Modeling Robot Differences by Leveraging A Physically Shared Context

Zsolt Kira and Kathryn Long
College of Computing,
Georgia Institute of Technology,
Atlanta, Georgia 30332-0250
Zsolt.Kira@cc.gatech.edu, gtg324x@mail.gatech.edu

Abstract

Knowledge sharing, either implicit or explicit, is crucial during development as evidenced by many studies into the transfer of knowledge by teachers via gaze following and learning by imitation. In the future, the teacher of one robot may be a more experienced robot. There are many new difficulties, however, with regard to knowledge transfer among robots that develop embodiment-specific knowledge through individual solo interaction with the world. This is especially true for *heterogeneous* robots, where perceptual and motor capabilities may differ. In this paper, we propose to leverage similarity, in the form of a physically shared context, to learn models of the differences between two robots. The second contribution we make is to analyze the cost and accuracy of several methods for the establishment of the physically shared context with respect to such modeling. We demonstrate the efficacy of the proposed methods in a simulated domain involving shared attention of an object.

1. Introduction

Developmental robotics attempts to study robotics from the perspective of building capabilities progressively via embodied interaction with the world. In the single-robot case, exploration in the world is performed alone and can involve trying to find cause-effect rules (Drescher, 1991) or exploration of the robot's own capabilities. With multiple robots, it is crucial for the robots to be able to share knowledge either through explicit communication or implicit means such as imitation. Such knowledge sharing speeds up development significantly and can allow more experienced robots to impart their wisdom to others. Social aspects of development have been recognized by developmental psychologists such as Vygotsky. It also allows one robot to be the teacher of another robot, reducing need for costly human interaction.

Several problems can prohibit effective sharing of knowledge, however. First, often-times the knowledge learned via exploration of the world is embodiment-specific. This is especially true for developmental learning methods, and can be problematic for robots because they may differ from each other. It is quite common to have some degree of heterogeneity among robots, and there can be slight perceptual differences even among two robots of the same model. For example, the camera color models may differ slightly. It is an even greater problem when different types of robots are used. Currently, there is a plethora of robotic systems in use in home environments (e.g. the Roomba and lawn mowing robots), research labs, and in various domains where task allocation to different robots is necessary (Gerkey and Mataric, 2004).

In order to allow knowledge sharing among such heterogeneous robots, they must first model their differences. The first insight of this paper is to propose leveraging *similarity* to deal with *heterogeneity*. Specifically, in order to ascertain their differences the robots engage in a protocol whereby they establish a *physically shared context* or joint attention. In other words, the two robots must make sure that the part of the environment that they are sensing is the same (the portion of the environment in question is currently assumed to be static). In an object-recognition domain, for example, the two robots should make sure they are looking at the same object in the environment. We distinguish this from a more *global* shared context, whereby knowledge of the tasks and intention are shared as well (this is similar to the notion of a shared intentional relation in (Kaplan and Hafner, 2006)).

There are several different methods for establishing such a shared context, and some of these have been studied previously in the joint attention and gaze following literature. However, a great deal of this research studies human-robot joint attention, focusing on perceptual problems of gaze following and making the interaction natural to humans. When two robots are involved, there is potential for additional methods and many assumptions can be made. For example, when localization is available, robots can easily transmit their headings to each other. Furthermore, each method has a cost associated with its establishment and the accuracy of the resulting shared context. The second contribution of this paper is to quantitatively analyze this cost/accuracy trade-off. In addition, we frame this analysis with respect to our first point, namely that the reason some methods are more accurate than others is that they differ in the guarantees they make about the similarity which is being leveraged.

More specifically, the three methods used are: 1) trading places, 2) following, and 3) pointing. We analyze these with respect to an object-recognition domain, where the goal is for both robots to have similar perspectives of the same object in the environment. The first method involves one robot locating an object and approaching it closely, taking a sensory snapshot, and sharing its location with the other robot. Note that sending raw sensory data at a particular location is an advantage afforded to two communicating robots which is not possible with a human and a robot. The first robot then moves out of the way, allowing the second robot to take its place in the same location. This method is very accurate in the sense that the two robots receive sensor readings from the same exact location, but is costly in terms of travel time and distance.

