notion of semantics is applied and its use varies across experimental conditions, then it is no longer clear whether the benefits derive from the visual salience component of the model or the semantic restrictions enforced by a world knowledge component. Clearly, extension of our model to richer task domains requires further development of our world knowledge component, and may also require significant research into scene analysis, object tracking, and the integration of richer task models.

FUTURE WORK

In the future, we plan to extend this work in several ways. First, a fully-automated version of the models is currently under development. This constitutes a fully automated parsing and resolution system that can then be applied to a range of new tasks with a variety of parameters. This will allow us to assess the generalizability of the model. A second area is to develop studies that expand our notion of collaborative visual salience. For example, objects may become activated multiple times in a short window of time, or be more or less salient depending on nearby actions. Future work will explore these parameters in detail. Finally, we plan to appreciably enhance the integrated model. It appears from both the initial data analysis and a qualitative examination of the model performance that the pairs make tradeoffs between reliance on the linguistic and visual context. Yet, our current understanding could be enhanced by taking a more theoretically informed approach to integrating the information from multiple streams.

ACKNOWLEDGEMENTS

This research was funded in part by NSF grants #99-80013, #02-08903, and by an IBM PhD Fellowship to the first author. We would like to thank Susan Fussell, Susan Brennan, Justine Cassell, Joel Tetreault, Donna Byron, Daniel Avrahami, and Aaron Bauer for their comments at various stages of this work.

REFERENCES

[1] Allen, J., Ferguson, G., Swift, M., Stent, A., Stoness, S., Galescu, L., Chambers, N., Campana, E., and Aist, G. (2005). Two diverse systems built using generic components for spoken dialogue. In *Proceedings of ACL '05, Comp. Vol.*, 85-88.

[2] Bangalore, S., and Johnston, M. (2004). Balancing data-driven and rule-based approaches in the context of a multimodal conversational system. In *Proceedings of HLT-NAACL '04*, 33-40.

[3] Brennan, S.E. (1995). Centering attention in discourse. *Language & Cognitive Processes, 10 (2)*, 137-167.

[4] Brennan, S.E. (2005). How conversation is shaped by visual and spoken evidence. In Trueswell, J., and Tanenhaus, M. (Eds.) *Approaches to studying world situated language use: Bridging the language-as-product and language-as-action traditions*, pp. 95-130. MIT Press, Cambridge, MA.

[5] Brennan, S.E., Friedman, M.W., and Pollard, C.J. (1987). A centering approach to pronouns. In *Proceedings of the ACL '87*, 155-162.

[6] Byron, D.K., Mampilly, T., Sharma, V., and Xu, T. (2005). Utilizing visual attention for cross-modal coreference interpretation. In *Proceedings of CONTEXT '05,* 83-96.

[7] Byron, D.K., and Stoia, L. (2005). An analysis of proximity markers in collaborative dialog. Appeared at *41st annual meeting of the Chicago Linguistic Society*.

[8] Cassell, J. (2004). Towards a Model of Technology and Literacy Development: Story Listening Systems. *Journal of Applied Developmental Psychology, 25 (1)*, 75-105.

[9] Cassell, J., and Stone, M. (2000). Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of International Natural Language Generation Conference*, 171-178.

[10] Chai, J.Y., Prasov, Z., Blaim, J., and Jin, R. (2005). Linguistic theories in efficient multimodal reference resolution: An empirical investigation. In *Proceedings of IUI '05*, 43-50.

[11] Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. MIT Press, Cambridge, MA.

[12] Clark, H.H., and Brennan, S.E. (1991). Grounding in communication. In Resnick, L., Levine, J., and Teasley, S. (Eds.) *Perspectives on socially shared cognition*, pp. 127-149. APA, Washington DC.

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

[14] Daelemans, W., Buchholz, S., and Veenstra, J. (1999). Memory-based shallow parsing. In *Proceedings of CoNLL Workshop*, 53-60.

[15] Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2001). TiMBL: Tilburg memory based learner,

version 4.0, reference guide. *Technical Report ILK Technical Report 00-01*, Tilburg University.

[16] Devault, D., Kariaeva, N., Kothari, A., Oved, I., and Stone, M. (2005). An information-state approach to collaborative reference. In *Proceedings of ACL '05, Comp. Vol,* 1-4.

[17] Donmez, P., Rosé, C.P., Stegmann, K., Weinberger, A., and Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In *Proceedings of CSCL '05*, 125-134.

[18] Eisenstein, J., and Christoudias, C.M. (2004). A salience-based approach to gesture-speech alignment. In *Proceedings of HLT-NAACL '04*, 25-32.

