"You two! Take off!": Creating, Modifying and Commanding Groups of Robots Using Face Engagement and Indirect Speech in Voice Commands

Shokoofeh Pourmehr, Valiallah (Mani) Monajjemi, Richard Vaughan and Greg Mori
School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
{spourmeh, mmonajje, vaughan, mori}@sfu.ca

***Abstract*— We present a multimodal system for creating, modifying and commanding groups of robots from a population. Extending our previous work on selecting an individual robot from a population by face engagement, we show that we can dynamically create groups of a desired number of robots by speaking the number we desire, e.g. "You three", and looking at the robots we intend to form the group. We evaluate two different methods of detecting which robots are intended by the user, and show that an iterated election performs well in our setting. We also show that teams can be modified by adding and removing individual robots: "And you. Not you". The success of the system is examined for different spatial configurations of robots with respect to each other and the user to find the proper workspace of selection methods.**

I. INTRODUCTION

In this paper we propose and demonstrate a new interaction mode for multi-robot HRI: standing in front of a population of robots, a user can designate a subgroup of determined size by looking at them and saying "You two!" (or three, or n). The robots cooperate to combine their independent observations of the user's face and determine which robots were intended. Membership of the group can then be modified by adding a robot with "And you" or removing one with "Not you". The team can then be commanded as a unit with e.g. "Take off!" (Figure 1). The user wears a bluetooth earpiece microphone but is otherwise uninstrumented. In a series of real-world experiments we show that the method works reliably for a wide range of relative poses of user and three robots.

To increase the efficiency and naturalness of interaction between humans and multi robot systems, we have been working on methods for uninstrumented humans to select and command individual and groups of robots. Inspired by the ways humans interact with each other or with animals, we have utilized face engagements, pointing gestures and spoken commands to interact with teams of robots. We have previously shown that users can select an individual robot from a group of robots by simply looking directly at it [1].

The contributions of this paper are (i) to propose a new interaction modality using indirect speech ("You two!"); (ii) to show that human-robot face engagement can be used to determine the subject or subjects of verbal commands using indirect speech. To do this we introduce two methods for selecting groups by face engagement. We also provide the first analysis of the reliability of selection by face engagement as the spatial arrangement of user and robots varies. The paper also serves as a case study of an integrated, complete and compelling multi-robot HRI system.

II. BACKGROUND

A. Multimodal human-robot interaction

Human-robot interaction (HRI) is an active area of research. Goodrich and Schultz [2] provide a survey of the field. Chen et al. [3] comprehensively examined human performance issues for interacting with teleoperated robots. Multimodal interaction approaches have recently received some attention by the HRI community, with speech inputs used in conjunction with other modalities to drive interaction. Draper et al. [4] compared manual and speech input when operators had to control UAVs. The study showed that using voice commands can significantly improve an operator's ability to control subsystems.

Some studies have elected to combine vision modalities with speech since they are natural means of communication for humans. Steifelhagen et al. [5] is a good example of an integrated system which includes speech recognition and vision for colour-based hand and face tracking to estimate pointing direction. Similarly, Perzanowski et al. [6] present a multimodal speech and gesture-based interface to work with teams of cooperative robots; they utilize the knowledge of spatial relations obtained by speech input along with gesture information from an active vision system to build context predicates. Recent work includes Briggs et al. [7] which incorporate spoken inputs and vision components to update the belief model of autonomous agents. Prasov [8] describes the role of shared gaze between a human and an individual robot during remote spoken collaboration.

Eye contact and gaze play an important role in initiating and regulating communication between people [9]. Throughout this paper, we will use the term *face engagement* as coined by Goffman to describe the process in which people use eye contact, gaze and facial gestures to interact with or engage each other [10]. The role of eye contact plays such an important role in the development of humans that the ability to detect eye contact is present at birth [11]. We therefore believe that face engagement could be an effective communication channel for human-robot interaction.

Fig. 1: An uninstrumented person selecting and commanding multiple robots out of a group by looking at them and saying the desired number of robots.

