Chapter 8

Diversity in Cooperative Movement

This chapter describes experiments involving teams of simulated robots learning a cooperative navigation task. The agents select from one of several formation strategies (including no formation at all) as they move across obstacle-strewn terrain. At issue is whether the agents benefit from formation behavior, and if so, whether teams perform best when all agents choose the same behavior. Teams using fixed homogeneous strategies are also evaluated for comparison.

Each robot is provided a common set of cooperative movement skills (motor schema-based behavioral assemblages) from which it learns to select using reinforcement learning. The agents learn individually to activate a particular behavioral assemblage given a reward signal. In contrast to the domains examined in earlier chapters, it is not necessary for the agents to learn a sequence of behaviors to succeed in this task. The agents learn which one of four cooperative movement behaviors to

Figure 8.1: A team of four robotic scout vehicles manufactured for DARPA's Demo II project. These robots were the target platform for earlier research in robot formations. Photograph courtesy of Lockheed-Martin.

activate; the same behavior is active for the entire trial.

The experiments in navigation simulations evaluate the agents in terms of *performance*, *policy convergence*, and *behavioral diversity*. As in foraging and soccer experiments (Chapters 6 and 7) the results show that robots will diversify by choosing heterogeneous behaviors. An interesting result is that teams using diverse movement behaviors perform better than homogeneous teams. In contrast to the results in other tasks, however, the degree of diversification does not depend on the reward structure. Navigating teams learn to perform equally well using local or global rewards.

The chapter proceeds with a discussion of the task, behaviors for accomplishing it and a description of the experimental results. Experiments follow the methodology introduced in Chapter 3.

8.1 Background and related work

The development of this task domain and the behaviors designed for it are extensions of previous research conducted in the Mobile Robot Laboratory at Georgia Tech [BA95b, BA99]. The earlier work was focused on developing behaviors for a team of robotic vehicles to be fielded as a scout unit by the U.S. Army (Figure 8.1). Formation is important in this and other military applications where sensor assets are limited. Formations allow individual team members to concentrate their sensors across a portion of the environment, while their partners cover the rest. Air Force fighter pilots for instance, direct their visual and radar search responsibilities depending on their position in a formation [For92]. Robotic scouts also benefit by directing their sensors in different areas to ensure full coverage [CGH96]. The approach is potentially applicable in many other domains such as search and rescue, agricultural coverage tasks and security patrols.

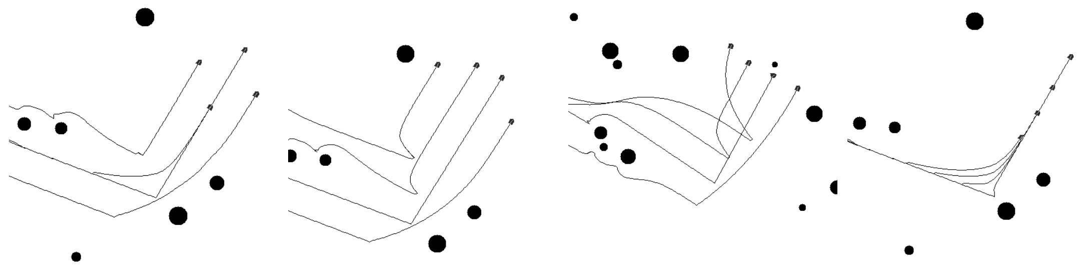

Figure 8.2: Four robots in leader-referenced *diamond, wedge, line* and *column* formations. These formation behaviors, developed in earlier research, are targeted for heterogeneous teams where each robot is assigned a specific position in formation.

Several formation strategies for scout robots were developed to enable the team to move cooperatively in military scenarios. In the scout domain the multi-robot team is heterogeneous because each agent is assigned a position in the formation according to an identification number. This is important in applications where one or more of the agents are dissimilar. In Army scout platoons for instance, the leader is not usually at the front of the formation, but in the middle, or to one side.

