Exploring the Use of Gesture in Collaborative Tasks

Isaac Wang
University of Florida
Gainesville, FL 32611, USA
wangi@ufl.edu

Pradyumna Narayana
Dhruva Patil
Gururaj Mulay
Rahul Bangar
Colorado State University
Fort Collins, CO 80523
prady@rams.colostate.edu
dkpatil@cs.colostate.edu
guru5@colostate.edu
rahul.bangar@colostate.edu

Bruce Draper
Ross Beveridge
Colorado State University
Fort Collins, CO 80523
draper@cs.colostate.edu
ross@cs.colostate.edu

Jaime Ruiz
University of Florida
Gainesville, FL 32611, USA
jaime.ruiz@ufl.edu

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3053239

Abstract
Personal assistants such as Siri have changed the way people interact with computers by introducing virtual assistants that collaborate with humans through natural speech-based interfaces. However, relying on speech alone as the medium of communication can be a limitation; non-verbal aspects of communication also play a vital role in natural human discourse. Thus, it is necessary to identify the use of gesture and other non-verbal aspects in order to apply them towards the development of computer systems. We conducted an exploratory study to identify how humans use gesture and speech to communicate when solving collaborative tasks. We highlight differences in gesturing strategies in the presence/absence of speech and also show that the inclusion of gesture with speech resulted in faster task completion times than with speech alone. Based on these results, we present implications for the design of gestural and multimodal interactions.

Author Keywords
Gesture interaction; communication

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

Introduction

Computers are becoming more like collaborative agents rather than computational tools. Dialogue-based systems such as Apple's Siri [21] and Amazon's Alexa [22] have changed the way we interact with computers by replacing the traditional keyboard and mouse with speech-driven interfaces. With the advancement of natural language processing, any user can now talk with these systems as they would with another human being, making interactions more natural.

However, the sole use of speech and audio for the medium of communication may be considered a limitation. Research has shown the usefulness of nonverbal aspects of communication (e.g., gestures, expressions, etc.), specifically for establishing common ground [3,4,8,12] as well as for conveying both complementary and redundant information [5]. Computers must utilize these additional channels of communication if they are truly to be more human-like in their interaction.

As a starting point, we first need to identify and model how gestures and speech are used in human to human communication. We focus on observing the natural interaction between two people when working together in a constrained, simplified task (specifically a physical construction task). In this abstract, we highlight results from an exploratory study comparing communication in the presence and absence of speech or gesture.

Related Work

Integrating the use of both speech and gesture in computer understanding has been an ongoing challenge since the introduction of the point and manipulate "Put-that-there" system by Bolt [1]. A large amount of work has focused on designing and developing gesture-based interaction systems ranging from mobile systems [2,9,10] to large-screen displays [14,19,20], as well as researching the relationships used in multimodal and gestural interaction [6–8,12,15–18].

In particular, Epps et al. [5] conducted a study investigating multimodal interactions for photo manipulation tasks. They identified gestures that provided complementary information (such as deictics used for specification) and gestures that provided redundant information (e.g., a rotational gesture used simultaneously with verbally describing the rotation).

Furthermore, Quek et al. [17] elaborate on the multimodality of human discourse, highlighting the temporal relationship of gestures to speech and describing specific instances of gestures that occur alongside spoken utterances. Others establish the usefulness of shared visual information in grounding, or establishing mutual understanding when communicating [3,4,6–8,12]. Veinott et al. [18] further raise the implication that the inclusion of video provides additional information, including gesture, which may be utilized to a higher degree in the presence of language barriers.

Method

The purpose of this study was to analyze the natural dyadic communication used by two people when engaging in solving a collaborative task.

Task and Apparatus

We asked pairs of participants to collaboratively build different pre-determined structures using wooden blocks. Participants were put in separate rooms with

Figure 1: Frames from two video feeds showing the view of the signaler (top) and actor (bottom).

similar setups. Each participant stood in front of a table facing a TV screen on the opposite end of the table. Microsoft Kinect v2 sensors were also set up on the opposite end, facing the participant (Figure 2). We developed software to stream live video (and audio) from the Kinect sensors between the two setups so that participants could communicate with each other as if they were facing each other at opposite ends of the same table (Figure 1). The Kinect sensors were also used to record the experiment, providing us with RGB video, depth data, and motion capture skeletons.