The second method is very similar, except that instead of occupying the exact same location, the second robot docks beside the first one and aligns its gaze. Intuitively, this method seems similar in cost, but less accurate because the viewing angles may differ somewhat. The third method is a longer distance shared gaze, where the first robot rotates until finding an object, no matter how far, and the second robot attempts to gaze at the same object. This method is hypothesized to be very cheap, but also grossly inaccurate due to possible occlusions and widely differing points of view. We perform quantitative experiments to characterize the trade-off between cost and accuracy of these methods.

2. Related Work

Establishing a physically shared context is a difficult problem and has been studied in numerous publications, usually using the terms *shared or joint attention* (Scassellati, 2000), or as a sub-problem of *gaze following* (Movellan and Watson, 2002). Most of this work deals with the human-robot interaction problems that differ when dealing with two robots instead. One exception is work done by Kaplan and Hafner which studied the development of skills necessary for joint attention between two robots. They recognized that this problem is not just one of gazing, but establishing a shared intentional relation to the world.

The main contribution of this paper is to utilize a physically shared context to learn about how two robots differ in their perceptual capabilities, a concept which has not been raised to our knowledge. Such a shared context is traditionally used for knowledge sharing or shared development of vocabulary. For example, Jung and Zelinsky studied two robots that are performing the same task (vacuuming) but that have different capabilities or roles; one robot can sweep small pieces and reach into corners, while the other can only vacuum the larger piles and cannot reach corners (Jung and Zelinsky, 2000). In that case, a shared ontology was developed by establishing a physically shared context: The two robots followed each other around the room and agreed on symbols for specific locations in the environment. Another example is work done by Luc Steels and his colleagues in the area of shared vocabulary development (Steels and Kaplan, 1999). They used shared attention to synchronize the two robot's symbols. This is a similar concept to ours, except that heterogeneity at a higher level (symbols) is being bridged. There have been attempts at realizing the work on real robots, and it has been noted that joint attention and context is crucial but can be problematic (Baillie and Nottale, 2005).

Some implicit knowledge sharing methods, such as imitation, have to deal with both shared attention (*what* to imitate (Dautenhahn and Nehaniv, 2002)) and differences in embodiment (Allisandrakis, 2002). One relevant study looked at an imitation task where a learner is trying to synchronize its vocabulary with a teacher by following it and listening to the teacher's utterances (Billard and Dautenhahn, 1998). In that case, it was observed that the fact that the learner was behind the teacher and not in exactly the same position (i.e. they did not have an accurate shared context) caused errors in learning. In terms of imitation among robots with differing embodiments, the approach used by Allisandrakis is to use a task-based approach whereby imitation was optimized based on task performance (Allisandrakis, 2002). Such learning, however, must be done for each task. We propose to explicitly model differences between robots and use these models to transfer skills in many varying tasks.

3. Domain and Robot Architecture

The environment used for our experiments is an indoor environment based on the layout of the robot laboratory, simulated in Player/Stage (Gerkey, et al., 2003). The environment and two robots are shown in Figure 1. In the environment, there are many objects that are recognized by the robots via their color. The object of interest, show in Figure 1, is a symmetric red/blue object composed of four red squares surrounding a larger blue square.

The robots themselves are simulated Pioneer 2DX robots and have simulated odometry, laser, and camera with a blob tracker. Odometry and laser sensors are the same in both robots, but heterogeneity arises from the fact that the two robots differ in visual sensing. Specifically, two aspects are changed: The zoom of the camera (resulting in smaller blobs for smaller scale)