Important contributions of this earlier work include behaviors for four-robot *diamond*, *line*, *column*, and *wedge* formation types and a performance analysis of each formation type in turns and across obstacle-strewn terrain. Results from the earlier work are compared with the performance of the new behaviors presented in this chapter. The four formation types developed previously are illustrated in Figure 8.2.

The earlier strategy works well, but it is limited to formations with a specific number of robots. The location of each robot in each formation is predefined and formations are not easily scalable to larger numbers of agents. The expectation is that in large-scale homogeneous teams agents should automatically move to the closest appropriate location. To provide this capability, a new, scalable formation technique is introduced here. An example large-scale robot formation using the new technique is given in Figure 8.3.

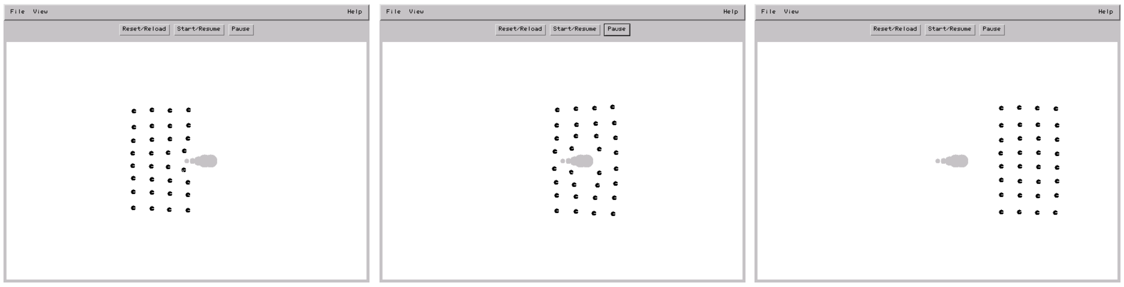

Figure 8.3: Large scale formation: 32 robots (black circles) moving from left to right in formation encounter an obstacle (grey object). These robots utilize the new scalable strategies introduced here. Sequence is from left to right.

The next section describes the task and experimental environment in detail. Following that, the new formation behaviors are introduced.

8.2 Task and performance metric

The task examined in these experiments is for a team of robots to move across a field as quickly as possible while avoiding collisions with obstacles and other robots. Performance is defined as

$$P = -t \tag{8.1}$$

where t is the time in milliseconds for the *entire* team of robots to move across the field. This is equivalent to the performance of the last agent to cross the field. Several other performance measures were considered, including the average time for all agents to complete the task and the time of the first robot to move across. The time for the last agent to complete the task was chosen because it indicates, to some degree, the extent of cooperation between the robots. Other measures might show improved performance when individual agents "abandon" their partners in an effort to cross the finish line more rapidly. Note however, that even though it may be advantageous for the robots to move in a group, this is not explicitly part of the performance measure.

In terms of the taxonomy introduced in Chapter 4 this task and performance metric have the following characteristics:

- `TIME_MIN` because the task must be completed in minimum time,

- `ROBOT_BASED` since performance is based on the location of the robots,
- `MOVEMENT_TO` because the robots must move to a location,
- `MULTI_AGENT` because the task implicitly requires all robots to complete the task,
- `SENSOR_LIM` since agents only have a short-range view of the environment (e.g. obstacles).

Figure 8.4 illustrates the JavaBots simulation environment used in the experiments. The field measures 20m by 60m. 30 obstacles, each 1m^2 in area, are distributed randomly about a 20 by 30 meter zone in the middle of the field (5% obstacle coverage). The robots are initialized on the left side of the field. They then navigate to the right side, through the obstacles to the finish line on the right. Timing stops when the last robot crosses the line.

