Awareness of Partner's Eye Gaze in Situated Referential Grounding: An Empirical Study

Changsong Liu
Department of Computer
Science and Engineering
Michigan State University
East Lansing, MI 48824
cliu@msu.edu

Dianna L. Kay
Department of Computer
Science and Engineering
Michigan State University
East Lansing, MI 48824
dnnkay855@gmail.com

Joyce Y. Chai
Department of Computer
Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

ABSTRACT

In situated dialogue, although artificial agents and their human partners are co-present in a shared environment, their representations of the environment are significantly different. When a shared basis is missing, referential grounding between partners becomes more challenging. Our hypothesis is that in such a situation, non-verbal modalities such as eye gaze play an important role in coordinating the referential process. To validate this hypothesis, we designed an experiment to simulate different representations of the shared environment. Our studies have shown that, when one partner pays attention to the other partner's naturally occurred eye gaze during interaction, referential grounding becomes more efficient. However this improvement is more significant under the situation where partners have matched representations of the shared environment compared to the situation with mismatched representations. This paper describes our experimental findings and discusses their potential implications.

Author Keywords
Situated Dialogue, Eye Gaze, Referential Grounding

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous

General Terms
Experimentation, Human Factors

INTRODUCTION

With recent advances in AI techniques, there is an increasing demand for artificial agents (e.g. robots) that can collaborate with humans in a shared environment. Examples of such applications include space exploration [7], military training [12], and autism therapy [5]. To develop this type collaborative agents, a crucial issue is to address the mismatched abilities between artificial agents and their human partners. Although artificial agents and their human partners are co-present in a shared environment, their perception, representation and knowledge of the shared environment could be significantly different. When a shared basis of the environment is missing, communication between partners becomes more challenging [4]. Language alone may be inefficient and other extralinguistic information will need to be pursued. In this paper, we investigate one type of non-verbal modalities – human eye gaze during speech communication.

Eye gaze serves many functions in mediating interaction [1, 4] and managing turn taking [17] and grounding [16]. Previous psycholinguistic findings have shown that eye gaze is tightly linked with language production and comprehension [10, 21, 15, 8]. Eye gaze has also been shown efficient for providing early disambiguating cues in referential communication [9], for intention recognition during object manipulation [2], and for attention prediction [6]. Specifically, related to the referential process, recent work has incorporated eye gaze in resolving exophoric referring expressions [18, 19].

Motivated by previous work, our hypothesis is that eye gaze plays an important role in referential grounding, especially between partners with mismatched representations of the shared environment. More specially, we are interested in the following questions:

1. *How difficult is it for partners with mismatched representations of the shared environment to collaborate?* When a shared basis of the environment is missing, partners may not be able to communicate as they normally do. We are interested in how the mismatched representations may impact collaboration, conversation, and automated language processing.

2. *To what extent does the collaboration benefit from the awareness of partner's eye gaze? Is such awareness more helpful for partners with mismatched representations?* Our hypothesis is that partners with mismatched representations could benefit more from gaze information. This is because on one hand verbal communication could be more difficult, and on the other hand gaze may allow many joint actions to be done non-verbally [3].

To validate this hypothesis and address the above questions, we designed an experiment where a director and a matcher

2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction.
February 13, 2011, Palo Alto, California, USA.

