Distributed Gesture Recognition by a Robot Swarm

Shelly Bagchi[1], Andrea Thomaz[2], and Magnus Egerstedt[3]

***Abstract*— In this paper we present a framework for controlling robot swarms through gestures. The use of gestures allows for a human to control a swarm without any additional equipment such as a joystick or computer. Flag semaphore was chosen as the gesture set due to its proven use in fields such as air traffic control. Each robot in the swarm participates in recognition without any adjustment, providing robustness to viewing angles and avoiding disruption of ongoing tasks. Overall recognition can be determined at a central node or leader robot, with all swarm members transmitting a vote based on their individual recognition. Several voting methods were examined, and the final method makes use of each robot's confidence level in their vote, as well as weighting votes subject to any occlusions that may be present. The developed control scheme is scalable to large swarms and shows the potential to be easily learned by novice users.**

I. Introduction

Multi-agent robotics has proven useful for collaborative tasks such as mapping or search & rescue. In particular, swarm robotics takes inspiration from biological swarms; each agent is fairly simple on its own, but coordinating large groups of agents can accomplish larger, more difficult tasks. Additionally, robot swarms should be robust to failures and scalable to both small and large sizes [1].

In practical applications, an element of human control is often necessary. However, control of individual robots becomes difficult when dealing with large swarms; performance has been shown to decrease as swarm size increases [2]. Rather, it is desirable for the swarm to be controlled as a group, whether through their collective behavior or in other ways that take advantage of the distributed nature of the swarm.

The simplicity of swarm robots compels the need for similar simplicity in control scheme. Often, robot systems require expert users who have been trained extensively over time in the particular interface, and this knowledge may not transfer well to any other systems. Interfaces such as joysticks have been explored, but can be difficult to use without prior experience. Our aim in this work is to examine the possibility of control without the use of any interface, but rather through the use of gestures.

Flag semaphore is a telegraphy system developed for visual, long-distance communication. It is often used in air traffic control or for emergency communication between ships. The system consists of a specific, static pose for each of the 26 English alphabetical characters, plus four additional signals. Poses are based on arm position, where each arm can be extended at any interval of 45-degrees. For visibility over long distances, the signals are commonly performed with a flag or lighted stick held in each hand, however no props are specifically necessary to use the semaphore system.

This method is ideal for control of a robot swarm by a novice user, due to its simplicity and lack of specialized equipment. Additionally, it shows potential for future work including long-distance swarm control. An example of the types of gestures used in the flag semaphore system is shown in Figure 1, and the full set can be found in [3].

Flag semaphore as a method of human-robot interaction was briefly explored by Nguyen et al. however their work relied on 3D camera data and recognition was done by only one robot [4]. Our approach simplifies the recognition process, requiring only a 2D camera, and employs distributed recognition over a swarm of robots.

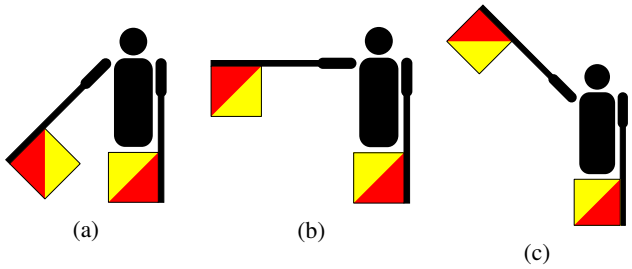

Fig. 1: Flag Semaphore signals for the letters A, B, C [5].

II. Related Work

A. Swarm Robot Control Methods

B. Gesture Recognition

[In progress]

III. Technical Approach

This section first outlines the recognition process for an individual robot, which will then be extended to a swarm through a voting process.