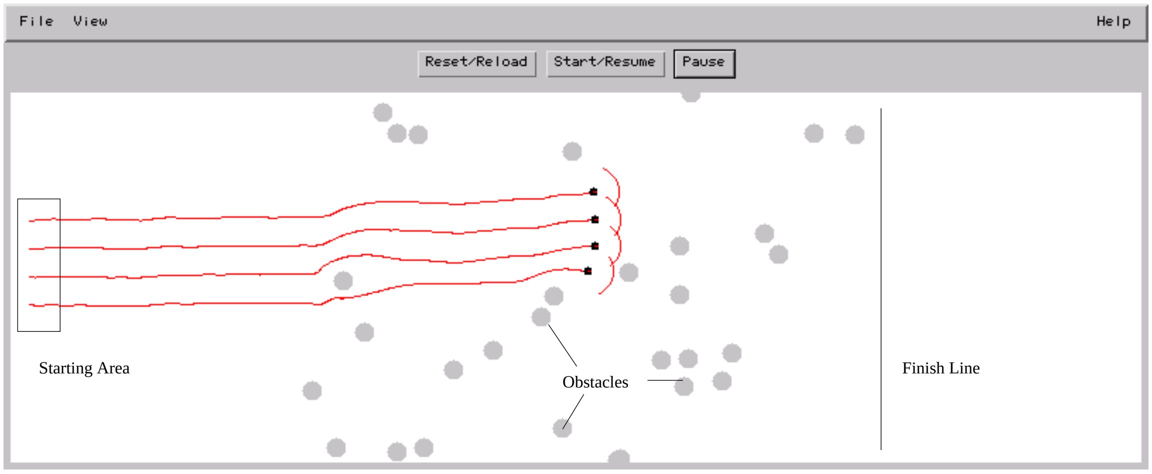

Figure 8.4: The simulation environment used in the experiments. Robots are initialized on the left. They navigate from left to right through the obstacles.

An example experimental run is illustrated in Figure 8.5. The agents are initialized line abreast on the left side of the field. This initial configuration was chosen because it ensures all robots are equidistant from the finish line. The first 20m of the field are clear of obstacles to enable the robots to settle into formation positions before encountering the obstacle field. After crossing the obstacle-free section the robots encounter a 30m long zone cluttered with hazards. As the figure shows, interaction with obstacles sometimes results in a rearrangement of the formation (the reason for this will become apparent as the behaviors are described in the next section).

Figure 8.5: Sequence of images from an experimental run with four robots programmed to use the *diamond* behavior. The top image illustrates how agents are initialized line abreast on the left side of the field. The agents settle into formation as they cross the obstacle-free area (second image). The robots regroup in a different arrangement after encountering an obstacle (third image). Finally (bottom) the team crosses the finish line.

Two aspects of the experimental setup should be considered when reviewing performance results. First, the arrangement of agents at the beginning of each run may bias the shape of the formation towards line abreast. Second, the measured time to complete the task includes the time taken for the agents to cross the initial, obstacle-free area. Thus overall performance is a combination of performance in obstacle-free and cluttered terrain.

Now consider how behaviors can be designed for this task.

8.3 Behavioral design

Formation maintenance is accomplished in two steps: first, a perceptual process, **detect_formation-position**, determines the robot's proper position in formation based on current sensor data; second, the motor process **maintain_formation**, generates motor commands to direct the robot toward the correct location. The motor schema paradigm enables the formation behavior to be simultaneously active in combination with other navigation behaviors.

The overall navigational strategy is similar to the approach developed in earlier research [BA95b]: several motor schemas, **move_to_goal**, **avoid_static_obstacles**, **avoid_robots** and **maintain_formation** implement the overall behavior for a robot to move to a goal location while avoiding obstacles, collisions with other robots and remaining in formation (a mathematical description of these motor schemas and the gains used in these experiments are provided in Appendix A). An additional background schema, **noise**, serves as a form of reactive "grease", dealing with some of the problems endemic to purely reactive navigational methods such as local maxima, minima and cyclic behavior [Ark89]. The key extension that distinguishes the new formation behaviors from previous work is the perceptual technique used to determine the proper formation position for each robot.