One participant was given the role of *actor* and was provided with a set of 12 wood cubes (with 4-inch sides). The other participant was given the role of *signaler* and was given an image of an arrangement, or layout, of these blocks. The signaler was assigned the task of communicating to and directing the actor to replicate the layout; the actor needed to respond to the signaler's commands by placing and arranging the blocks on the table. The table acted as a shared workspace, as blocks placed on the table could be seen by both participants (although from opposite perspectives). Not all 12 blocks were used for every layout, and the signaler was not allowed to show the layout to the actor.

Since we wanted to observe natural communication in action, participants were also not allowed to talk or strategize beforehand, and no instruction on how to speak/gesture was given from the experimenter. A trial began when the experimenter presented a new block layout to the signaler and ended when the participants replicated the block layout.

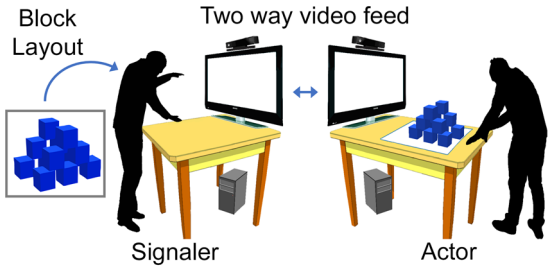

Figure 2: The experiment setup. The participant on the left (Signaler) was asked to direct the other participant (Actor) to construct a block layout.

Experimental Conditions

The type of communication allowed (speech, gesture, etc.) was restricted based on the following conditions.

- *Speech Only:* In this condition, participants were not able to see each other, and could only rely on speech to communicate; the video feed was restricted so that the signaler could only see the blocks on the table, and the actor could not see anything at all.
- *Gesture Only:* Participants could only see each other using the video presented on the TVs; the audio was muted, restricting participants to using only non-verbal communication to accomplish tasks.
- *Gesture and Speech:* Both audio and video were enabled, allowing participants to use any natural combination of speech and gesture they wished.

Block Layouts

We used a set of 80 different block layouts for this experiment. To obtain these layouts, we developed and deployed a web-based "game" allowing people to submit their own layouts which were then scored based on the uniqueness of the layout. This was done to elicit

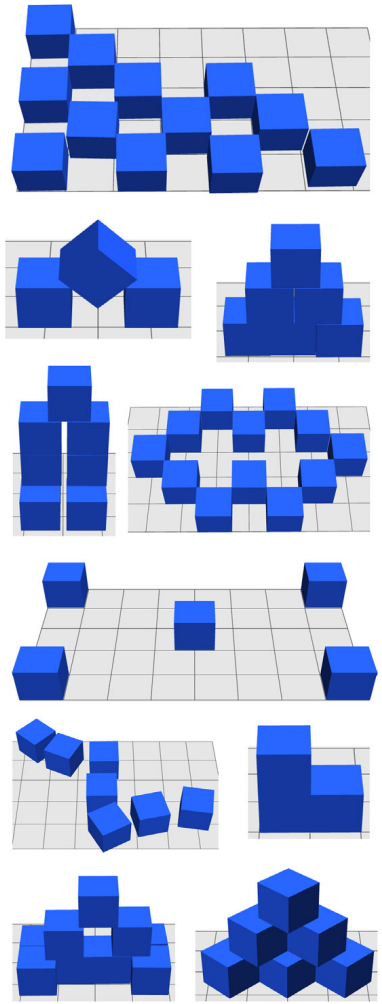

Figure 3: Examples of the wide variety of block layouts used in the experiment

a large variety of layouts. Different layouts would afford different intents; for instance, a layout with multiple layers may involve gestures/speech to stack blocks on top of each other, and a layout with rotated blocks would involve gestures/speech specifically to rotate blocks. Thus, a large user-submitted collection of layouts allowed us to capture gestures and utterances for a wide range of possible intents.

Through this crowdsourcing, we collected a corpus of over 200 layouts; out of these, 80 unique ones were chosen for use in the experiment. The layouts used vary in their difficulty and complexity (Figure 3).