B. Robot Selection and Task Delegation

There is little work on human-robot interfaces for multi-robot systems. Examples can be broken up into two general cases:

1) Traditional Human-Computer Interfaces: Rather than interacting directly with robots, a traditional human-computer interface is used to represent the spatial configuration of the robots and allow the user to remotely interact with the robots. Examples include McLurkin et al. [12] that uses an overhead-view of the swarm in a traditional point-and-click GUI named "SwarmCraft", and work by Kato that displays an overhead live video feed of the system on an interactive multi-touch computer table, with which users can control the robots' paths by drawing a vector field over top of the world [13]. Chernova et al. [14] propose a multi-robot demonstration learning framework named "FlexMLfD" where users can teach individual policies to multiple robots at the same time.

2) Embodied, World-Embedded Interactions: Embodied, world-embedded interactions occur directly between the human and robot, through mechanical or sensor-mediated interfaces. A useful property of this type of interaction is that since robots observe humans directly using their onboard sensing, they may not need to localize themselves in a shared coordinate frame in contrast to the GUI-based interfaces. Also, human users can walk and work among the robots, and are not tied to an operator station. Examples include work by Payton that uses an omnidirectional IR LED to broadcast messages to all robots, and a narrow, directional IR LED to select and command individual robots ([15], [16]). Naghsh et al. [17] present a similar system designed for firefighters, but do not discuss selecting individual robots. Zhao et al. [18] propose the user interacts with the environment by leaving fiducial-based "notes" (for example, "vacuum the floor" or "mop the floor") for the robots at work site locations. Xue et al. [19] introduce a clever fiducial design for imperfect visibility conditions and combines this with user-centric gestures in an underwater scenario . In our previous works, we developed face engagement [1] and pointing-gesture [20] techniques for single-robot and circling-gesture [21] techniques for multi-robot selection. However the vision system for interpreting circling gestures lacked robustness. Our novel system allows human operator to interact with multiple robots in a shared environment by only using voice and visual and linguistic cues.

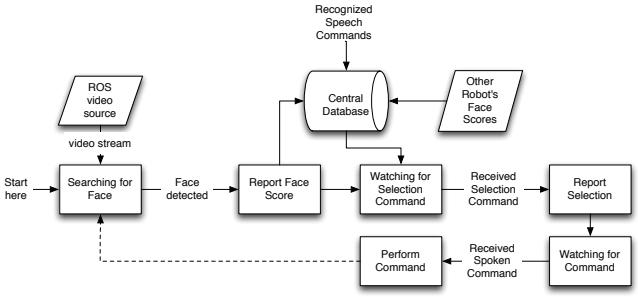


Fig. 2: System diagram: the system runs on each robot.

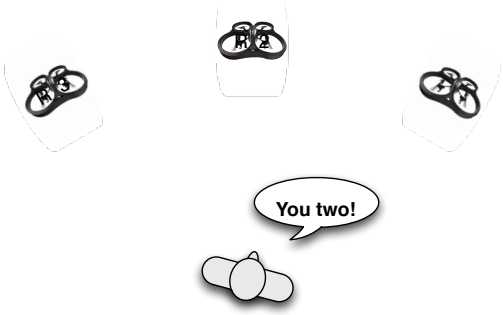
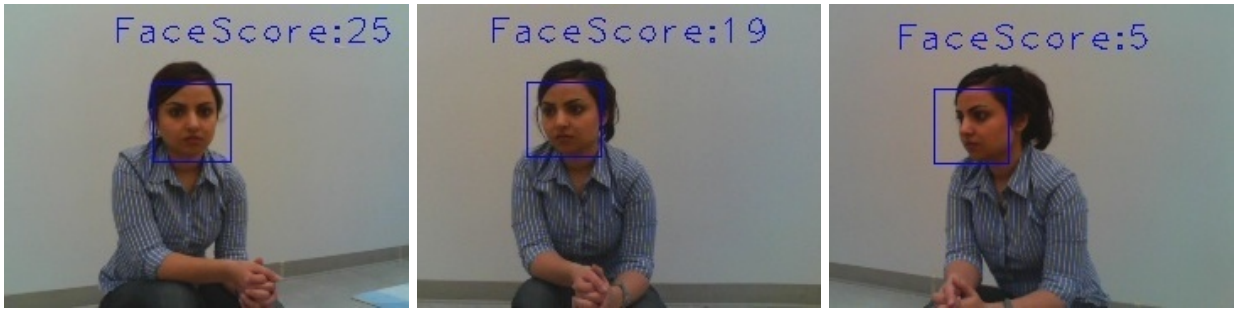


Fig. 3: A human operator creates a team of robots by looking at them and uttering the desired number of robots.