Instead of having each agent assigned to a particular position as in the previous approach, it may be advantageous to develop a more general technique. Design goals for the new formation strategy include

- **scalability:** the approach should easily scale to any number of agents,
- **locality:** the behaviors should depend only on the local sensors of each agent,
- **flexibility:** the behaviors should be flexible so as to support many formation shapes.

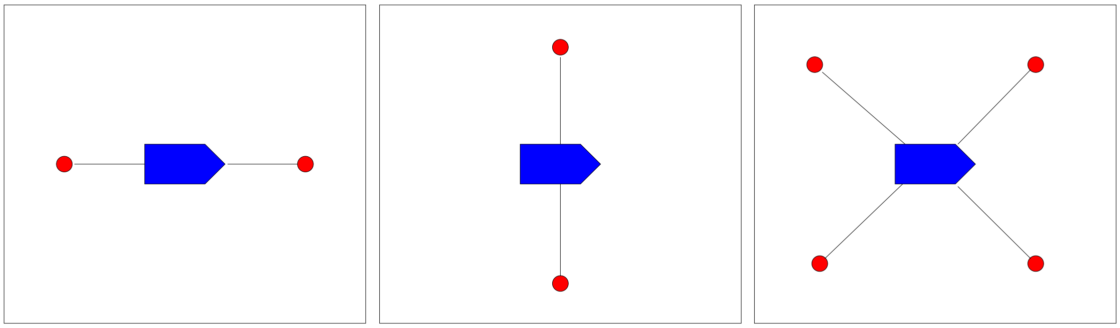

Figure 8.6: Attachment site geometries for different formations. From left to right: *column*, *line* and *diamond*. Robots are represented as five-sided polygons while attachment sites are shown with small circles.

This new strategy is based loosely on the way molecules form crystals. From the point of view of each robot in the group, every other robot has several local "attachment sites" other robots may be attracted to. Different formation shapes are created when different attachment site geometries are employed. Figure 8.6 illustrates the three attachment site geometries examined in this work. To determine a formation position each robot builds a list of potential attachment sites for all of the robots within sensor range based on the formation type it is using. An attractive vector is generated towards the closest site.

In addition to the motor schemas mentioned earlier, a low-gain attractive force, **move_to_unit_center**, is added to draw all of the robots together. As the team converges, the robots "snap" into position, and a regular geometric structure emerges. Example formations resulting from the integration of these behaviors are illustrated in Figure 8.7.

Note that for each attachment site geometry there are many potential robot team arrangements. It is also possible for interaction with obstacles to "unsnap" the formation into smaller sub-formations. In many cases however, the formations re-group after splitting around obstacles.

Performance of these behaviors are now examined in homogeneous teams of navigating robots.

8.4 Fixed formation strategies

As a baseline for comparison with learning teams, four fixed strategies were developed and evaluated. In each case, all the robots utilize the same attachment geom-