Procedure

Each pair of participants were assigned to one of the experimental conditions at the start of the session, and one participant was designated the actor. The pair was asked to complete 10 tasks within 30 minutes. After 10 tasks (or 30 minutes, whichever came first), the participants switched roles and completed another 10 tasks for a total of up to 20 tasks. Each task in a session used a unique block layout, randomly selected from the corpus of 80 different layouts.

Participants

We recruited 60 participants (10 pairs per condition) through visits to computer science classes at a local university and through word-of-mouth. Participants were between the ages of 19-64 (mean = 24.2, SD = 7.7), and 17 participants were female. Out of the 60 participants, six were left-handed, and one person was ambidextrous. Half of all participants had experience with motion gesture systems like the Microsoft Kinect or Nintendo Wii. Out of the 30 pairs of participants, 16 pairs were previously acquainted with each other. Each individual received a $10 Amazon gift card as compensation.

Data Labeling

We annotated the video recordings from the sessions that included gestures (Gesture Only, Gesture and Speech). We used an internally developed tool for video annotation called EASEL to allow us to identify and mark the start/end of each observed gesture and assign it a label, for gestures made by the signaler. We assigned each gesture a label describing the physical movements of body parts. A specific labeling language was used to keep labeling consistent and descriptive. For instance, motion labels followed the format of "*<body part>: <motion>, <orientation>*". For example, "RA: move, left" describes a gesture in which the right arm (RA) moves in the left direction.

Additionally, the implied semantic intent of each gesture was also annotated, for example, "slide left," if the gesture was to signify that the current block(s) needed to be moved left. Intents included labels for sliding/moving blocks, stacking, signifying OK, stop, etc. The labels used for both gesture and intent were kept consistent across trials in the Gesture Only and Gesture and Speech conditions, in order to compare instances of gestures between the conditions. Only gestures from these two conditions were labeled; speech utterances were not transcribed at this point for any of the conditions containing speech.

Results

Labeling Results

We labeled 24,503 gesture instances from all trials, with 5,060 unique gestures. However, due to the nature of our labels as compound physical descriptions,

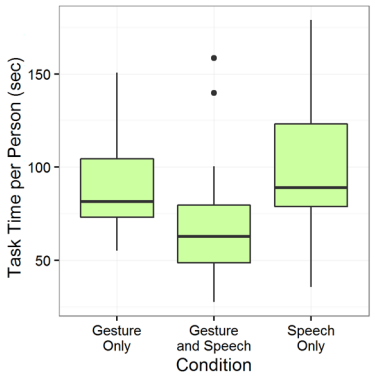

Figure 4: Average task time per person across the three conditions.

Gesture	Count
body: still	1850
RA: move, down	428
arms: move, down	265
head: nod	215
head: rotate	213

Table 1: The five most commonly occurring gesture labels in the Gesture Only condition.

Gesture	Count
body: still	2457
head: rotate	521
head: nod	190
RA: move, down	182
arms: move, down	172

Table 2: The five most commonly occurring gesture labels in the Gesture and Speech condition.

many of these only occurred once in the entire set. In actuality, only 1,427 gestures occurred two or more times, and only 110 gestures occurred at least twenty times and were used by more than one participant.

Gestures and Intents Used

We looked at the most common gesture labels from both the Gesture Only and Gesture and Speech conditions to compare differences in the observed physical motions. Tables 1 and 2 show the five most commonly occurring gesture labels and Tables 3 and 4 show the common intent labels for each condition. The majority of labels are "body: still" for both conditions (i.e. when the signaler participant is not moving and is either thinking or waiting for the actor to make a move); the most common intents are subsequently "wait" and "think." After these, there are a number of single arm and head movements, and a few notable gestures such as making a thumbs up gesture or moving the hands apart.

Our preliminary analysis of gesture frequency shows differences in intents used between the two conditions. Similar gestures and intents were condensed into 25 discrete actions and organized into four types: Numeric (counting gestures), Action (translate, rotate, etc.), Reference (this block, there, here, etc.), and Social Cues (start, done, OK, no, stop, etc.). For each action, we looked at the number of times the action occurred in the Gesture Only condition and compared it against the number of times it occurred in the Gesture and Speech condition. We found eight actions that were used at least five times as much in the Gesture Only condition than in Gesture and Speech.

