Control of a Robot Swarm via Static Gestures

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, search & rescue, and [FINISH THIS]

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 alphabetical characters, plus four additional special signals. Each pose is 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.

FIGURE: FLAG SEMAPHORE EXAMPLE

II. Related Work

[In progress]

III. Gesture Recognition

This section will outline the recognition process for an individual robot, which will be extended to a swarm in the next section.

A. Problem Setup

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: Error, H/8, I/9, O, W, X, and Z. These gestures were dissimilar from the others due to not being static or having two arms on one side of the body as opposed to one arm on either side.

As one of the aims of swarm robotics is for each robot to have fairly simple hardware, it was assumed that each robot would only have access to a small monocular camera (i.e. no depth information is available). For this reason, a Nao robot was used to collect experimental data. Each gesture is static, so a single image is the only input into the recognition process. 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. The same process was 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 was considered as a separate data point for training and testing purposes.

FIGURE: EXPERIMENTAL SETUP

B. 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. [1], 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 was chosen instead. Because no depth information was available, skin color thresholding is employed for this process. This does result in a few caveats: The controlling human must be wearing short

[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

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 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.

FIGURE: RECOGNITION SCREENSHOT

IV. Swarm Consensus

A. 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. Frames where no human was found or no arms were detected were not considered, since if that data was obtained from a swarm member, 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 results in 27,405 simulated swarms. 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.

B. Voting Methods

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.

MORE DISCUSSION HERE, OR IN RESULTS?

V. Results

- recognition results? cross validation / recall, precision, etc.

FIGURE: POSITION-BASED CONFIDENCE MATRIX
FIGURE: VOTING RESULTS

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] Navneet Dalal and Bill Triggs. Histogram of oriented gradients for human detection. CVPR 2005.