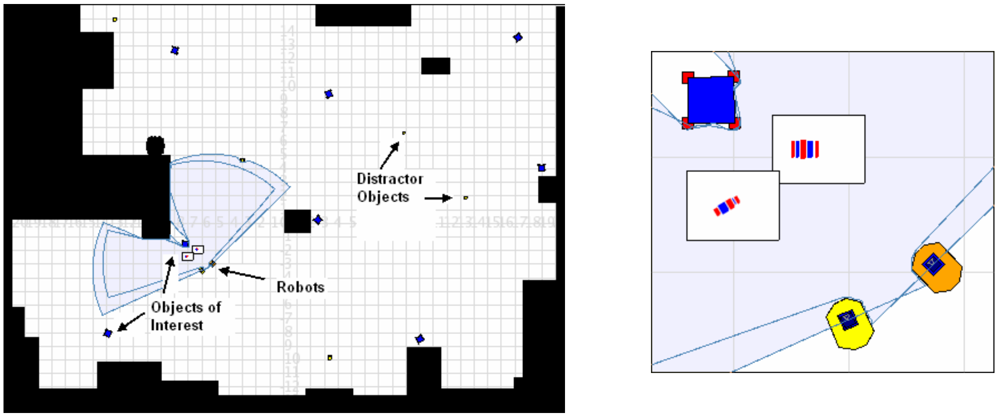

Figure 1 – Left : Map of the environment, corresponding to the layout of the robot laboratory. Right : Zoom in of robots, showing the object (upper left), two robots, and their points of view in the blob finder (squares). Note the differences in scale (due to the zoom of the camera) and camera angle.

and the angle of the camera (resulting in skewed objects). These differences can be seen in the figure.

Robot control is performed via a behavior-based system based on the *AuRA* architecture (Arkin, 1997). Lower level primitive behaviors, such as obstacle avoidance, can be combined into higher level motor schemas such as a **GoTo** behavior that includes movement towards the goal, obstacle avoidance, and a small amount of noise. We also make use of schemas that do not necessarily move the robot, but perform computations such as storing sensory readings or sending messages to other robots. In our diagrams, motor schemas will be represented as ovals and computational schemas as rectangles. An entire mission can be specified via a Finite State Automata, where control moves from one motor or computational schema to another via perceptual triggers. This is based on *MissionLab*, a mission specification system utilizing the *AuRA* architecture (MacKenzie et al., 1997). However, our implementation is not geared to usability and hence does not utilize a GUI for specifying the FSAs.

4. Leveraging Physically Shared Context

4.1 Leveraging Similarity to Deal with Heterogeneity

There are many levels at which two robots can be heterogeneous, all the way from sensorimotor capabilities to higher level knowledge. In this paper, we are dealing with low level perceptual differences. In order to deal with this heterogeneity, we contend that there must be some similarity between them. Furthermore, similarity with respect to the environment and task can be leveraged as well. Two robots that have completely different sensing (e.g. laser versus a heat sensor) will find communication incredibly difficult. However, even robots with small variations in their sensing capabilities (e.g. different camera angles) must furthermore share aspects of the environment and task.

Specifically, we propose to leverage similarity in location, whereby similar aspects of the world are attended to by both of the robots. We achieve this via joint interaction between the robots, where they actively make sure that they are in the same location (and hence perceiving the same object). We also assume that the environment is static near the location of interest. We show in our experiments that this allows two robots to learn simple models of their sensing differences. Note that these same mechanisms may also be used for establishing shared context for the purpose of the actual knowledge transfer once these models are built.

4.2 Methods of Establishing a Physically Shared Context

As discussed in previous sections, the problem of establishing such joint attention is not a trivial problem. Many methods exist, and these methods can furthermore be implemented in various ways. In this paper, we select three different methods and implement them in a behavior-based architecture. The second contribution of this paper is then to analyze the cost of each method and the accuracy of the resulting physically shared context (our metrics for cost and accuracy will be defined later). Furthermore, the differences between these methods can be characterized by how well they guarantee that aspects of the environment are similar.

The three methods we use are: 1) trading places, 2) following, and 3) pointing. For each method, there is a "trainer" and a "learner". The trainer is the robot that first selects a given object to focus on. This is done by exploring the environment until the object is detected via the blob detection system. The Finite State Automatas for two of the methods (following and

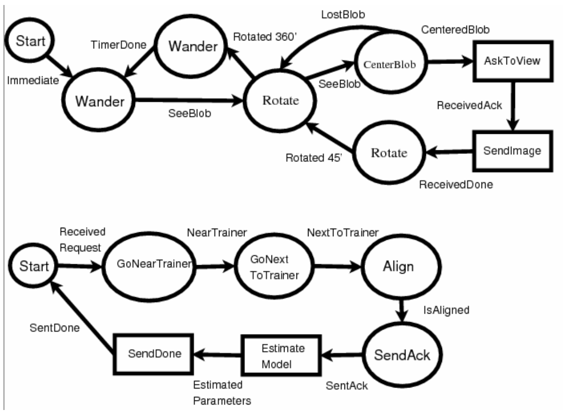

Figure 2 - FSA for "pointing" method. On the top is the trainer FSA and on the bottom is the learner FSA.

pointing) and each role (trainer and learner) are shown in Figures 2 and 3.