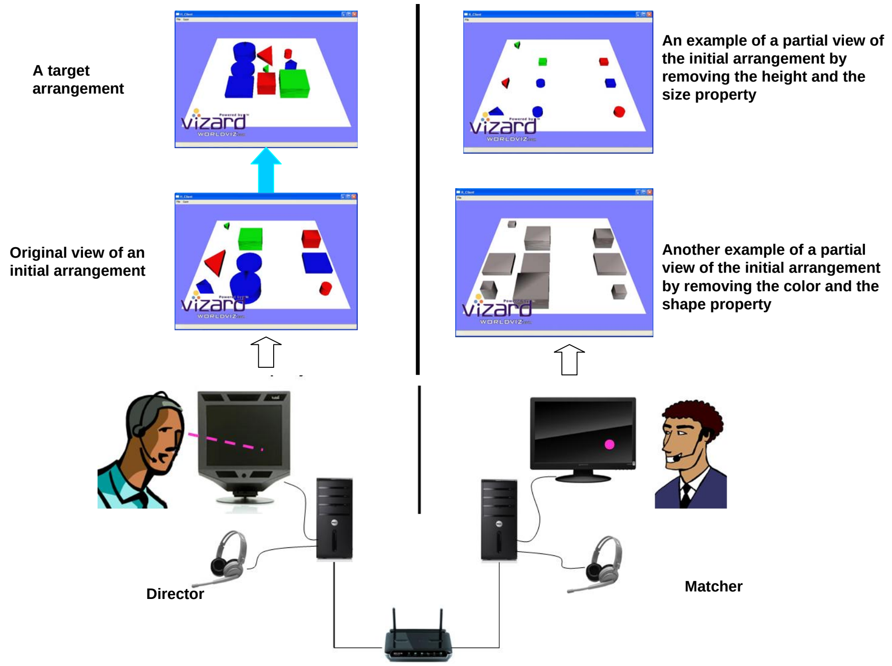
A target arrangement
Original view of an initial arrangement
An example of a partial view of the initial arrangement by removing the height and the size property
Another example of a partial view of the initial arrangement by removing the color and the shape property
Director
Matcher

Figure 1. Architecture of our collaborative system. Two partners in the same room are separated by a divider. The director is seated in front of a display-mounted Tobii 1750 eye tracker (Tobii Technology) and the matcher in front of a regular computer. Two computers are connected and synchronized via an Ethernet hub. The director's eye gaze positions are captured and can be displayed as gaze cursors (a 32 by 32 pixels pink dot) superimposed over the matcher's display based on experimental conditions. A bi-directional microphone-speaker system was used as the speech channel for two partners to verbally communicate.

collaboratively solve a block game. The director has a complete view of the shared environment. By controlling what the matcher "sees" from the environment, we are able to simulate different representations of the shared environment between the director and the matcher (i.e., matched representations or mismatched representations). In addition, during the interaction, we keep track of the director's eye gaze and alter the conditions based on whether the matcher is aware of the director's eye gaze by controlling whether the director's eye gaze will be made available on the matcher's screen. These settings allow us to investigate the role of eye gaze in different experimental conditions.

Our results indicate that human partners can overcome the mismatched representations by switching their communicative strategies. The matcher's awareness of the director's eye gaze during interaction improves referential grounding. However, such an improvement is more significant under the matched representations compared to mismatched representations. This finding is somewhat surprising given our original hypothesis. This is possibly due to the more complex gaze pattern that interacts with the spatial information occurred in utterances, which opens up new interesting questions for future research.

METHOD

Apparatus

An architecture of our collaborative system is shown in Figure 1. Two partners (a director and a matcher) collaborate on a block arranging task. The director instructs the matcher to arrange a set of blocks according to a given configuration (i.e., a target configuration) which is only available to the director. They both face the same virtual environment, however, it can be displayed differently on their respective screens to simulate different internal representations of the environment. The director's eye gaze positions are captured by a TOBII display-mounted eye tracker. Depending on experimental conditions, the director's naturally occurred gaze will be made available to the matcher (shown as gaze cursors on the matcher's screen) real time during interaction.

Experiment Design

A pair of participants collaborate to arrange a set of blocks in a given order. In each task, there are seven blocks with randomly assigned color (red, green or blue), shape (triangle, circle or pentagon), size (small, medium or large) and height (short, medium or tall). At the beginning of the task, seven blocks are randomly placed on a plane (Figure 2a). The director is given a target arrangement of the blocks (Figure 2b). The director can view the final arrangement at any time by pressing the 'space' key. The director can not move the blocks by himself. He has to instruct the matcher to pick up one block and move it to the desired location, and then continue with the next one till all the blocks are in the right positions. They also have to follow a given order when they are moving the blocks one by one. The block that should be moved at the current step is indicated by an arrow pointing to it, which can only be seen by the director.