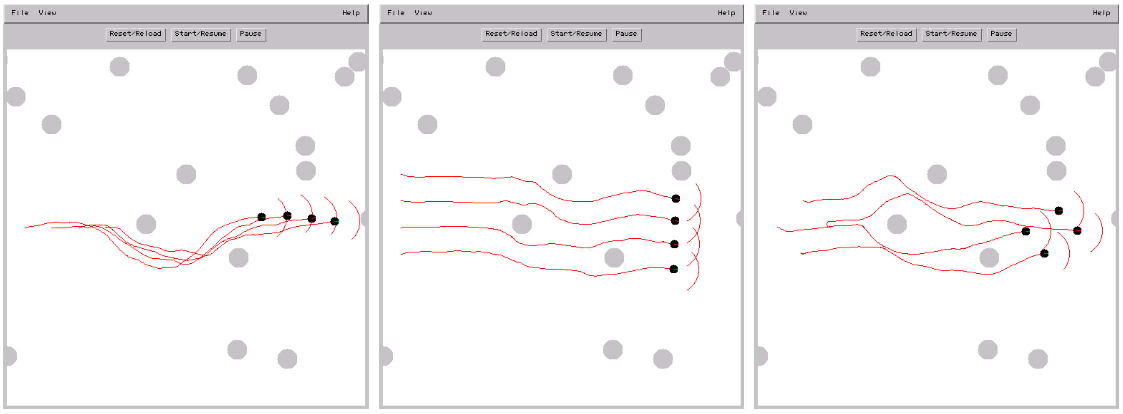

Figure 8.7: Example four-robot formations resulting from the use of different attachment site geometries. From left to right: *column*, *line* and *diamond*. In each of these short demonstration runs the robots were initialized in proper formation positions, experimental runs are over a longer course.

etry. Experiments were run with one to eight robots using *diamond*, *line*, *column* and *no_formation* geometries. The *no_formation* assemblage utilizes the same navigational behaviors and gains as in the other assemblages, except **maintain_formation** is not activated. The group of robots are still attracted to one another because the **move_to_unit_center** motor schema is activated.

Performance was evaluated by running each simulated robot team through five different randomly generated worlds 50 times. A total of 250 simulations were run for each number of robots for each formation geometry, or a total of 8000 trials overall. The average time for robots to complete the traverse is plotted for each strategy in Figure 8.8.

The relative performance of teams using *diamond*, *line* and *column* geometries mirrors similar results reported earlier [BA95b]. The earlier experiments in navigation across an obstacle field showed that *column* formation provides the best performance (in terms of path length). Even though performance is measured in these experiments using time instead of path length, the results are similar. The *column* geometry provides the best performance for navigation across the obstacle field. This is because the formation as a whole presents a smaller cross section to the obstacles as it moves across the field. For similar reasons, the *line* formation performs worst; it presents the broadest cross section.

In contrast to the performance of other strategies, the performance of teams using *no_formation* improves consistently as the number of robots increases. This

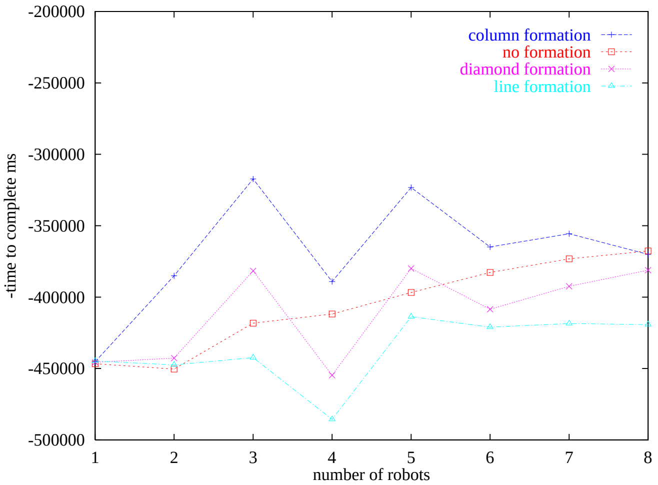

Figure 8.8: Average performance for fixed homogeneous teams.

is probably because in all strategies except *no_formation*, when a robot gets stuck, other robots are likely to remain near it and get stuck also. In the *no_formation* strategy, the low-gain **move_to_unit_center** behavior slows progress of the other agents, but it will not stop them. In addition the **move_to_unit_center** behavior provides the side-effect of pulling "stragglers" out of the areas they may be stuck in.

8.5 Learning cooperative movement strategies

Learning teams were developed using the same behavioral assemblages used in the fixed systems. This ensures that the performance of learning teams in comparison to the fixed teams is due only to differences in policy.