III. Method

We assume that before assigning a task to a team of robots, the human operator must select some robots from the population to form the team. We seek to design HRI systems that make this easy. We believe that a good approach is for uninstrumented humans to interact with teams of autonomous robots as they would with teams of humans, since this is familiar. This motivates our choice of face engagement and spoken commands.

In our system, each robot runs a face detector and communicates with a centralized voice recognition subsystem



(a) Right (b) Middle (c) Left

Fig. 4: An example of three robots' simultaneous camera views while arranged around a human operator. The user intends to engage the right-hand robot (view a) and it has the highest face score.

and database to coordinate their activity. Interfacing and message passing is enabled by ROS[1] [22]. The system layout is summarized in Figure 2.

A. Face Detection and Tracking

The first step of robot selection is to detect and track the human face. Each robot is equipped with a video camera and faces are detected in each frame. Face detection is done using the OpenCV [23] implementation of the Viola-Jones face detector [24]. Once a face is detected, a Kalman Filter is used to track the detected face to achieve robustness to occasional false negative and false positive detections.

The robots use the face detector to understand if they are currently being looked at by the human. One challenge in a multi-robot system is that the human face can be visible to multiple robots at the same time. To solve this problem, we use a mechanism developed and successfully used earlier by our group [1]. The face detector, a cascade Haar classifier, finds a group of neighbouring sub-windows around each candidate face. Since, the classifier is trained on the frontal-faces only, the number of such sub-windows increases when the human is directly looking at the camera (Figure 4). We use this number as a score to assess the quality of the currently tracked face. In the next section, we will describe how our system uses this so-called "face score" to determine which robot is currently being engaged by the user.

B. Voice Recognition

We employed the PocketSphinx library [25] to do speech recognition. PocketSphinx is an open source speech recognition system which matches voice commands with a predefined vocabulary. The vocabulary we used is defined with the words and phrases necessary for our system. It is a very small vocabulary which makes speech recognition very accurate in practice but requires the human operator to learn the set of allowed words and phrases.

C. Robot Selection

Our interaction design calls for the user to announce the desired number of robots (e.g. *"you two"*) and look at them (Figure 3). When the keyword "you" is detected, all robots currently tracking the user's face announce their ID and face score to the central database (and display their state by changing the LED colors they show to the user). In the experimental section below we describe two variations of our basic election method where we define a team of size n as the either the n robots with the highest simultaneous face scores, or the best single face score, iterated n times, with the winner of each round not participating in subsequent rounds. In either case, the intention is that the n robots that are most attended to by the user's face form the group.

IV. EXPERIMENTAL RESULTS

To demonstrate and validate the system, we performed several experiments with different spatial configurations of robots with respect to the user and other robots (Figure 5). Since we are using face engagement to attract the robot's attention, the spatial arrangement of the workspace is important. Experiments are designed to find the spatial arrangements of user and robots that work for our system. For convenience in these experiments we used laptops with integrated webcams to stand in for robots; however, our video demonstration shows the system working with three low-cost UAV robots.

The human operator uses a small lexicon for announcing the desired number of robots (e.g. "you three"), modifying the group (adding a new member (e.g. "and you") or removing (e.g. "not you")), getting robot's attention or regourping (e.g. "again" or "robots") and commanding the selected group (e.g. "take off").

As shown in Figure 5, for each experiment robots are located l m from the user with θ degrees of separation and the user at the centre. l ranges from **1** to **2.5** m with **0.5** m steps and θ ranges from **15** to **90** degrees with **15** degree steps. In each configuration, the user attempts to select single or multiple robots as required by the trial. Each experiment is repeated five times, so that overall 198 experiments were performed.