Among these eight, five of them were numeric gestures (one through five). The other three were the action "servo translate" (which defines the act of continually translating an object in a direction until feedback—as in a stop or OK—is given), the reference "this column," and the social cue "no."

Task Performance

As the number of trials may have differed between participants, task performance was evaluated by taking the average task/trial time for each participant. One outlier (greater than three standard deviations away from the mean) was omitted from the Gesture Only condition. For the Gesture Only condition, the average trial time for a participant was 90.3 seconds (SD = 26.5). For the Gesture and Speech condition, the average was 69.6 seconds (SD = 33.5), and for Speech Only, the average was 97.4 seconds (SD = 39.5). Figure 4 shows the average task times per person across conditions.

Analysis of variance (ANOVA) revealed a significant effect of condition on task/trial completion time ($F_{2,56}$ = 3.659, $p < 0.05$). Post-hoc analysis using TukeyHSD correction revealed that the Gesture and Speech condition was significantly faster than Speech Only ($p < 0.05$). However, it showed no difference between Gesture Only against the two other conditions ($p > 0.10$ in both cases).

Discussion

Differing Gesture Strategies

We interpret the differences in gesture occurrences as differing gesturing strategies. Due to the presence of speech, numeric gestures may not be needed in most cases; the number of blocks needed can be stated

Intent	Count
wait	2019
think	813
here	757
OK	652
yes	474

Table 3: Common intent labels for the Gesture Only condition.

Intent	Count
talk	1850
wait	1412
think	1220
here	372
yes	194

Table 4: Common intent labels for Gesture and Speech condition.

verbally, and numeric gestures would likely fill a redundant role. Likewise, it would be easier to say "no" or verbally correct the other participant than to perform a full gesture. In the case of "this column" and "servo translate," representing these through gestures may not be necessary with speech because there are better words to describe the actions of specifying a column, and continually translating an object.

Further analysis is needed to compare our observations with the formal gesturing strategies (Gesticulation, Pantomime, Emblems, Sign, etc.) that appear in the presence/absence of speech as referenced in McNeill's extensions to Kendon's Continuum of gestures [13].

Performance of Gesture and Speech

We noted that task performance in the Gesture and Speech condition was significantly faster than Speech Only. These results suggest that there may be information encoded in the gestures used, as suggested by [16–18]. However, [6,11,12] emphasized that shared visual information is key and views of people do not matter as much in terms of task performance. In our study, we kept a shared workspace in all conditions (both participants retained their view of the blocks on the table) and still saw a difference in performance with the inclusion of gesture and a full view of both people. Thus, it is necessary to pursue transcription and analysis of speech utterances in addition to gesture instances, in order to fully understand their specific relationship and see what information is conveyed through each communication channel.

Implications to Gestural Interaction/Interfaces

When interacting with a lifelike computation agent on a collaborative task, our work suggests that designers should use both speech and gesture rather than speech alone as the interaction modality. We base this recommendation on the fact that using gesture and speech resulted in faster task times than speech alone.

It is also important to consider the use of social cues (e.g., acknowledgement such as OK, start task, etc.). Such social cues are critical in communicating feedback before, during, and after a task. We saw that these cues were frequently used between humans, but are not typically found in gesture sets.

Conclusion and Future Work

In this abstract, we presented results from a study exploring the use of gesture in collaborative tasks. We highlighted differences in gesturing strategies and presented implications for the design of gestural interactions.

Future work will involve analyzing the relationship of speech to gesture, requiring transcription. This will be done similar to the temporal analysis by [5,17]. This will enable us to understand exactly why certain gestures were not used as often in the presence of speech, and help identify why having both speech and gesture resulted in faster task performance times than speech alone. The interplay between the signaler and actor also needs to be explored.

Acknowledgements

This work was partially funded by the U.S. Defense Advanced Research Projects Agency (DARPA) and the U.S. Army Research Office (ARO) under contract #W911NF-15-1-0459.