A. Gesture Recognition Process

As previously mentioned, the gestures used were taken from the established English-language flag semaphore system. In order to keep the gesture set consistent, seven gestures were removed from the set: H/8, I/9, O, W, X, Z, and Error. These gestures were dissimilar from the others due to having two arms on one side of the body as opposed to one arm on either side. As one of the aims of swarm

[1]Institute of Robotics & Intelligent Machines, Georgia Institute of Technology, Atlanta, GA, USA. shelly.bagchi@gatech.edu
[2]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA. athomaz@cc.gatech.edu
[3]School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA, USA. magnus.egerstedt@ece.gatech.edu

robotics is for each robot to have fairly simple hardware, it is assumed that each robot would only have access to a small monocular camera (i.e. no depth information is available). Each gesture is static, so a single image is the only input into the recognition process.

The first step in gesture recognition is to determine whether the robot is able to see the controlling human. This is done using the histograms of oriented gradient (HOG) descriptors method developed by Dalal et al. [6], which makes use of a linear SVM that has been trained with a large number of human images. This method has been shown to have good performance detecting humans with a variety of poses and backgrounds, and the OpenCV implementation was used during experimentation. If a human is not found using the HOG descriptors, the robot will not proceed with the remaining recognition steps, and will not contribute towards a swarm consensus. If a human is found, the area of the image containing a human is used for the rest of the recognition process.

Initially, flag detection was considered to obtain features for recognition. However, after observing the results from the HOG descriptors, it was apparent that the flags were nearly always outside the area returned as containing a human. Due to this and in order to provide more flexibility to the controlling human, arm detection is used instead. Because no depth information is available, skin color thresholding is employed for this process. This does result in a few caveats: The controlling human must be wearing short sleeves, and a variety of skin tones must be present during training to ensure recognition.

To obtain features for classification after thresholding, first contours and then the rotated bounding boxes for each contour are found. Some filtering is necessary at this point in order to remove background noise as well as other areas of skin (e.g./ face or legs). This is accomplished by examining the aspect ratio and location of each bounding box, as well as the relative sizes - anything much smaller than the largest box found is likely noise. Additionally, there can be a maximum of two arms found, so only the two largest boxes are kept after filtering. The gestures are dependent on the angle of each arm, thus the angle of the filtered bounding boxes are used as features for classification.

Classification was done using a multi-class Support Vector Machine (SVM). Separate classifiers were used for data points where only one arm was visible versus those where both arms were detected. This was due to the fact that multiple gestures could have the same left-arm angle, for example, implying that left-arm data collected from those gestures should be classified together. Data showing both arms, however, should be able to yield one specific gesture rather than a subset. A 10-fold cross validation was done to obtain initial performance, and then a specific subset of points were held out to simulate swarms as described in the next section.

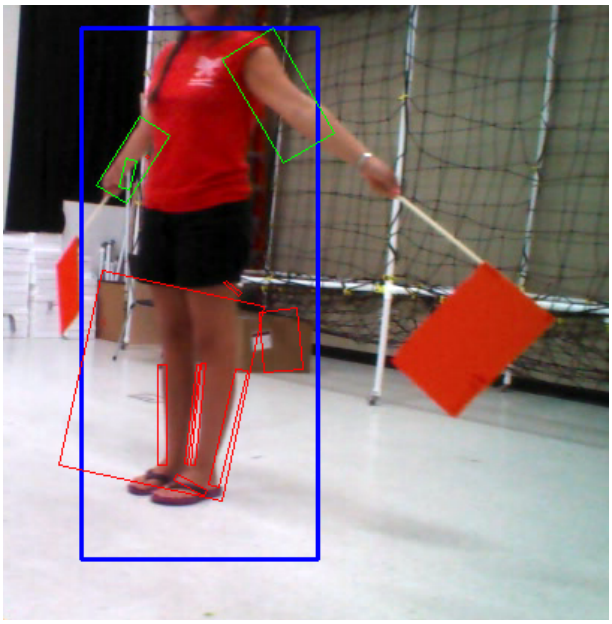

Fig. 2: Example of arm recognition. Small or overlapping bounding boxes are filtered in post-processing.

B. Swarm Consensus

In order for the swarm to come to a collective decision, voting was chosen as the method of combining information. This simplifies communication and distributes the required processing, rather than having each robot transmit an image to a central node or leader robot. Four different types of voting were implemented and compared to obtain recognition results for each swarm as a whole.