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

[20] Gergle, D. (2006). *The Value of Shared Visual Information for Task-Oriented Collaboration*. Unpublished Ph.D., Carnegie Mellon University.

[21] Gergle, D., Kraut, R.E., and Fussell, S.R. (2004). Action as language in a shared visual space. In *Proceedings of CSCW '04*, 487-496.

[22] Gergle, D., Kraut, R.E., and Fussell, S.R. (2006). The Impact of Delayed Visual Feedback on Collaborative Performance. In *Proceedings of CHI '06*, 1303-1312.

[23] Grosz, B.J., Joshi, A.K., and Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of ACL '83*, 44-50.

[24] Grosz, B.J., Joshi, A.K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21 (2)*, 203-225.

[25] Grosz, B.J., and Sidner, C.L. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics, 12 (3)*, 175-204.

[26] Gundel, J.K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69 (2)*, 274-307.

[27] Gweon, G., Rosé, C.P., Zaiss, Z., and Carey, R. (2006). Providing support for adaptive scripting in an on-line collaborative learning environment. In *Proceedings of CHI '06*, 251-260.

[28] Heeman, P.A., and Allen, J. (1995). Dialogue transcription tools. *Tech. Report, Trains TN 94-1*, Univ. of Rochester.

[29] Hobbs, J.R. (1978). Resolving pronoun references. *Lingua, 44*, 311-338.

[30] Hudson, S.B., Tanenhaus, M.K., and Dell, G.S. (1986). The effect of the discourse center on the local coherence of a discourse. In *Proceedings of the Cognitive Science Society '86*, 96-101. Lawrence Erlbaum Associates.

[31] Huls, C., Bos, E., and Claassen, W. (1995). Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics, 21 (1)*, 59-79.

[32] Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI 2000*, 685-689.

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

[34] Kraut, R.E., Gergle, D., and 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.

[35] Kumar, R., Rosé, C.P., Aleven, V., Iglesias, A., and Robinson, A. (2006). Evaluating the effectiveness of tutorial dialogue instruction in an exploratory learning context. In *Proceedings of the Intelligent Tutoring Systems Conference*.

[36] Mitkov, R. (2000). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of DAARC 2000*, 96-107.

[37] Müller, C. (2006). Automatic detection of non-referential it in spoken multi-party dialog. In *Proceedings of EACL 2006*, 49-56.

[38] Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., and Sclabassi, R.T. (1993). Turning away from talking heads: The use of video-as-data in neurosurgery. In *Proceedings of INTERCHI '93*, 327-334.

[39] Ou, J., Oh, L.M., Fussell, S.R., Blum, T., and Yang, J. (2005). Analyzing and predicting focus of attention in remote collaborative tasks. In *Proceedings of ICMI '05*, 116-123.

[40] Oviatt, S., Levow, G.-A., Moreton, E., and MacEachern, M. (1998). Modeling global and focal hyperarticulation during human-computer error resolution. *Journal of the Acoustical Society of America, 104 (5)*, 3080-3091.

[41] Oviatt, S., MacEachern, M., and Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24 (2)*, 87-110.

[42] Oviatt, S.L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction, 12*, 93-129.

[43] Oviatt, S.L., DeAngeli, A., and Kuhn, K. (1997). Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proceedings of CHI '97*, 415-422.

[44] Ranjan, A., Birnholtz, J., and Balakrishnan, R. (2006). An exploratory analysis of partner action and camera control in a video-mediated collaborative task. In *Proceedings of CSCW '06*, 403-412.

[45] Scholl, B.J. (2001). Objects and attention: the state of the art. *Cognition, 80*, 1-46.

[46] Scott, S.D., and Carpendale, S. (2006). Guest editors' introduction: Interacting with digital tabletops. *IEEE Computer Graphics & Applications, 26 (5)*, 24-27.

[47] Strube, M. (1998). Never look back: An alternative to centering. In *Proceedings of ACL '98*, 1251-1257.

[48] Strube, M., and Hahn, U. (1999). Functional centering: Grounding referential coherence in information structure. *Computational Linguistics, 25 (3)*, 309-344.

[49] Tetreault, J.R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics, 27 (4)*, 507-520.

[50] Tetreault, J.R. (2005).*Empirical evaluations of pronoun resolution*. Unpublished Ph.D., University of Rochester.

[51] Walker, M.A. (1989). Evaluating discourse processing algorithms. In *Proceedings of ACL '89*, 251-261.

[52] Whittaker, S. (2003). Things to talk about when talking about things. *Human-Computer Interaction, 18 (1-2)*, 149-170.