References

1. Richard A. Bolt. 1980. "Put-that-there": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '80), 262–270. https://doi.org/10.1145/800250.807503
2. Alex Butler, Shahram Izadi, and Steve Hodges. 2008. SideSight: Multi-"Touch" Interaction Around Small Devices. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (UIST '08), 201–204. https://doi.org/10.1145/1449715.1449746
3. Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, Lauren Resnick, Levine B., M. John, Stephanie Teasley and D (eds.). American Psychological Association, 13--1991.
4. Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1: 1–39. https://dx.doi.org/10.1016/0010-0277(86)90010-7
5. Julien Epps, Sharon Oviatt, and Fang Chen. 2004. Integration of Speech and Gesture Inputs during Multimodal Interaction. In *Proceedings of the Australian Conference on Human-Computer Interaction* (OZCHI '04).
6. Susan R. Fussell, Robert E. Kraut, and Jane Siegel. 2000. Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (CSCW '00), 21–30. https://doi.org/10.1145/358916.358947
7. Susan R. Fussell, Leslie D. Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D. I. Kramer. 2004. Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Hum.-Comput. Interact.* 19, 3: 273–309. https://doi.org/10.1207/s15327051hci1903_3
8. Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2004. Action As Language in a Shared Visual Space. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (CSCW '04), 487–496. https://doi.org/10.1145/1031607.1031687
9. Chris Harrison and Scott E. Hudson. 2009. Abracadabra: Wireless, High-precision, and Unpowered Finger Input for Very Small Mobile Devices. In *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology* (UIST '09), 121–124. https://doi.org/10.1145/1622176.1622199
10. Sven Kratz and Michael Rohs. 2009. Hoverflow: Exploring Around-device Interaction with IR Distance Sensors. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (MobileHCI '09), 42:1–42:4. https://doi.org/10.1145/1613858.1613912
11. Robert E. Kraut, Susan R. Fussell, and Jane Siegel. 2003. Visual Information As a Conversational Resource in Collaborative Physical Tasks. *Hum.-Comput. Interact.* 18, 1: 13–49. https://doi.org/10.1207/S15327051HCI1812_2
12. Robert E. Kraut, Darren Gergle, and Susan R. Fussell. 2002. The Use of Visual Information in Shared Visual Spaces: Informing the Development of Virtual Co-presence. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (CSCW '02), 31–40. https://doi.org/10.1145/587078.587084

13. David McNeill (ed.). 2000. *Language and gesture*. Cambridge University Press, Cambridge ; New York.
14. Jörg Müller, Gilles Bailly, Thor Bossuyt, and Niklas Hillgren. 2014. MirrorTouch: Combining Touch and Mid-air Gestures for Public Displays. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services* (MobileHCI '14), 319–328. https://doi.org/10.1145/2628363.2628379
15. Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and Synchronization of Input Modes During Multimodal Human-computer Interaction. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI '97), 415–422. https://doi.org/10.1145/258549.258821
16. Thies Pfeiffer. 2011. Interaction between Speech and Gesture: Strategies for Pointing to Distant Objects. In *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, Eleni Efthimiou, Georgios Kouroupetroglou and Stavroula-Evita Fotinea (eds.). Springer Berlin Heidelberg, 238–249. https://doi.org/10.1007/978-3-642-34182-3_22
17. Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal Human Discourse: Gesture and Speech. *ACM Trans. Comput.-Hum. Interact.* 9, 3: 171–193. https://doi.org/10.1145/568513.568514
18. Elizabeth S. Veinott, Judith Olson, Gary M. Olson, and Xiaolan Fu. 1999. Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '99), 302–309. https://doi.org/10.1145/302979.303067
19. Robert Walter, Gilles Bailly, and Jörg Müller. 2013. StrikeAPose: Revealing Mid-air Gestures on Public Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 841–850. https://doi.org/10.1145/2470654.2470774
20. Robert Walter, Gilles Bailly, Nina Valkanova, and Jörg Müller. 2014. Cuenesics: Using Mid-air Gestures to Select Items on Interactive Public Displays. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services* (MobileHCI '14), 299–308. https://doi.org/10.1145/2628363.2628368
21. iOS - Siri - Apple. Retrieved January 11, 2017 from http://www.apple.com/ios/siri/
22. Amazon Alexa. Retrieved January 11, 2017 from http://alexa.amazon.com/spa/index.html