The results are shown in Figures 6-8. The graphs show the success rate of selecting the desired robots located at l m from user and at θ degrees from other robots. Due to symmetry between the left and right hand robot cases, right and left results are combined. The results are presented as

[1] http://www.ros.org/

a heat map, where a white colour indicates 100% success rate and black colour 0% success rate. No experiments were performed in the hatched area, as it was either too close or too far for the face detection to work, or there was not room to fit three robots in that space.

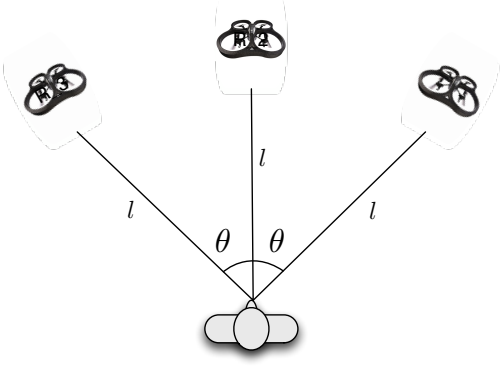


Fig. 5: Robots' configuration with respect to user and each other

A. Single robot selection

As shown in our previous work, one can select an individual robot from a population by making face engagement with it. In order to find the spacial workspace of this method in the presence of other robots, we vary the robots' spatial configuration and tried to select one of them. In every experiment, the human operator directly looked towards the desired robot and uttered *you*. To isolate the spatial effects, for this experiment we assume that the voice recognition module works perfectly. The results are shown in Figure 6.

Figure 6-a shows the average success rate over 5 repeats of selecting either of the left or right robots (i.e. at $\pm\theta$ degrees from the human pose) and Figure 6-b shows the same measure for the middle robot (i.e. at 0 degrees from the human pose). The results show that when robots are very close together or very far from the human operator the success rate of selecting the desired robot decreases. This is due to their face scores becoming similar so one robot is selected effectively at random. The failure rate is higher when the user tries to select the middle robot, because there are two sources of error (selection of the right or left robot instead of the middle robot).

B. Multi-Robot Selection

To select a subgroup of robots from a population, we investigate two different ways of making face engagement with multiple robots. One is by looking toward the whole group and trying to make face engagement with all of them simultaneously. The other is to select the desired robots one by one. We repeat the experiment above, this time selecting two or three neighbouring robots from our population of three. Again we vary the spatial layout to find how sensitive the method is to the spatial layout of the workspace. Results are shown in Figure 7 and Figure 8.

1) Method 1: Simultaneous Selection: In order to select a group of robots, one instinctively looks toward them. We call this method *simultaneous selection*. In this method, the human operator looks toward the whole desired group of robots and tries to make face engagement with all of them at the same time. The number of robots the user has asked for, will be selected simultaneously by electing the n robots with the highest face scores.

In this method, our observed success rate varied strongly with the human-robot distance and the angle between robots. Figure 7-b shows that when the robots stand less than 15 degrees apart and all three robots can completely see the user's frontal face the success rate is very high. As the angle between robots increases, the workspace is limited to shorter distances since the user cannot have face engagement with all of them at the same time. So the spatial workspace of this method is limited.

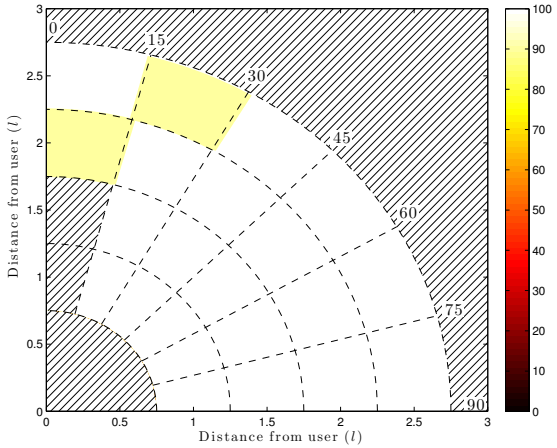
As in the first experiment selecting one-from-three robots, when selecting two-from-three robots with this method the close proximity of the third robot can cause an incorrect selection. Figure 7-a shows that with the robots located very close together (θ less than 15 degrees apart and thus with very similar face scores) only a 40% success was observed in this situation.

