Wearable Multi-modal Interface for Human Multi-robot Interaction

Boris Gromov[1], Luca M. Gambardella[1], Gianni A. Di Caro[2]

***Abstract*— A complete prototype for multi-modal interaction between humans and multi-robot systems is described. The application focus is on search and rescue missions. From the human-side, speech and arm and hand gestures are combined to select, localize, and communicate task requests and spatial information to one or more robots in the field. From the robot side, LEDs and vocal messages are used to provide feedback to the human. The robots also employ coordinated autonomy to implement group behaviors for mixed initiative interaction. The system has been tested with different robotic platforms based on a number of different useful interaction patterns.**

I. Introduction

In this work, we envisage *search and rescue* (SAR) scenarios where human agents and multiple robots are both part of the rescue team and work "shoulder-to-shoulder", sharing the same physical environment and interacting as peers. In these forthcoming scenarios, the success of a SAR mission will be greatly affected by the performance of the bidirectional interaction between humans and robots, assuming that the robots have a certain degree of autonomy while performing the assigned tasks.

Based on this view, we design and prototype a *wearable multi-modal interface* that enables effective bidirectional interactions from humans to multi-robot systems and vice versa. No external infrastructure or hand-held instrumentation is needed as often used in non-SAR research [1]. The interface is functional to set up a *mixed initiative* system [2], with a special focus on real-world search and rescue operations. The multi-robot system is empowered with autonomous group strategies for providing feedback to the human and for effectively actuating as a collective system based on human directives and commands.

More specifically, *from the user side*, we take advantage of human natural ability to concurrently use different interaction modalities and blend them in proportions appropriate for the information being conveyed, following similar research approaches [3, 4]. We use *deictic arm gestures* to convey information about spatially-related notions, such as indicating directions, pointing to objects and structures, selecting humans, robots, or groups of them. *Hand gestures* are used to express iconic commands and simple mobility controls. Finally, *speech commands* are used to express more complex notions, as well as to reinforce and confirm the basic notions expressed through gestures. Based on this categorization of modalities, a *multi-modal vocabulary and a basic grammar* have been designed, with the goal of maximizing reliable information transfer and minimizing humans cognitive load and the time needed for communication acts. To transform gestures-related spatial information (e.g., directions), into a common reference frame for the humans and the robots, we have developed a *relative localization system*, which is set up and maintained dynamically on the spot. No GPS is required.

A fundamental component of an effective system for human-robot interaction is represented by the feedback provided to the human. Since we consider a multi-robot scenario, coordination needs to be in place to let the multiple robots providing a coherent aggregate feedback to the user (in general how to effectively get information from a group of robots is not a trivial task, e.g., [5]). Therefore, *from the multi-robot side*, distributed group strategies for providing feedback to the human have been developed based on *coordinated movements, lights, sounds, and voice messages*. Moreover, the robots have been empowered with basic autonomous mechanisms to perform in-group selection, formation control during group motion, and, more in general, to respond as a unit to human inputs and commands.

In this work we stress the fact that *multiple robots* are expected to be concurrently deployed and locally interact with humans. Compared to the single robot case, a number of complicating matters arise in this context, that require to combine multiple modalities from the human side to issue requests [6, 7], and set up coordinated group strategies from the robot side to provide feedback without exhausting human's sensory and cognitive capabilities.

In order to perform interaction in a natural and robust manner, our approach is to rely on the use of off-the-shelf wearable devices, that *locally* fuse and decode human signals (i.e., gestures and speech). Local wireless ad-hoc networking between the human and the robots takes care of reliable transmission of the processed inputs. The full setup of the system and the robots that have been used for the experiments are shown in Figure 1 and described in detail in the following sections.

In practice, during interaction, we attempt to make the robots *as passive as possible*. This approach is determined by the need to make interaction robust to different external conditions, such as high variability of illumination and background acoustic noise, and leave as much autonomy as possible to the agents. For the success of SAR missions it is important that robots use their cameras for primary tasks, not to keep watching humans to catch possible command gestures. In this way, we can potentially achieve much

[1] Boris Gromov and Luca M. Gambardella are with Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland, {boris,luca}@idsia.ch

[2] Gianni Di Caro is with Department of Computer Science, Carnegie Mellon University (CMU), gdicaro@cmu.edu.