1) Each robot receives one vote, so that equal consideration is given to all members of the swarm. Robots that only see one arm vote for multiple gestures equally (based on prior knowledge of the gesture set). This method is used as a baseline.
2) Each robot votes based on its confidence in its recognition result. This is obtained directly from the classifier as the score, or the distance from the separator for the chosen gesture. A higher score implies more confidence in the classification result. As before, one-arm results contribute equally to multiple gestures.
3) Voting is again based on confidence, but robots able to see both arms are weighted twice as much as others.
4) Voting is based on confidence, but robots with a view of one arm have their vote weighted lower; in effect, their individual votes are divided between all possible gestures. For example, if three gestures can have the left arm at 90 degrees, a robot that sees only the left arm will contribute one-third of its confidence vote towards all three gestures.

IV. Simulation & Experimental Setup

A. Data Collection

A Nao robot is used to collect experimental data. The Nao contains two 2D color cameras in its forehead; the

top camera is used here to better capture the controller's entire body. The camera is capable of capturing images at up to 1280x960 resolution, or recording video at up to 30 frames per second. To obtain large set of training data, video was taken with the Nao moving in a semi-circle around the human performing a gesture; this was in order to simulate different possible viewing angles. A visualization can be seen in Figure 3. Video is recorded in 640x480 resolution at 20 frames per second. The same process is repeated at five different distances away from the human, from 1 to 3 meters at 0.5 meter intervals. Then, each frame of each video is considered as a separate data point for training and testing purposes.

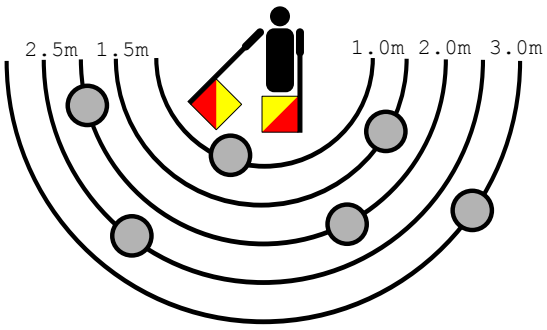

Fig. 3: Experimental Setup.

B. Simulation of Swarms

As previously mentioned, data was collected in video form covering the area in front of the human controller. By going through the recognition process for each video frame, about 600 to 800 data points per video (3000 to 4000 per gesture) were obtained. Each data point can be considered a simulated robot's view of the controller. Therefore, frames where no human was found or no arms were detected were not saved, since that particular robot would not participate in voting or contribute to the swarm's overall recognition.

In order to simulate swarms, evenly-spaced points were sampled from each set of data points and held out during the training phase of recognition. Then, swarms of a specific size were constructed by taking all possible combinations from the sampled subset of locations. For example, 6 viewing angles from 5 distances yields 30 locations, and all combinations of 4-robot swarms within those locations results in 27,405 simulated swarms. This is done in order to trim the data set and avoid having to simulate millions of swarms; it also has the advantage of enabling the classifier training to occur only once, speeding up the process. Then, within each swarm, the individual robots go through the recognition process separately before coming up with a vote for their result. Votes can be compiled on either a leader robot or a central node, dependent on the system infrastructure.

V. Results

[Recognition results - stats, learning curve, confusion matrix]

Figure 4 shows a visual representation of 30 robot locations spread out around the controller. Recognition confidence is represented by color intensity, where blue shows a correct vote and red an incorrect vote. Here, results have been collated over all gestures. It was initially hypothesized that robots with "bad" viewing angles - perhaps too close to the controller or too far on either side - would have low-confidence or incorrect recognition results. However, from Figure 4 there does not seem to be any clear pattern which would validate that hypothesis.