In contrast to the experiments in soccer and foraging, this task does not require a sequence of behaviors. Each agent selects a single behavior to follow for an entire trial, at which point it receives a reward. For the purposes of incorporating Q-learning, however, the problem can be viewed as a sequential task with one step. The Q-learner automatically tracks previous rewards to refine its choice of action for each trial.

8.5.1 Reinforcement functions for cooperative movement

The policy an agent learns will depend on the reward function used to train it. One objective of this research is to discover how *local* versus *global* reinforcement impacts the diversity and performance of learning teams. Global reinforcement refers to the case where the reinforcement signal is simultaneously delivered to all agents, while with local reinforcement each agent is rewarded individually. To that end, we consider two reinforcement functions for the learning robots. The local reinforcement function is:

$$R_{\text{local}} \quad = \quad -t \tag{8.2}$$

where t is the elapsed time in milliseconds from the start of the trial until the robot crosses the finish line. This effectively rewards minimum-time completion because shorter times result in a less negative reward. In terms of the reward taxonomy R_{local} is classified as an `INTERNAL_SOURCE`, `PERFORMANCE`, `IMMEDIATE`, `CONTINUOUS` and `LOCAL` reward.

The global reinforcement function is:

$$R_{\text{global}} \quad = \quad -t_{\text{team}} \tag{8.3}$$

where t_{team} is the time when the last agent on the team crosses the line. In terms of the taxonomy presented in Chapter 4 R_{global} is an `INTERNAL_SOURCE`, `PERFORMANCE`, `IMMEDIATE`, `CONTINUOUS` and `GLOBAL` reward function.

Experimental data were gathered by running thousands of trials and monitoring robot performance. The learning robots are evaluated on three criteria: task performance ($-t$), policy convergence, and diversity of behavior. For each type of reward and each number of robots (1 to 8), experiments were conducted in 5 randomly generated environments. In each environment the learning robots were initialized with Q-values set to random values between -1 and 1. The agents were then trained over 300 trials. Information on policy convergence and performance was recorded after each trial. The robots retain their learning set between trials. Overall, a total of 24,000 trials were run for the learning systems.

8.5.2 Task performance

Performance was evaluated for each number of robots for each reward type by averaging the results of the final 50 trials in each of the five experimental environments. Each data point therefore represents average performance in 250 trials. Performance for locally and globally rewarded teams is plotted in Figure 8.9.

The difference in performance between teams trained with local versus global rewards is not statistically significant. But both types of team out-perform the best homogeneous fixed strategy. This is interesting because it means the agents have discovered a better strategy than the homogeneous *column* formation for navigation across cluttered terrain. The *column* formation was shown in previous work ([BA95b]) and again in this research (Figure 8.8) to provide the most efficient (homogeneous) team navigation across cluttered terrain. Reasons for this result are examined in Section 8.6.

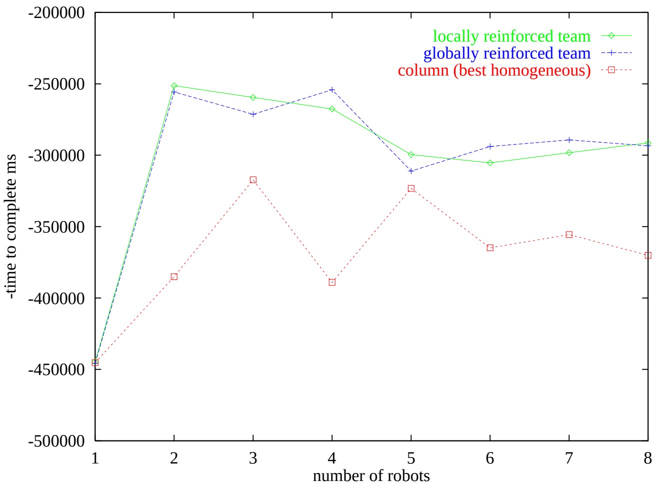

Figure 8.9: Performance for learning navigating teams.