2) Method 2: Incremental Selection: To improve on the first method, we devised a second method in which face engagement is iterated over the set of desired robots. We name this method *incremental selection*. In this method, after announcing the desired number of robots n, the robots with the highest face scores will get selected one after each other in n rounds, with the winner of each round not participating in later rounds.

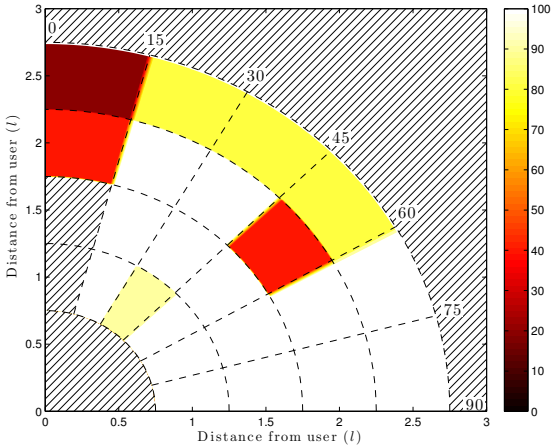
By incrementally selecting robots, the user can group robots located far from each other because she is able to have face engagement with each of the desired robots separately. The success rate of incremental selection of multiple robots is illustrated in Figure 8. The results indicate that this method of multi-robot selection has wider spatial workspace: it is robust to a wider set of mutual poses. Since the human operator can look individually at all the desired robots for selecting them, their configurations have less effect on the success rate. The only source of failure we saw using this method is when robots are posed very close together, which is in common with the single-robot selection mode. According to Figure 8-b we can conclude that the workspace of selecting all three robots incrementally is the whole area in which the face detector works.

C. Combining group and individual engagement

Since we can select groups and individuals, and issue keyword commands by voice, we can combine these parts. We add the keywords "And you" and "Not you" to add and remove individual robots, currently face-engaged, to and from the team. This allows us to create teams of neighbouring robots and add individual distant robots afterwards. It also allows us to recover from incorrect group allocations, since if the wrong robot was added to the team

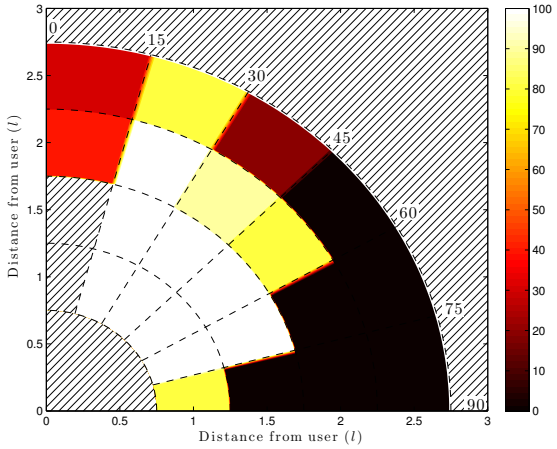


(a) Selection of one of the side robots

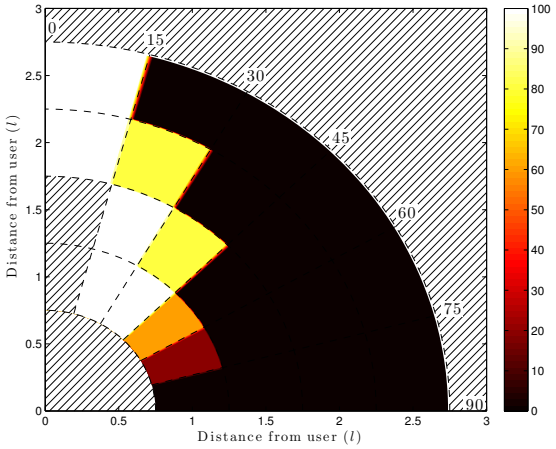


(b) Selection of middle robot

Fig. 6: Success rate of selecting an individual robot.

