Wearable multi-modal interfaces for mixed-initiative interaction in human multi-robot teams

B. Gromov, L. M. Gambardella, Gianni A. Di Caro

I. INTRODUCTION

In a *mixed team*, humans and one or more, possibly heterogeneous, robots share the same physical environment and work '"shoulder-to-shoulder" to perform a common mission. To allow a smooth and effective collaboration in such spatially distributed and heterogeneous system, *interaction* plays a fundamental role and needs to be suitably designed. When the physical environment is that of a *search and rescue mission*, then the scenario poses specific challenges that need to be addressed when designing interaction: usually, the embedding environment is unstructured and possibly harsh/hostile (e.g., after a natural disaster), communication infrastructure is hardly available and light and visibility conditions can vary very much. There, human agents are professionals who need to focus on their job (e.g., rather than focusing on how to "talk" to the robots. However, in spite of the many practical challenges that arise, it is apparent that the deployment of a fleet of heterogeneous robots can result in a net advantage while performing the mission, as long as it is possible to effectively use their artificial sensory-motor and processing capabilities.

In typical search and rescue operations *global mission planner* assigns to the individual agents the task to be performed, possibly with an associated time schedule. A task assignment could consist in telling rescuers A, B, C to search for survivors in a specific mission area (e.g., inside a collapsed house) for the next 30 minutes, and telling flying robots R_1, R_2 to collect aerial images of another area. However, the same resources could be used in a mixed group of humans-robots, to search for survivors in the same area. Assuming that robots have enough locomotion and decisional autonomy to implement the search request from the global planner, this would be a basic example of a mixed team in action. However, in order to get the best out of the deployed resources, and locally create a *coalition* among humans and robots, some explicit interaction can be profitably used. E.g. the human can use his/her own superior cognitive abilities to locally direct the actions of the robots. This would amount to a form of *local, high-level action planning* that would complement the global mission planning. On the other side, robots could explicitly ask for support from the human to execute complex actions (e.g., open the door) or to decide where to search next. In more general terms, a *mixed initiative system* can be profitably locally set up, in which robot have enough autonomy to implement basic directives given by humans and can also directly include the humans in the proximity of their action/decision loop.

On these premises, the objective of this work is to design and prototype different solutions for enabling efficient and natural interactions in both directions–from humans to multi-robot systems and vice versa, with the aim of paving the road towards *mixed initiative interaction in real-world search and rescue scenarios*.

II. MULTI-MODAL VOCABULARY FOR INTERACTION FROM HUMAN TO MULTI-ROBOT SYSTEMS

Based on these basic insights, our approach to enable effective and practical mixed initiative interaction between humans and multiple robots is the definition of a *multi-modal vocabulary* and a basic *grammar*. The primary goal is to maximize information transfer in both directions, minimizing human's cognitive load, and the time needed for communication acts. For humans the vocabulary is built combining means that are natural and intuitive for humans, such as *speech* and *arm* and *hand gestures*. For the multi-robot side, the goal is achieved by distributed strategies for group communications using *coordinated movements, lights, sounds, and voice messages*. Moreover, robots are empowered with basic autonomous mechanisms to perform in-group selection and to respond as a unit to human commands.

We use different modalities for referring to different notions or entities when constructing a sentence. In particular, from the human side: *hand gestures* should be used to express iconic commands and implement simple mobility controls; *arm gestures* are essential to convey information about spatially-related notions, such as indicating directions, pointing to objects and structures, selecting humans, robots, or groups of them; *speech* is conveniently used to express complex commands and behaviors, or to effectively broadcast critical messages, such as alarms. The different modalities are used concurrently and then *fused* together, in order to be mutually confirmed and to reduce decoding and classification errors.

III. USE OF WEARABLE DEVICES TO MAKE THE ROBOT *passive* DURING INTERACTION

In this respect, instead of making the robots decoding and classifying the multi-modal signals issued by a human, as is common practice (e.g., the robots use vision to "see" human gestures and then classify them), our approach is to rely on the use of *wearable devices*, in combination with sensors

(*) All authors are with the Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland. Email: {boris,luca,gianni}@idsia.ch.This work was partially supported by the Swiss National Science Foundation (SNSF) through the National Centre of Competence in Research (NCCR) Robotics.