8.5.3 Learning rate

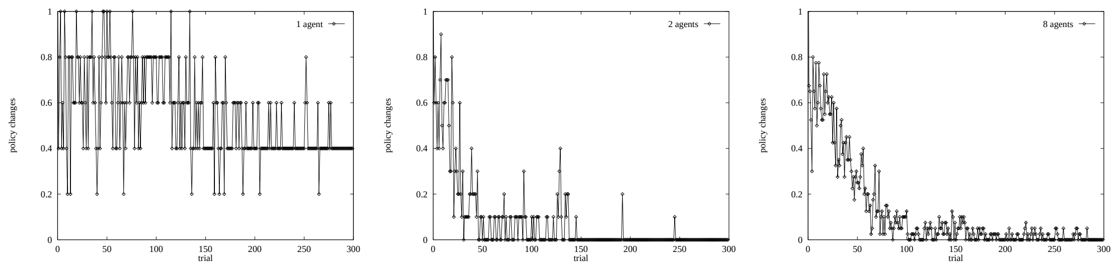

Figure 8.10: Policy convergence measured as average number of policy changes per trial for teams using local and rewards. Left to right: one agent, two agents, eight agents. Plots for teams using global rewards are similar.

Learning rate is evaluated by checking for policy convergence. Convergence is tracked by monitoring how frequently an agent's policy changes. At the end of a trial, after receiving its reward, an agent may switch from one behavior, say *diamond*, to another, perhaps *column*. Switches like this are tracked as policy changes. Because each trial is only a single step, a robot can only switch policies zero or one times per trial.

The data, depicted in Figure 8.10 shows convergence properties for one, two and eight robots using local rewards (plots for global rewards are similar). For two and eight robots, convergence is good, with policy changes dropping off to zero in both cases. In the case of one agent, however, convergence is poor. This is because, in the absence of any other robots to move in formation with, all navigational strategies are equally beneficial. None of the four strategies provides any advantage over the others, so the agents oscillate from one strategy to the other.

8.5.4 Diversity

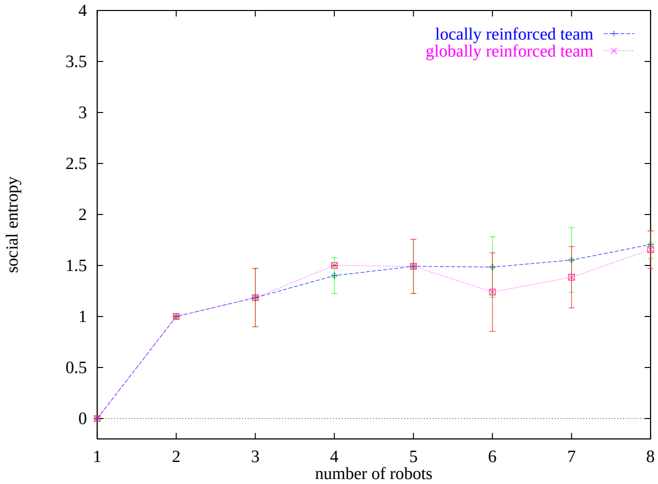

Figure 8.11: Hierarchic social entropy of learning navigational teams. Error bars indicate 95% confidence intervals.

After the training phase, robots are evaluated for behavioral diversity by examining their policies. Diversity is measured after the learning phase is complete using hierarchic social entropy (Chapter 5). For teams of five robots, entropy can range from a minimum of 0.0 (all agents are identical) to 2.32 (all agents are different). The graph in Figure 8.11 plots diversity for the learning navigational teams as the number of robots varies from one to eight. The data indicate no significant difference in diversity between the teams trained using local rewards and those trained using global rewards.

8.6 Discussion

The results of experiments in this task raise several challenging questions. First, why doesn't the selection of local versus global rewards impact diversity as it does in other tasks?