This work was partially supported by the Swiss National Science Foundation (SNSF) through the National Centre of Competence in Research (NCCR) Robotics.

higher robustness and reliability compared to most of the existing approaches, in which it is the task of the robots to remotely detect, decode, and classify the multi-modal signals issued by the human. For instance, this is the typical approach in popular vision-based human-robot and human-multirobot interaction [8, 9, 10]. The main drawback of using vision is that it may easily fail in the challenging and varying environmental conditions of SAR scenarios. However, whenever it is the human who controls position of the camera some of these problems can be circumvented.

The *contributions of the work* can be summarized as follows. (i) A fully working human-wearable system for interaction with multiple robots, based on human-centered relative localization and robot sensing, and fusion of speech, arm and hand gestures. (ii) Use of each modality for different purposes and fusion for their mutual confirmation. (iii) A set of interaction patterns for: individual and group robot selection, autonomous group motion towards spatial entities (direction, objects, locations), human-aided group motion. (iv) Robot group strategies for mixed-initiative classification of inconsistent combination of modalities for self-organized group recruitment during interaction.

II. Interacting with multiple robots: the need for multi-modality

The interface that we have developed makes use of multiple interaction modalities–speech, vision, arm and hand gestures–to generate commands for the robots. However, one might wonder if such a multi-modal interface is really necessary for the SAR applications we are considering, if, for instance, only basic commands need to be issued to the robots.

The answer is in the fact that *multiple robots* are expected to be concurrently deployed and locally interact with humans on the field. Compared to the single robot case, a number of complicating issues arise in this context, that can be robustly dealt with using the fusion of different modalities.

First, when a human wants to address a specific robot or group of robots, to express for instance a command like "Robot XYZ, go to room A and search for survivors", he/she needs to either name or point to the robots, or do both actions (and it is also required that the robot knows where room A is). However, when a multitude of robots is present in the scene, it is not reasonable that the human rescuer knows or keeps in mind the names or IDs of the robots. While a hand-held interface like a tablet would provide a better way for robot selection, in SAR domain human agents need their hands to be free for rescue activities. It becomes apparent that the multi-robot scenario asks for different interaction modalities compared to the single robot case, in which there would be no ambiguity when referring to the robot. In particular, the desired result can be obtained by fusing information from pointing gestures and speech. For instance, the sentence "You, go to room A" associated to pointing to an individual robot would allow to specifically select that robot for the task of going to room A, even if the name/ID of the robot is not explicitly mentioned in the sentence. If it is, then this would just increase the robustness of the procedure. Moreover, the two modalities would confirm (or disconfirm) each other, increasing the overall reliability of the system. How gestures and speech are used is discussed in the next section.

The presence of multiple robots also creates additional problems when a specific spatial notion, such as a direction has to be conveyed (e.g., "Search there for survivors", where "there" is pointed using an arm gesture). In fact, spatial directions have to be correctly understood by all the selected robots, accounting for their different poses in the environment with respect to the human pointing to the desired direction. This issue is addressed with the relative localization procedure discussed in Section III-C, that combines arm gestures and vision modalities. Once a common frame of reference is set, then either gestures or speech, or both, can be used to express spatially-related notions.

III. Wearable multi-modal interface

The hardware components of the implemented wearable system is shown in the left part of Figure 1, which consists of *two Myo Armbands* from Thalmic Labs [11] (a), a *video camera* paired with a *laser pointer* (b), a *headset* (c), and a *single-board computer* (ODROID [12]) with network interfaces running ROS (d).

The two Myo Armbands are used to reconstruct arm configurations and to provide spatial 3D vectors of pointed directions, while a single Myo, located on the forearm, is used to capture the hand gestures. The headset is used to record human speech commands on the wearable computer to eventually acquire a textual representation of the command using speech recognition technology. The video camera paired with the laser pointer is used to acquire information about the robots and to implement *relative localization*—a fundamental component of our system that allows to represent arm and hand gestures in a common reference frame for human and the robots. Finally, the ODROID computer performs all necessary computations and provides network connectivity during the interaction.