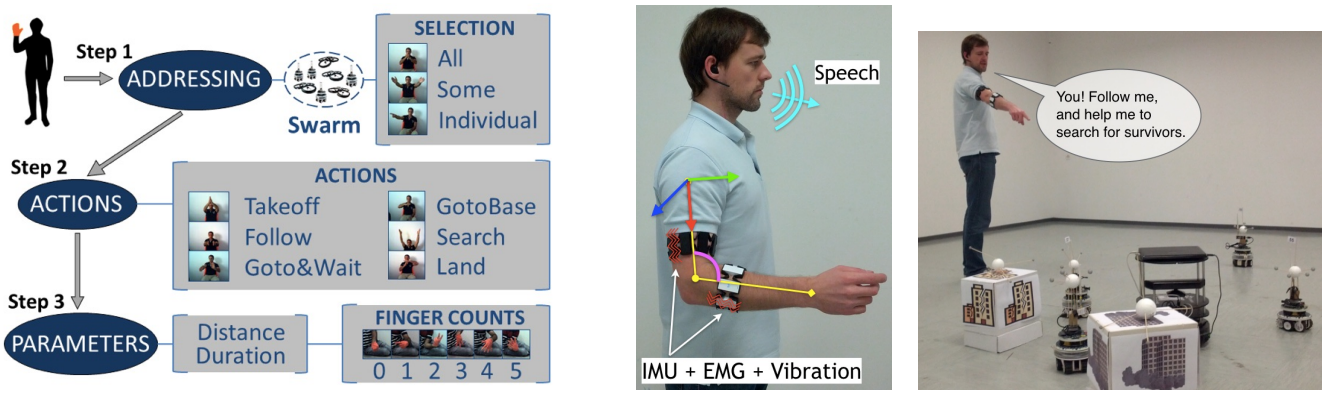

Fig. 1. (*Left*) The gesture vocabulary defined in previous work [1], [2], [3]. The vocabulary is mostly intended for search and rescue missions and assumes the presence of flying robots. Robots need to "see" the human and rely on machine vision to classify the gesture. (*Right*) A first multi-modal system employing wearable devices: two Thalmic Lab's Myo Armbands and one wireless headset microphone for speech recognition. The Myo devices comes with IMUs and EMGs. IMUs data allow the reconstruction of the 3D pose of the arm and are used to indicate spatial entities, such as directions and robots. EMG data permit to implement hand gestures for the fine control of robots' motion. Speech recognition is used to encode complex commands. Robots receive via radio already decoded information and do not need to "watch" the human.

on-board of the robots, to fuse and decode human signals (e.g., gestures and speech) directly *on-board of the human*. *Local ad hoc networking* between the human and the robots takes care of reliably transmitting to the robots the processed inputs. This way of proceeding is dictated by the need to make interaction *robust to different external conditions* (e.g., in terms of illumination and background noise) and leave as much *autonomy* as possible to the agents (i.e., robots should use their cameras for mission tasks, not to keep watching humans to catch possible command gestures). Therefore, a solution based, for instance, on the the robots visually detecting and classifying human gestures, which we adopted in previous work [1], [2], [3] and which is extremely popular in the literature, would not be really satisfactory and could easily fail when visual conditions are degraded. In Fig. 1 the gesture vocabulary developed in previous work is shown vs. a first implementation of a multi-modal system, where robots are mostly "passive" regarding interaction.

We present the system in supplementary video[1], where a human interacts with robots (Foot-bots / marXbots) using voice commands, arm and hand gestures.

Selection of individual robots performed with *addressing* voice commands accompanied by pointing gestures. The command consists of the ID of a robot, that allows us to query respective robot about its location in the world and thus localize the user w.r.t. to its pose. In principle, such command has to be used only once, unless the user wants to move to another location.

Once a robot is selected and human is localized *relative* spatial commands and gestures, e.g. "go there", can be used.

Selection of a group of robots with a single selection gestures is done by 'seeding' one of the robots in the group and making it to recruit a particular number of robots around. The selected group then can be controlled in exactly the same way as a single robot.

[1] https://youtu.be/FWMCxARQYhY

Due to accumulation of pointing and localization errors pointed locations can be very imprecise. To intervene and adjust final destination of a robot or a group we utilize *manual control based on hand gestures*.

Multiple individual selections of robots, i.e. cherry-picking, is useful when the robots of interest are surrounded by others and therefore group selection by seeding is not feasible. By keep quickly addressing individual robots they are gathered to a group.

In the video we also demonstrate how fast the actual *arm tracking and selection of individual robots* can be done.

The *two-armed spatial selection of groups of robots* is meant to be used whenever a large group of robots has to be subdivided, e.g. selecting a half of the pool of available robots.

As can be seen the multi-modal interface that we propose can easily be used with any types of robots as well as in heterogeneous groups. We demonstrate this by using the same interface simultaneously on Foot-bots and a quadcopter.

Finally, we show a basic approach to *voice and gestures fusion* in the context of error reduction. The system analyses the compliance of issued voice commands and corresponding arm gestures. In the trivial case the voice commands that assume presence of specific pointing gestures are ignored if the later are missing.

References

[1] A. Giusti, J. Nagi, L. Gambardella, and G. A. Di Caro, "Cooperative sensing and recognition by a swarm of mobile robots," in *Proceedings of the 25th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, October 7–12, 2012, pp. 551–558.

[2] J. Nagi, H. Ngo, A. Giusti, L. Gambardella, J. Schmidhuber, and G. A. Di Caro, "Incremental learning using partial feedback for gesture-based human-swarm interaction," in *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Paris, France, September 9–13, 2012, pp. 898–905.

[3] J. Nagi, A. Giusti, L. Gambardella, and G. A. Di Caro, "Human-swarm interaction using spatial gestures," in *Proceedings of the 27th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Chicago, IL, USA, September 14–18, 2014.