The director and the matcher both face the same environ-

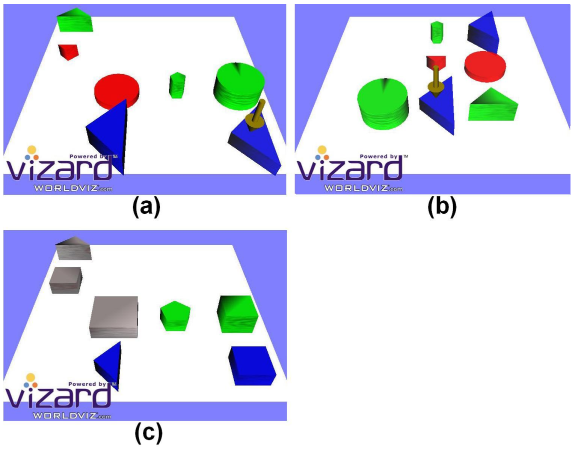

Figure 2. Example of different views in our experiment. (a) shows a random initial arrangement of the blocks on the director's screen at the beginning of the task. The block that should be moved at the current step is indicated by an arrow pointing to it. (b) shows the target arrangement of the blocks. (c) shows the corresponding partial view displayed on the matcher's screen to simulate mismatched representations between the director and the matcher.

ment, but may have different representations of this environment. The blocks can be displayed identically or differently on their respective screens, depending on the experimental conditions based on the following dimension:

- Whether the director and the matcher has matched (*view*+) or mismatched (*view*–) representations of the shared environment. Under the mismatched condition, the display on the matcher's screen is different from the director's in the following way: each attribute (color, shape, size or height) of a block has a 50% chance to be set to an 'unknown' default status (i.e., color is set to gray, shape is set to square, size is set to medium and height is set to medium). Figure 2(c) shows a mismatched-view display of the same environment as in Figure 2(a).
- Whether the matcher is aware of the director's gaze direction during interaction. This is controlled by whether or not displaying the director's naturally occurred eye gaze as cursors on the matcher's display. This results in two levels: the matcher is aware of the director's eye gaze (*gaze*+) and the matcher is not aware of the eye gaze (*gaze*-)

Based on the above two dimensions, we have a total of four experimental conditions:

- (*view*+,*gaze*+): matched-view with gaze awareness
- (*view*–,*gaze*+): mismatched-view with gaze awareness
- (*view*+,*gaze*–): matched-view without gaze awareness
- (*view*–,*gaze*–): mismatched-view without gaze awareness

Participants and procedure

Sixteen (eight pairs) undergraduate/graduate students from Michigan State University were recruited to participate in our studies. In each pair of participants, one played the role of the director and the other played the role of the matcher throughout the entire experiment. Each pair first had two practice trials. The first practice trial was the (*view*+,*gaze*+) condition and the second was (*view*–,*gaze*–). After the first two practice trials, all our participants had no problem understanding the task and familiarizing themselves with gaze cursors. They then proceeded with four actual trials, each of which was based on one of the four conditions, in a completely random order. The experiment lasted approximately 40 minutes.

RESULTS AND DISCUSSION

During the experiment, each trial was logged by keeping track of both the director's and the matcher's displays and speech communication. Two kinds of information were extracted from the logged data to facilitate our investigation: the total time spent to finish each trial and the total number of utterances issued by both partners in each trial. Table 1 shows the time (seconds) spent to finish each trial. Table 2 shows the number of utterances in each trial.

The Role of Mismatched Representations

We compare the means of time and number of utterances between different conditions. Surprisingly, we did not find significant difference between (*view*–,*gaze*–) and (*view*+,*gaze*–) (for time, $t = -0.25, p < 0.594$; for number of utterances, $t = -0.46, p < 0.671$). The result implies that collaboration based on mismatched representations of the environment may not be more difficult. A further investigation of our data indicate that partners employed different communicative strategies in the (*view*–,*gaze*–) condition compared to the (*view*+,*gaze*–) condition. There were basically two kinds of strategies for collaboratively referring to the intended block: one is to describe its object-based properties (color, shape, size and height), and the other is to describe its spatial information locally (with respect to another block(s) close by) or globally (with respect to the environment). Sometimes the two strategies can also be used together. Here are examples of the two strategies from our data:

> *Object-based properties*: the red pentagon / the big blue one / the tallest one
> *Spatial information*: the object underneath the one we just moved / the middle one of the three objects on the top / the object in the right bottom corner
> *Combined*: the green one to the left of that one / the triangular-looking one on the right side of this environment