A. Arm and hand gestures

Arm gestures are issued to represent spatial entities (e.g., directions), locations, objects, robots: the arm configuration defines a 3D spatial vector that can be passed to the robots for performing the desired spatially-related action (e.g., search for survivors in the indicated location). However, in order to do this, first, the arm configuration must be reconstructed, second, the human and the robots need to set a *common reference frame* for representing the 3D vector associated to the gesture (this second step is discussed in Section III-C). The first step is performed using the two Myo Armbands, based on a simplified kinematics model of the human arm: the upper arm (first link) is connected to the stationary shoulder with a 3-DOF ball joint and to the forearm (second link) with a 1-DOF hinge joint. Therefore, the arm configuration is fully described with four generalized coordinates. Data from the two Myo Armbands, each attached

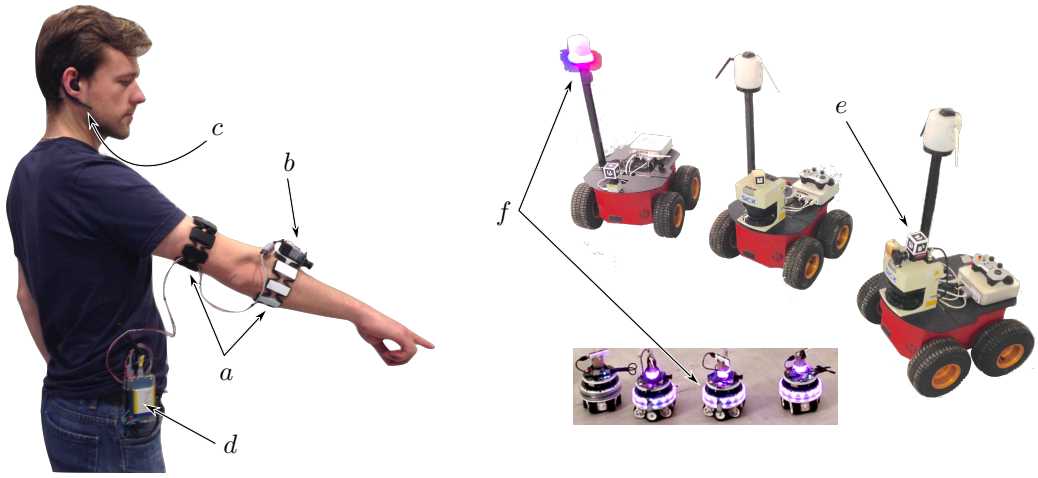

Fig. 1. The components of the wearable system for human-multirobot interaction, and the robots employed for the experiments. (Left) Two Myo Armbands (a) are used for arm configuration reconstruction and hand gestures recognition. A video camera (b) is used to identify robots, while a laser pointer helps the human to correctly point to the robots. A wireless headset (c) transmits voice data to the computer (d) for speech recognition. A wearable single-board ODROID computer (d) processes data from individual modalities, fuses them and sends high-level commands to the robots via wireless ad-hoc networking. (Right) Robots used for experiments. Three Pioneer P3AT and four Foot-bot robots used in experiments are shown. The Foot-bots use on-board LEDs (f) to provide feedback during interaction, while Pioneers have been equipped with special LED-rings (f) for the same purpose. Visual markers (e) are used to obtain ID and pose of the robots.

to the corresponding arm link, are used to acquire these coordinates. In general, a Myo Armband provides 9-DOF motion data and the measurements from 8 electromyography sensors (EMG). However, the onboard inertial measurement unit (IMU) also internally filters and fuses the motion data from its gyroscope, accelerometer and magnetometer into an accurate absolute 3D orientation anchored to the magnetic north. Thus, knowing the 3D orientation of each link in the global reference frame, the constraints imposed by the arm kinematics and sensors' positions on the arm, it is possible to calculate the arm configuration parameters. In particular, the three parameters of the ball joint are exactly the ones acquired from the upper arm Myo, while the hinge joint angle is calculated as a difference between upper and forearm Myo orientation.

We exploit the built-in capabilities of the Myo Armband for the *hand gestures* to locally adapt the motion of individual and group of robots. In practice, waving the hand right or left makes the associated robots to turn accordingly, based on their own direction of motion. Thus, the motion of individual or groups of robots can be locally adapted or manually guided.

1) Accuracy of gesture reconstruction: In order to evaluate the applicability of the chosen wearable sensors we have estimated the accuracy of the arm configuration reconstruction by comparing their measurements with a ground-truth provided by a commercial vision-based motion capture system (Optitrack). To do so, retroreflective markers were co-located with the Myo IMUs. The person performing the experiments was requested to perform periodic motions by swinging his forearm roughly in the range of $90°$. Both sets of measurements were processed by the same arm reconstruction algorithm and the two sets of elbow angle parameters were acquired (Figure 2). To quantify the error, the mean and standard deviation values of the signed difference between two sets were calculated. We observed the mean error $\mu = -11.3°$, while the standard deviation was $\sigma = 2.35°$. While the measured errors are not large, they can be significant for the far away objects or directions. In this case, additional manual controls can be applied to the motion of the robots on the way using the hand gestures. Alternatively, additional information can be supplied through the speech, such that the robots can match gestures and speech and compute the right direction. An example of this way of proceeding is discussed in the next sub-section.