The first method, trading places, involves the two robots attempting to occupy exactly the same position and vantage point one after another. Note that this results in a more accurate shared context than just shared attention, yet is not that much more expensive as we will see. In this method, the trainer travels to the object once it is detected, and then asks the learner to follow. At this point, the trainer sends its captured viewpoint to the learner, and wanders until a set time has expired, after which it waits for the learner to finish. Meanwhile, after receiving the request from the trainer, the learner moves to the trainer's previous position. Once there, the learner aligns to the object and compares the image the trainer sent to its current view. Through this comparison, the learner estimates the model and notifies the trainer of its completion.

In the second method, following, the learner docks close to the trainer and attempts to obtain a vantage point similar to that of the trainer. As can be seen in Figure 3, the trainer travels close to the object after detecting the object. Once at the object, the trainer asks the learner to 'follow', at which point the learner approaches, and docks next to, the trainer. Once docked, the learner aligns to the object and sends acknowledgement. In response, the trainer sends its current vantage point, with which the learner is able to estimate a model using the received image and its own view.

The third method, pointing, involves the two robots focusing on a distant object from adjacent locations. As shown by the FSA in Figure 2, the trainer wanders until detecting the object, and then rotates until the object in view. At this point, the trainer centers the object in its view, and then asks the learner to 'view'. Upon receiving this request, the learner approaches, docks next to, and aligns with the trainer. Once aligned, the learner sends the trainer a message,

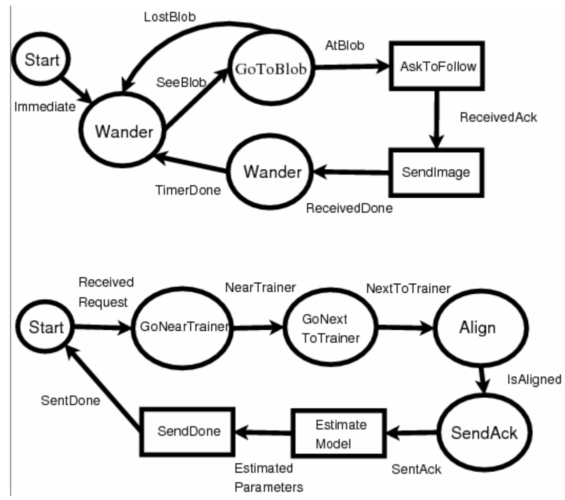

Figure 3 - FSA for the "follow" method. On the top is the trainer FSA and on the bottom is the learner FSA.

to which the trainer responds with an image of its current view. The learner then uses its vantage point and the received image to estimate the model, notifying the trainer when complete.

The first method, given previously stated assumptions such as accurate localization, guarantees that differences in sensor data between the two robots result from differences in the robots themselves. The second method, however, reduces this guarantee by introducing errors due to perspective differences. Objects, for example, can look very different depending on the angle from which they are viewed. This difference is minimized somewhat by making sure that the robots view the object from the same distance. The third method introduces even more error due to a higher degree of perspective differences and lower resolution. This characterization of similarity guarantees leads us to conclude that intuitively, trading places will be more accurate than following, which will be more accurate than pointing.

5. Experimental Design and Results

5.1 Modeling of Robot Differences

In this paper, we are focusing on methods for shared context establishment and showing that such similarity can be leveraged to understand heterogeneity. Hence, the models we are using for the differing robots are extremely simple, namely the direct estimation of the angle and scale differences of objects in the cameras. Other models can be used as well with additional types of heterogeneity, for example color differences in cameras or even at higher levels qualitative differences in sensing.