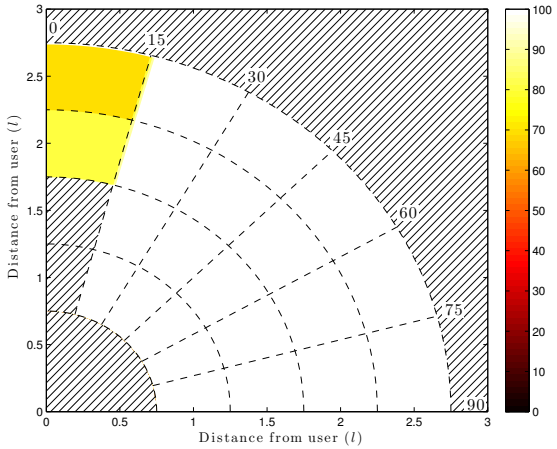


(a) Selection of two robots

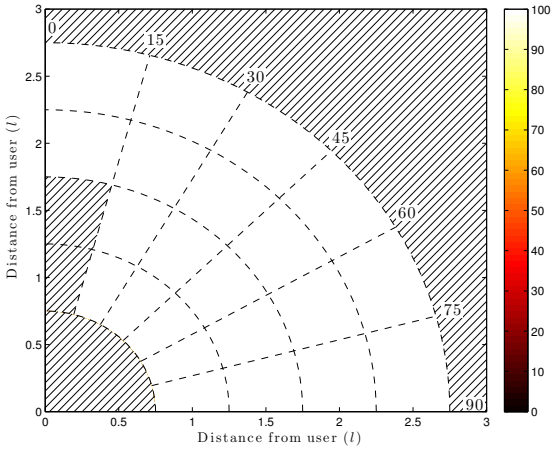


(b) Selection of three robots

Fig. 7: Success rate of simultaneous selection of multiple robots.



(a) Selection of two robots



(b) Selection of three robots

Fig. 8: Success rate of incremental selection of multiple robots.

we can remove it ("Not you") then add our preferred robot ("And you"). Once a group is correctly assembled, we can command the group as a unit. In the video demonstration accompanying this paper we show a team of UAVs being created, modified, and commanded to "Take off".

V. CONCLUSIONS AND FUTURE WORK

We have described a system which integrates spoken commands and face engagement to create, modify, and command teams of robots. We introduced two modes of selecting multiple robot and compared them, concluding that iterated election is much more robust to spatial layout compared to simultaneous election. This is because in iterated election the user can look around from robot to robot in the team rather than having to look at their centre of mass.

In future work we will demonstrate the practicality of our methods on working outdoor robot systems including heterogeneous teams of robots. We will examine the problem of normalizing face scores between robots whose distance to the user is not the same. And we will extend this work to designate teams of robot by name, so we can say "You three are Red Team", "You three join Blue Team", and "You switch to Green Team". Further, we aim to test whether the face engagement approach is practical when humans and robots are moving relative to each other, where the face score will continuously change: this is a condition for use in fixed-wing UAVs, for example.

In all this work, we aim for simple, robust methods that are easy and intuitive to use. The data in this paper show that the method is robust as long as the face detector is working and the robots are not too close together, but we suggest that the video shows what the data can not: the simple and natural feel of our interaction design.

ACKNOWLEDGMENT

This work was supported by the NSERC Canadian Field Robotics Network, and other NSERC funding.

REFERENCES

[1] A. Couture-Beil, R. T. Vaughan, and G. Mori, "Selecting and commanding individual robots in a vision-based multi-robot system," in *Proc. of the Canadian Conf. on Computer and Robot Vision*, May 2010, pp. 159–166.

[2] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: A survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[3] J. Y. Chen, E. C. Haas, and M. J. Barnes, "Human performance issues and user interface design for teleoperated robots," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1231–1245, 2007.

[4] M. Draper, G. Calhoun, H. Ruff, D. Williamson, and T. Barry, "Manual versus speech input for unmanned aerial vehicle control station operations," in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, 2003, pp. 109–113.

[5] R. Stiefelhagen, C. Fugen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *Proeedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Sendai, Japan, Sept. 2004, pp. 2422–2427.