Overall, the Myo Armbands showed reasonable performance as compared to the Optitrack-based ground truth and can be used as a part of the wearable interface.

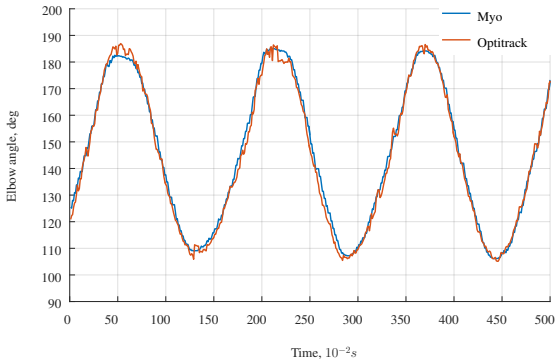

Fig. 2. Comparison of *Myo*- and *Optitrack*-based arm configuration reconstruction. To make the comparison more clear, the Myo plot was aligned by the amount of the mean error ($\mu = -11.3°$).

2) Perceptual errors when using deictic gestures: One of the interesting issues we faced during the initial trials of the arm reconstruction algorithm is a discrepancy between *human's perception* and the measurements. At first, it may

seem that the pointed object should always lay on the same line as the axis of human's arm. However, we identified that it is almost always *not* the case. In fact, a significant body of research in the past 40 years has addressed this issue of *systematic human perceptual error* when using deictic gestures [13, 14, 15]. Using the same experimental setting as before and the configuration shown in Figure 3, we have measured the pitch and yaw angular errors as a function of the distance from the pointed object. Numerical results are reported in Figure 4. It can be observed that the error in pitch is systematically larger and also relatively less stable than that in yaw. Both errors grow with the distance, with the error in pitch showing a larger rate of increase, and becoming quite large for relatively large distances. Unfortunately, the magnitude and direction of these errors largely fluctuate from person to person, such that it is quite difficult to devise a general filtering procedure for bringing them to zero. In order to overcome this issue, we have included a *laser pointer* to the system (Figure 1, b), which is used to help the human to visually correct the way he/she is pointing.

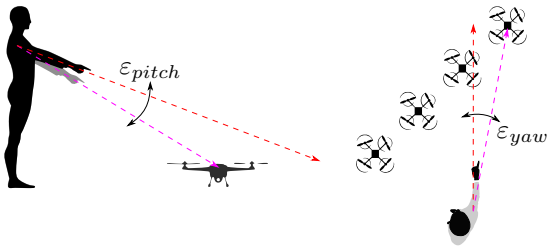

Fig. 3. Illustration of the systematic errors committed by humans when using pointing gestures to refer to spatial entities. Errors are both in pitch and yaw, and usually have different relative magnitudes.

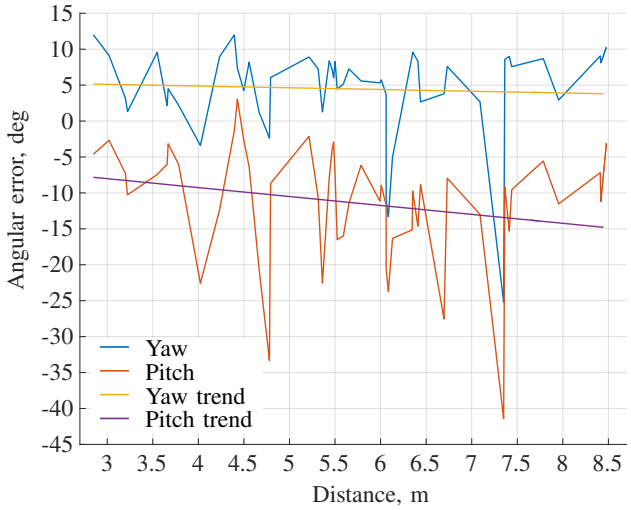

Fig. 4. Experimental measure of the systematic pointing errors committed by humans using deictic gestures. Errors are calculated for both pitch and yaw as a measure of the distance from the pointed object.