For each pair of image instances (one from each

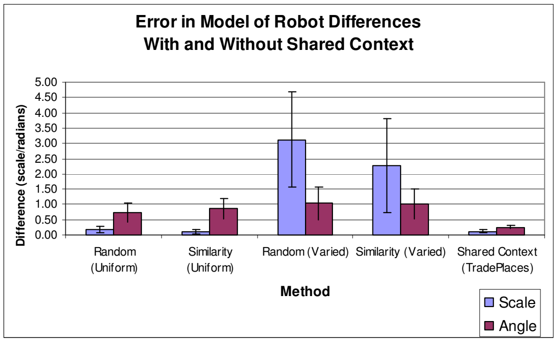

Figure 4 – Error in model learned by 1) Random Selection : Selecting random instances from object training of trainer and learner, 2) Similarity-Based Selection : Selecting most-similar instances from object training of trainer and learner, and 3) Shared Context : Where the learner and trainer establish a shared context first and then trade instances. In the uniform condition, images of the object were acquired from exactly the same distance away (one meter), while in the varied condition object instances were acquired from different distances.

robot), the scale difference is estimated by dividing color histograms obtained from each image. Color histograms are obtained by summing up counts of a discretized color space, in this case into three bins: red, green, and blue. Since a scale difference results in different sized blobs in simulation, this suffices. Formally, given two images ***I*** and ***I'***, the scale of size difference is estimated to be:

$$\frac{1}{n}\sum_{j=1}^{n} H_j(I)/H_j(I')$$

where there are n color bins (in our case $n = 3$) and $H_j(I)$ is the count in image I of pixels having color j. In other words, we are averaging the fraction of counts for each color.

The difference in angle is estimated by discretizing the angle space into 16 bins, and finding the pixel-wise image difference between the two images for each angle. The angle with the lowest difference (highest similarity) is chosen as the best-fitting angle. In the following experiments, accuracy is measured as the difference between the actual angle and the scale of size difference compared to the estimated values for these parameters.

5.2 Leveraging Similarity to Deal with Heterogeneity

The first experiment tests the hypothesis that obtaining key instances in the environment that are shared will allow the robots to learn the parameters of the differences in their sensing more accurately than when a physically shared context is not guaranteed. In the control condition, the two robots explore the environment and store ten instances of the target object, and attempt to model their differences by 1) randomly

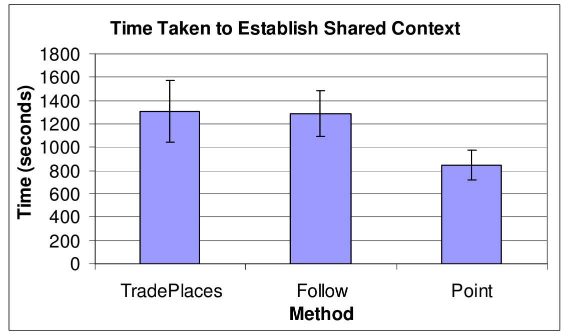

Figure 5 – Time taken to establish a physically shared context using the three methods.

selecting among these instances or 2) selecting instances based on their similarity. Specifically, the ten most similar pairs of images from the two sets are chosen. We did this in two different conditions: one where the object instances were capture when the robot was a uniform distance from it, and one where the distance from the object varied. Similarity is measured using the intersection of the color histograms, a standard appearance-based approach (Swain and Ballard, 1990). Formally, given two images ***I*** and ***I'***, the intersection of their two normalized histograms is:

$$H(I) \cap H(I') = \sum_{j=1}^{n} \min(H_j(I), H_j(I'))$$

where there are n color bins (in our case $n = 3$) and $H_j(I)$ is the count in image I of pixels having color j.

In the experimental condition, we use the first method (trading places) to establish a shared context among the two robots, after which they exchange sensory snapshots and learn their differences using these instances. The hypothesis is that performance will be poor in the control conditions relative to the experimental condition. As stated, the metric for accuracy is based on the actual difference between the model parameters and ground truth, which is known.