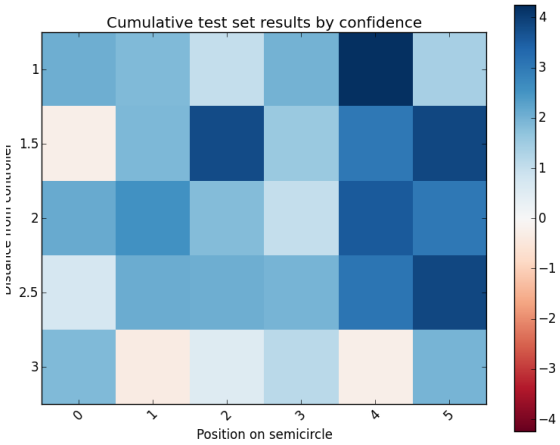

Fig. 4: Matrix showing recognition confidence relative to physical location.

Figure 5 shows an example of voting results for all possible combinations of 3-robot swarms. In this particular example, the gesture set was limited to nine gestures for testing purposes: B, D, G, M, A, C, F, J, N (in the same order as in Figure 5). These gestures were used due to the overlap in single-arm positions; for example, M, A, and N all have the same left-arm angle.

Within each swarm, each individual robot goes through the recognition process and contributes a vote towards the swarm's consensus, as described previously. Votes are collected and the gesture with the most votes is taken as the swarm's overall decision. Then, each swarm's deciding vote is normalized and averaged to show how many swarms chose each gesture. The results show that even for swarms as small as three robots, the majority were able to detect the correct gesture by a clear margin. Figure 5 in particular shows the swarms participating in recognition for gesture 8 (N), however the results for other gestures display the same trend.

Figure 5 also shows a comparison of the four voting methods discussed earlier. Although all modes show the same trend with very similar performance, mode 4 is slightly superior due to a higher confidence shown for the correct result as well as a lower confidence for any incorrect votes. This leads to the conclusion that a confidence threshold may be beneficial in order to avoid errors in situations where no

members of the swarm have a good view of the controller, or perhaps situations where the controller is not clearly displaying a distinct gesture.

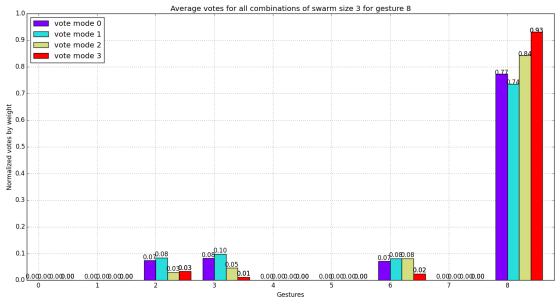

Fig. 5: Example of voting results for all 3-robot swarms, participating in recognition for gesture 8 (N).

VI. Future Work

An extension of this method may be to allow communication between individual robots during the recognition process. If there is a pair of robots where one can see the gesture's left arm and the other can see the right, they may be able to come up with a more confident vote than either separately. However, this would require a more complex communication scheme within the swarm.

The use of flag semaphore in long-distance human communication naturally results in the question of whether this method allows for long-distance control of a swarm. This could be useful for lessening reliance on wireless communication to remote-control a swarm.

Another possible application for future work is task learning for a swarm. Individual gestures may be mapped to "primitive" swarm actions, such as a specific formation or motion. Multiple gestures could then be performed sequentially to teach the swarm a larger task as a chain of primitive actions.

References

[1] E. Sahin, "Swarm robotics: From sources of inspiration to domains of application.," *SWARM ROBOTICS*, vol. 3342, pp. 10 – 20, 2005.

[2] J. Y. Chen and M. J. Barnes, "Humanagent teaming for multirobot control: A review of human factors issues," in *Human-Machine Systems, 2014. IEEE Transactions on*, vol. 44, IEEE, 2014.

[3] U. N. S. Cadets, "Flags, the nato phonetic alphabet, and morse code." (Visited on 08/13/2015).

[4] N. Nguyen-Duc-Thanh, D. Stonier, S. Lee, and D.-H. Kim, "A new approach for human-robot interaction using human body language," in *Convergence and Hybrid Information Technology*, pp. 762–769, Springer, 2011.

[5] "Semaphore alpha - wikimedia commons." (Visited on 08/13/2015).

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.