B. Speech sentences and multi-modal fusion

While the deictic gestures are very efficient to describe directions, locations and spatial entities, they are not sufficient to communicate semantically complex notions and requests. For instance, the requests may include information about the properties of the objects, like color or size; spatial relations, e.g. "on the left", "behind"; or even temporal notions. Attempt to use gestures alone would require quite a clumsy gesture vocabulary, or, indeed, the use of a specialized sign language. It is apparent that the requirement to learn such a language would be impractical during a SAR mission. Also the decoding of a full sentence would be quite hard to perform in a robust way. Conversely, these types of complex sentences can be naturally expressed using *verbal communication*. Thus, we use speech to both express notions that could hardly be expressed using gestures and to reinforce and support the gestures themselves. That is, speech is used for *modality fusion*: a sentence may refer to other modalities using direction and/or selection semantics (e.g. "there", "you"). For instance, while using a gesture a direction is pointed to, a sentence such as "move to the indicated direction" uttered at the same time can support the gesture and allow to express a more complex request.

The audio speech data is captured by a headset (Figure 1, c) and stored on the ODROID computer (Figure 1, d). In turn, in the current implementation, the embedded computer sends the data to the cloud-based *Microsoft Speech Recognition* services and obtains the textual representation of a command. Once the text is received back, it is matched against a list of predefined command templates and the associated entities are extracted. At the moment, the command templates are limited to basic operations that could be performed during a SAR operation, but can be expanded with more sophisticated commands or can be integrated with natural language processing engines.

Currently the system can process information regarding robot's ID, distance to be traveled, actions to be performed, directions or locations, position, color of objects, time allocated to perform tasks. For instance, the human user can utter a sentence like "Robot A, you, and you, move to this direction and search for red bins for 10 minutes" and accompany it with the appropriate gestures to first select the robots and then point to the desired direction. The uttered sentence is transformed into text using the Microsoft services, then all the relevant action elements are extracted, and finally the individual commands are sent in clear to the selected robots using local wireless networking.

To fuse the two modalities the timestamped streams of arm configuration messages and of the recognized speech commands are stored in a local cache. The timestamps of the beginning and the end of the utterance are preserved from the original audio stream and then used to extract the corresponding slice of continuous arm configuration data, which, in turn, is filtered and utilized to estimate the pointed direction or location with respect to the user. The fusion is triggered automatically based on the type of speech command, i.e. when the sentence contains spatial information.

Multi-modal information fusion is also used to detect inconsistencies and provide local feedback: whenever spatial

information is contained in the decoded sentence, the elbow angle (hinge joint) and the shoulder orientation (ball joint) are matched against predefined margins to check whether the arm configuration is compatible with the provisioning of spatial information or not. For instance, if the decoded speech sentence "Go to that door" is not accompanied by a pointing gesture, an error is raised up and communicated to the human through a vibration of the Myo Armbands and by vocal message to the headset.

C. Vision-based robot identification and relative localization

In order to interact with a specific robot or a subset of robots from a group, first they have to be identified and selected. Then, if the request or command from the human involves a spatial entity, a *relative localization* has to be performed which essentially sets a *common reference frame* between the human and the robots.

Our system can perform relative localization on the spot by combining the reconstructed arm configuration and a *vision-based* recognition of robot poses. The human starts the process by pointing to the desired robot to be identified and selected. Let's say this is robot *1*. The camera attached to the Myo Armband on the forearm (Figure 1, b) is used to detect the multi-faceted visual marker attached to the robot (Figure 1, e), in order to obtain its ID and estimate the pose with respect to the camera frame[1]. As noted in Section III-A.2, the laser pointer paired with the camera allows the human to align the arm in the direction to the robot. Using the reconstructed arm configuration, the full 6-DOF pose of robot *1* is computed with respect to the human footprint. This information is then stored both on the wearable system (Figure 1d), and transferred to the available robots via local ad-hoc wireless networking. Each time the process is repeated for another robot *i*, its relative pose is also computed in human's frame and therefore can be also bound to robot *1* coordinate frame and vice versa. Exploiting coordinates' transformation properties, a tree of coordinate systems rooted in the human reference frame is created. The tree can also be dynamically maintained by the robots performing on-board dead-reckoning and/or using additional localization measures. Robots can share their own coordinate updates with each other using ad-hoc networking, dynamically maintaining the systems coordinate transformation tree. Updates can also be performed by the human repeating the above localization procedure.