[6] D. Perzanowski, A. C. Schultz, W. Adams, M. Bugajska, E. Marsh, G. Trafton, D. Brock, M. Skubic, and M. Abramson, "Communicating with teams of cooperative robots," in *Proc. from the 2002 NRL Workshop on Multi-Robot Systems*. Kluwer, 2002, pp. 16–20.

[7] G. Briggs and M. Scheutz, "Multi-modal belief updates in multi-robot human-robot dialogue interaction," in *Proc. of 2012 Symposium on Linguistic and Cognitive Approaches to Dialogue Agents*, 2012, pp. 67–72.

[8] Z. Prasov, "Shared gaze in remote spoken hri during distributed military operations," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE Int. Conf. on*, March 2012, pp. 211–212.

[9] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, pp. 22–63, 1967.

[10] E. Goffman, *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, September 1966.

[11] T. Farroni, G. Csibra, F. Simion, and M. H. Johnson, "Eye contact detection in humans from birth," in *Proc. of the National Academy of Sciences of the United States of America*, 1999, pp. 9602–9605.

[12] J. McLurkin, J. Smith, J. Frankel, D. Sotkowitz, D. Blau, and B. Schmidt, "Speaking swarmish: Human-Robot interface design for large swarms of autonomous mobile robots," in *Proc. of the AAAI Spring Symposium*, 2006, pp. 72–75.

[13] J. Kato, D. Sakamoto, M. Inami, and T. Igarashi, "Multi-touch interface for controlling multiple mobile robots," in *Proc. of the 27th Int. Conf. on Human factors in Computing Systems (Extended Abstracts)*. ACM, 2009, pp. 3443–3448.

[14] S. Chernova and M. Veloso, "Confidence-based multi-robot learning from demonstration," *Int. Journal of Social Robotics*, vol. 2, no. 2, pp. 195–215, 2010.

[15] D. Payton, "Pheromone robotics," http://www.swarm-robotics.org/SAB04/presentations/payton-review.pdf, 2004, slides from a presentation at the Swarm Robotics Workshop, SAB04. Retrieved September 28, 2009.

[16] M. Daily, Y. Cho, K. Martin, and D. Payton, "World embedded interfaces for human-robot interaction," in *Proc. of the 36th Annual Hawaii Int. Conf. on System Sciences (HICSS'03) - Track 5 - Volume 5*, ser. HICSS '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 125.2–. [Online]. Available: http://dl.acm.org/citation.cfm?id=820752.821587

[17] A. M. Naghsh, J. Gancet, A. Tanoto, and C. Roast, "Analysis and design of human-robot swarm interaction in firefighting," in *IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, Munich, Germany, Aug. 2008, pp. 255–260.

[18] S. Zhao, K. Nakamura, K. Ishii, and T. Igarashi, "Magic cards: a paper tag interface for implicit robot control," in *Proc. of the 27th Int. Conf. on Human Factors in Computing Systems (CHI)*. ACM, Apr. 2009, pp. 173–182.

[19] A. Xu, G. Dudek, and J. Sattar, "A natural gesture interface for operating robotic systems," in *Proc. of the 2008 IEEE Int. Conf. on Robotics and Automation (ICRA '08)*, Pasadena, California, USA, May 2008, pp. 3557–3563.

[20] S. Pourmehr, V. Monajjemi, J. Wawerla, R. Vaughan, and G. Mori, "A robust integrated system for selecting and commanding multiple mobile robots," in *IEEE Int. Conf. on Robotics and Automation*, 2013.

[21] B. Milligan, G. Mori, and R. T. Vaughan, "Selecting and commanding groups in a multi-robot vision based system," in *6th ACM/IEEE Int. Conf. on Human-Robot Interaction (Video Session)*, 2011.

[22] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Int. Conf. on Robotics and Automation*, 2009.

[23] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, 2008.

[24] P. Viola and M. Jones, "Robust real-time face detection," *intl. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[25] D. Huggins-Daines, M. Kumar, A. Chan, A. Black, M. Ravishankar, and A. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proc. 2006 IEEE Int. Conf. on*, vol. 1, May, pp. I–I.