In contrast to the other tasks examined in this dissertation, this one is not internally competitive (COMP_INT). In soccer, for instance, a greedy forward might deny other agents opportunities to score. In this task however, a "selfish" agent seeking to maximize its own reward will not penalize other other agents on the team. It is likely that agents striving to maximize a local reward in this task would behave in the same way as agents striving to maximize a global reward. This is why we see little or no difference in performance and diversity between teams using local and global rewards.

It is also surprising that teams using a diverse set of formation behaviors can perform better than those using the best homogeneous strategy. How can this be? It would seem that formation can only work if the agents agree on the same formation geometry. The answer is that the agents learn to exploit each other's behavior to speed themselves across the terrain. Figure 8.12 illustrates.

In this example the agent at the bottom of the figure attempts to maintain a *line* formation with respect to the other robot. At the same time the robot at the top tries to maintain a *diamond* formation with respect to the bottom robot. The top robot can never reach its formation position because as it attempts to move there it pulls the other agent along with it. The resulting interaction is very much like a "carrot on a stick" for both robots. The **maintain_formation** behavior contributes a forward vector to the motion of both agents, thus speeding them up.

Note that the agents only benefit from this effect when they select differing formation behaviors. Otherwise they would quickly settle into equilibrium and be driven forward only by other forces (e.g. **move_to_goal**). **The agents diversify in response to one another's behavior.**

Another strategy observed in the learning teams is for a "leader" robot to select the *no_formation* assemblage while a follower utilizes the *column* or *diamond* behavior. The leading robot moves more quickly than if it used a formation behavior because it is repelled slightly from the trailing agent (due to the **avoid-robot** motor schema) and not pulled back by a formation force. The trailing robot in turn is "pulled" forward by its **maintain_formation** motor schema.

The resulting team behaviors do not provide the regular geometric arrangements one would expect in robot "formations." In most systems, the group breaks into pairs of agents that move across the field in an irregular formation like that in Figure 8.12. Note that regular geometry is not a performance criteria, nor is it part of either of the reward functions. There is really no reason to expect it in the resulting multi-robot teams.

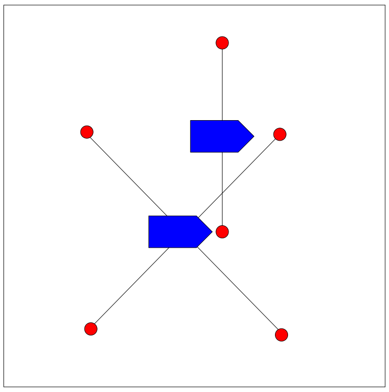

Figure 8.12: Agents exploit each other's formation behavior to move more quickly.

Finally, it must be pointed out that the experimental approach used may have contributed to the extent of diversity in the systems. Although each team was trained in a different random environment, the environments were not re-randomized between trials. It is likely that the robots adapted to the specific environment they were trained in. This bias could be removed by rearranging the obstacles at the start of each trial.

8.7 Summary

Both fixed and learning teams were evaluated in their ability to navigate quickly across an obstacle field. The experiments utilize a new scalable and flexible strategy for multi-robot formation and cooperative movement. Behaviors for four types of formation geometries were developed and evaluated: *diamond*, *line*, *column* and *no_formation*.

The results for teams following fixed homogeneous policies agreed with results from earlier work [BA95b]. In particular, *column* formations are best for navigation across cluttered terrain. Line formations perform worst.

In separate experiments, robots were trained to navigate using local and global reward strategies. In contrast to the other task domains examined in this work, the performance and diversity of teams was not significantly impacted by the choice of reward. The key results for this task are:

- team performance is about the same for local and global rewards,
- both types of rewards lead to diversification in the robot teams,
- the learning robots find ways to exploit each others' behavior in order to move more quickly across the terrain.

Different types of rewards do not impact diversity or performance because the task is not internally competitive (`COMP_INT`). Agents striving to maximize individual performance do not penalize each other, and will also tend to maximize team performance.