The entire process is depicted in Figure 5, where the coordinate transformation trees associated with individual robots are represented by the *blue directed arcs*, while the *red arcs* constitute the transformation tree acquired with the help of the camera.

IV. Multi-robot side: multi-modal coordinated feedback

As it has been discussed in Section II, the presence of multiple robots brings additional challenges to the design

[1]In the current implementation we use the *AR Track Alvar* ROS-package with bundled cubic markers.

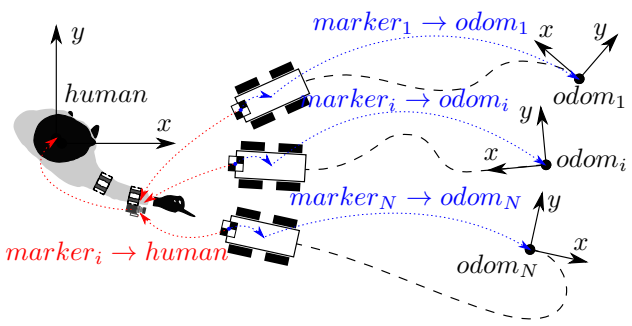

Fig. 5. Relative localization scheme. *Checker boxes* depict visual markers. *Colored dashed arcs* constitute the coordinates transformation trees (TF-trees) with the arrows pointing to corresponding parent frames. *Blue arcs* are the robots' individual TF-trees rooted in their odometry frames. *Red arcs* are the transformations acquired during the localization process and rooted in the human frame. The *long-dashed curves* are the robots' traveled paths.

of interaction modalities and requires special care for the *feedback* provided by the robots.

At the current stage of development, we are focusing on the fact that the human needs to get some feedback after selecting one or more robots and issuing a request. The feedback should clarify that the robots have been correctly selected and that the request is going to be correctly executed. Although the robots are almost passive during the interaction, things can still go wrong, and the human needs to promptly intervene if this is the case.

When multiple robots are in the scene, feedback needs to be provided in a coordinated way, otherwise the resulting effect could be extremely confusing. For instance, if all selected robots would individually reply to human's input by a synthetic voice message (e.g., "Robot XYZ1 acknowledges the request for going searching in room B"), the net effect would be an overlapping stream of voices. If 10 robots have been selected, this would just result in big noise for the human. Therefore, we have developed *coordinated multi-robot strategies* for providing informative feedback to the human.

In the current implementation, robots provide an audio-visual feedback which is both fully informative and acceptable in terms of sensory and cognitive load to the human. To give a visual feedback, the robots are equipped with controllable RGB LED-rings. Once a robot or a group of robots are selected, they set their LEDs to a predefined color. The confirmation messages are sent via local ad-hoc network to the human wearable PC, which aggregates the acknowledgments and generates a synthetic voice feedback. Similarly, after accepting a command, the robots trigger animated color patterns, while the voice feedback additionally confirms the action to be performed.

V. The system at work: Interaction patterns

Based on the technologies described in the previous sections, a number of *interaction patterns* between human agents and multi-robot systems have been designed with the

goal of addressing typical useful situations in SAR scenario[2].

a) Specific selection and localization of robots: To start interacting with a robot or a group of robots, the user has to point to the visual marker on one of them and use the addressing voice command "You". The laser pointer mounted on the forearm helps to correctly point to the visual marker and acquire robot's ID. In this way, a *single* robot is selected. To visually confirm the selection, the robot turns on its on-board LEDs (and, if the voice option is On, a vocal confirmation is also provided). In background, the selection process is accompanied by the relative localization process, that allows to put the robots into a common reference frame with the user. To localize *additional robots*, the voice command "Start localization" can be issued. Following it, the user can point to any additional robot and say "Localize" to confirm the selection. Once the process is done, the robots' local coordinate transformation trees are be joined together via the root tree node at human's current position. Moreover,

b) Generic group robot selection: When it is not important which particular robots to use for a task, a group of robots can be conveniently selected with a *seeded selection* approach, that realizes a form of mixed initiative interaction, relying on robots' decision autonomy. In fact, the user has to point to one of the robots, the *seed*, and say "You and another N", where N can be any number of robots. The seed, in turn, communicates (via ad hoc networking) with other localized robots and recruits the N required ones based on some appropriate criteria (e.g., closeness, battery status, current engagement, skills).