It is natural to use object-based properties with matched views, whereas using spatial information is a better strategy for mismatched views. In our experiment, participants had no problem switching to and relying on the spatial information when they detected they might have mismatched representations. Sometimes they even tried to explicitly query or request their partners about the condition or strategy at the beginning of a trial, e.g., "*Can you see color this time?*", "*I only have gray squares, so you would better always describe the spa-*

	Pair of participants								
Condition	1	2	3	4	5	6	7	8	$\bar{y}_{i.}$
view+,gaze–	126.2	158.6	129.8	208.2	156.0	70.1	188.7	160.4	149.8
view–,gaze–	71.4	107.8	217.0	150.2	169.2	103.2	257.2	77.0	144.1
view+,gaze+	72.8	134.8	155.3	155.8	97.1	94.0	120.3	53.6	111.5
view–,gaze+	86.7	129.9	135.7	210.2	84.9	83.6	146.8	78.9	119.6
$\bar{y}_{.j}$	89.3	132.8	159.5	181.1	126.8	87.7	178.3	92.5	$\bar{y}_{..} = 131.0$

Table 1. Time (seconds) spent to finish each trial. The last row of the table is the average time of each pair of participants ($\bar{y}_{.j}$). The last column of the table is the average time of each experimental condition ($\bar{y}_{i.}$).

	Pair of participants								
Condition	1	2	3	4	5	6	7	8	$\bar{y}_{i.}$
view+,gaze–	49	45	45	99	121	23	52	65	62
view–,gaze–	24	44	90	53	103	30	84	29	57
view+,gaze+	17	41	51	43	72	33	27	19	38
view–,gaze+	40	33	48	98	66	24	38	30	47
$\bar{y}_{.j}$	33	41	59	73	91	28	50	36	$\bar{y}_{..} = 51$

Table 2. Number of utterances in each trial. The last row of the table is the average number of utterances of each pair of participants ($\bar{y}_{.j}$). The last column of the table is the average number of utterances of each experimental condition ($\bar{y}_{i.}$).

tial location". The spatial information based strategy, although shown to be less preferable when object-based properties were available [13], appears equally efficient compared to the object-based strategy in our experiment. Actually, it provides an easier and more reliable way for collaborative referring when the shared basis was missing as in the mismatched-view case. Therefore, all our participants spontaneously relied on spatial information and accomplished the task with little trouble under the mismatched-view condition.

The Role of Gaze Awareness

The other aspect we are interested in is the role of gaze awareness in this collaborative task. As we expected, when the two partners' views were matched, allowing the matcher to see the director's gaze positions significantly reduced the time and the number of utterances needed to accomplish the task. This is demonstrated by the comparison between (*view+,gaze+*) and (*view+,gaze–*). Under the (*view+,gaze+*) condition, participants spent 39.3 seconds shorter time ($t = 2.43, p < .05$) and issued 24.5 fewer utterances ($t = 2.69, p < .05$) compared to the (*view+,gaze–*) condition. This result is in accordance with previous findings (e.g. [11, 1]) that shared gaze facilitates the communication of task-relevant spatial information.

Is gaze more helpful when two partners' representations of the environment are mismatched? Our results have shown some interesting observations. Although the (*view–,gaze+*) condition on average took 24.5 seconds shorter time and 10 fewer utterances than the (*view–,gaze–*), the differences are only marginally significant (for interaction time, $t = 1.14, p < 0.146$; for number of utterances, $t = 0.9, p < 0.199$). Why does gaze appear to be less helpful under the mismatched-view condition compared to the matched-view condition? To answer this question, more in-depth analysis has to be done to investigate how the gaze patterns are different under two conditions and how the difference may affect the partner's acceptance of gaze information. This is left for our future research. Here we only present one possible explanation based on our intuition.

We hypothesize that gaze patterns under the mismatched condition can be different from the patterns under matched condition. In the matched-view case, since the speaker's strategy is mainly based on describing the referent's own properties, his/her gaze should mostly fixed on the object that is being described at the moment. In other words, the object that draws most of the speaker's attention is very likely the current referent, and thus allows the listener to directly use this cue to infer the intended referent. However, such a pattern may be weakened in the mismatched-view case as the speaker switches to the strategy based on spatial information. Using spatial language often involves complex mental computations [14], which can possibly interact with gaze patterns. For example, instead of steadily fixating on the object that is being referred to, the speaker may need to look back-and-forth between two objects while he is selecting a proper spatial term to describe the relation. Also, the gaze could be circling among a group of objects if the speaker intends to use a group-based description (e.g., "*the middle object*", "*the third one from left to right*"). The gaze may not fixate on any object but rather scan through the environment if the speaker intends to describe a global direction (e.g., "*the northwest one*", "*the one that is to the right bottom corner*"). All these possible gaze behaviors may be closely related to the internal computations of spatial language, and thus may not precisely indicate the intended referent.