c) Selecting a specific sub-set of robots: To select a number of specific robots from the pool, a *cherry-picking* process can be performed by quickly pointing and addressing the robots with the voice commands "You", "you", ..."and you". While doing this, as before, in background the relative localization process is active to create the coordinate transformation tree.

d) Commanding for useful tasks: Once selection is done, the selected robots can be controlled with various commands including *spatial components*, specified through arm gestures. At the moment the system allows to express requests like "Go there", "Move N meters in this direction", "Search for survivors for 20 minutes", "Go to the red bin", and so on.

e) Discrete and continuous manual control: Due to accumulation of pointing and localization errors, pointed locations can be sometimes imprecise. To intervene and adjust final destination of a robot or a group we utilize manual control based on *discrete* hand- or *continuous* arm-gestures. To switch to manual control, the user hold the arm along the body, with the elbow bent roughly at 90°. The discrete hand gestures provided by Myo, i.e. *Wave In*, *Wave Out*, *Fist*, *Fingers Spread*, are directly mapped to the *turn left*, *turn right*, *accelerate*, *decelerate* motion commands of the robot. The continuous control is *mimicking a joystick* and performed by holding the fist while displacing the hand from initial point in any direction on the virtual plain. In this case, moving the fist forward sends positive forward velocity commands to the robot, while moving the fist on the sides makes the robot turning to corresponding direction. The velocities commanded to the robot are proportional to the displacement. Implemented manual control is exposed to ROS with `sensor_msgs/Joy` interface, where discrete commands are mapped to buttons and continuous to the axes, therefore virtually any robot supporting this interface can be controlled with our virtual joystick.

[2]A subset of these patterns is shown in an annotated companion video accessible at the https://www.dropbox.com/s/1yimezfdqlrsdi0/wearable-hri.mp4?dl=0. The video was previously put on display at the *Fielded Multi-robot Systems Operating on Land, Sea, and Air* workshop (ICRA '16). In the implementation shown in the video, the robots are localized by means of an optical tracker.

REFERENCES

[1] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 16–21, Jan. 2001.

[2] M. A. Hearst, "Trends & controversies: Mixed-initiative interaction," *IEEE Intelligent Systems*, vol. 14, no. 5, pp. 14–23, 1999.

[3] A. Finzi, J. Cacace, R. Caccavale, and V. Lippiello, "Attentional multimodal interface for multi-drone search in the alps," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2016.

[4] T. Kollar, A. Vedantham, C. Sobel, C. Chang, V. Perera, and M. Veloso, *A Multi-modal Approach for Natural Human-Robot Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 458–467.

[5] S. Nagavalli, S.-Y. Chien, M. Lewis, N. Chakraborty, and K. Sycara, "Bounds of neglect benevolence in input timing for human interaction with robotic swarms," in *ACM/IEEE International Conference on Human-Robot Interaction*, March 2015, pp. 197–204.

[6] J. Xavier and U. Nunes, "Interfacing with multiple robots using environmental awareness from a multi-modal HRI," in *Proc. of Conference on Mobile Robots and Competitions (ROBOTICA)*, 2007.

[7] G. Jones, N. Berthouze, R. Bielski, and S. Julier, "Towards a situated, multimodal interface for multiple uav control," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2010, pp. 1739–1744.

[8] S. Pourmehr, V. Monajjemi, J. Wawerla, R. Vaughan, and G. Mori, "A robust integrated system for selecting and commanding multiple mobile robots," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

[9] J. Alonso-Mora, S. H. Lohaus, P. Leemann, R. Siegwart, and P. Beardsley, "Gesture based human - robot swarm interaction applied to an interactive display," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[10] 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)*, 2014.

[11] Myo Armband. Official web-page. [Online]. Available: https://www.myo.com/

[12] ODROID-C1+. HardKernel official web-page. [Online]. Available: http://www.hardkernel.com/main/products/prdt_info.php?g_code=G143703355573

[13] J. Foley and R. Held, "Visually directed pointing as a function of target distance, direction and available cues," *Perception & Psychophysics*, vol. 12, no. 3, 1972.

[14] D. Droeschel, J. Stckler, and S. Behnke, "Learning to interpret pointing gestures with a time-of-flight camera," in *Proc. of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 481–488.

[15] S. Mayer, K. Wolf, S. Schneegass, and N. Henze, "Modeling distant pointing for compensating systematic displacements," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, 2015, pp. 4165–4168.