CONCLUSION

This paper investigate the role of eye gaze in the referential process in situated dialogue. Our findings indicate that, when the partners' representations of the shared environment are matched, eye gaze and awareness of eye gaze facilitate the communication and significantly reduces the time and verbal communication needed for grounding references. But when the partners have mismatched representations of the shared environment, the effect of awareness of eye gaze appears less significant. This is possibly due to more complex patterns of gaze behaviors in production of utterances involving complex spatial information. However, more in-depth investigation has to be done to have a clear understanding.

Our experiments further indicate that, participants spontaneously employed communication strategies based on spatial information to describe objects of interest when the shared representation was missing. By using spatial information, participants grounded references as efficiently as applying strategies based on object properties. This finding again (see [20] for another example in human robot interaction) has revealed the importance of spatial language, especially when the artificial agents and their human partners have mismatched perceptions and representations of the shared environment. Therefore, spatial language understanding and spatial information based dialogue management serve as key components to enabling collaborative and situatd interaction between humans and artificial agents.

This paper only describes some initial results of our investigation. To have a more clear understanding of the role of eye gaze in this collaborative referential process, we also need to capture the gaze behaviors from the matcher and examine how that may affect the director's behaviors and thus the overall discourse of interaction. In addition, the interactions between the gaze from both partners will be even more interesting. Our future work will explore these directions.

ACKNOWLEDGMENT

This work was supported by CNS-0957039 and IIS-1050004 from the National Science Foundation.

REFERENCES

1. M. Argyle, M. Cook, and M. Argyle. *Gaze and mutual gaze*. Cambridge University Press Cambridge, 1976.
2. T. Bader, M. Vogelgesang, and E. Klaus. Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 199–206. ACM, 2009.
3. S. Brennan, X. Chen, C. Dickinson, M. Neider, and G. Zelinsky. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477, 2008.
4. H. H. Clark. *Using language*. Cambridge University Press, 1996.
5. K. Dautenhahn and I. Werry. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, 12(1):1–35, 2004.
6. R. Fang, J. Chai, and F. Ferreira. Between linguistic attention and gaze fixations in multimodal conversational interfaces. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 143–150. ACM, 2009.
7. T. Fong and I. Nourbakhsh. Interaction challenges in human-robot space exploration. *interactions*, 12(2):42–45, 2005.
8. Z. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11(4):274, 2000.
9. J. Hanna and S. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615, 2007.
10. M. Just and P. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480, 1975.
11. A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1):22, 1967.
12. P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, and D. Piepol. Building interactive virtual humans for training environments. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, volume 2007. NTSA, 2007.
13. S. Kriz, J. Trafton, and J. McCurry. The Role of Spatial Information in Referential Communication: Speaker and Addressee Preferences for Disambiguating Objects. In *D. S. McNamara & J. G. Trafton (Eds.), Proceedings of the 29th Annual Cognitive Science Society Austin, TX: Cognitive Science Society*, 2007.
14. C. Liu, J. Walker, and J. Chai. Ambiguities in Spatial Language Understanding in Situated Human Robot Dialogue. In *2010 Fall Symposium on Dialogue with Robots*, 2010.
15. A. Meyer, A. Sleiderink, and W. Levelt. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2):B25–B33, 1998.
16. Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 553–561. Association for Computational Linguistics, 2003.
17. D. Novick, B. Hansen, and K. Ward. Coordinating turn-taking with gaze. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1888–1891. IEEE, 2002.
18. Z. Prasov and J. Chai. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29. ACM, 2008.
19. Z. Prasov and J. Y. Chai. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 471–481, October 2010.

20. M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock. Spatial language for human-robot dialogs. *IEEE transactions on systems, man and cybernetics*, 34:154–167, 2004.

21. M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632, 1995.