Figure 4 shows the accuracy results for the three conditions, plotting the average absolute difference between the model and the known ground truth. The results are averaged over five runs for each condition. For scale estimation, this term is unitless while for angles it is in radians. As can be seen, establishing a shared context using the most accurate method results in better overall performance in terms of modeling robot differences. Especially when the object distance is varied, which is more realistic, there is substantially larger error without shared context. Even when some context is guaranteed by the control mechanism (i.e. the snapshot of the object is always taken from a fixed distance), the performance is comparable for estimating scale. Estimating angle, however, is more sensitive and the direct establishment of shared context has lower error. Using image similarity seemed to help in terms of modeling scale, although the variances involved are

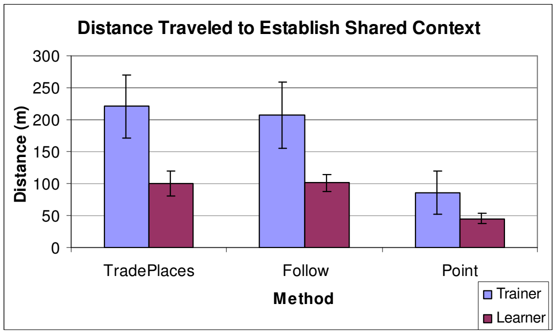

Figure 6 – Distance traveled while establishing a physically shared context using the three methods.

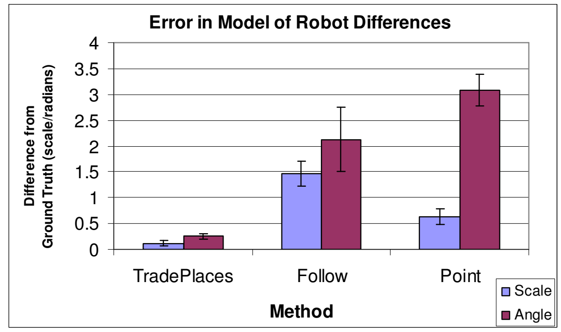

Figure 7 – Error in models of the differences between the two robots, when using instances obtained while establishing a physically shared context using the three methods.

high and more experiments are necessary to confirm this. Image similarity did not make a difference in terms of angle. This is probably because the similarity metric does not take into account the differences between the two robots.

As shown, error reduction occurs even when all of the instances in the control conditions were taken in a situation where the robot was the same distance from the object. Of course, if a larger library of instances is built, this error can be reduced. However, it is not practical to explore all parts of the environment or all views of an object. Furthermore ascertaining which instances are similar is made difficult by robot differences, as shown in the similarity-based selection method. Establishing a shared context and learning the model is quicker and results in very low error rates.

5.3 Analyzing the Trade-off between Cost and Accuracy

The second experiment seeks to analyze the trade-offs between the three methods used to establish a physically shared context. Based on the discussion analyzing their varying levels of guarantees of similarity, the hypothesis is that pointing will be the least accurate but cheapest, following will be more accurate but much more expensive, and trading places will only be slightly more costly than following but significantly more accurate. Here, our metric for accuracy is based on the difference between the model parameters and ground truth knowledge of their differences. Standard deviations are also calculated to analyze statistical significance, and are shown via error bars

Figures 5 and 6 show the average time taken and distance traveled to establish a shared context for ten instances. As predicted, pointing is superior according to both metrics. This is because it is a longer distance interaction and the robot can establish a shared context for multiple objects that are around. In our environment, the same object was located in multiple places, leading this strategy to be quicker. There is not much difference in time taken for trading places and following, and only a slight difference that is not statistically significant (due to the variance) in distance traveled. Despite this, the two methods vary greatly in terms of modeling accuracy, as can be seen in Figure 7.

In terms of estimating the angle difference, trading places was better than all the other methods, and following was second but much worse. Pointing resulted in the worst performance. Surprisingly, this role between pointing and following was reversed in terms of estimating the scale difference. This is one data point that did not agree with our hypothesis. The reason for this is that since blob detection is performed in simulation, the scale can still be estimated despite being very far away. In fact, since both robots were far away, the difference in perspectives was not as great as in the following method.

6. Discussion and Conclusions

The experimental results we have demonstrated have confirmed most of our hypotheses. Establishing a physically shared context allowed the robots to build accurate models of their differences. Furthermore, we have analyzed several methods of doing so, and their trade-offs in terms of cost and accuracy.

One surprising result was that, in some cases, pointing at an object can result in more accurate models than following. This is because close to an object, differences in perspective can greatly change what one sees. If only a crude characteristic, such as scale , is needed then longer distances are not a problem. This demonstrates that depending on what types of differences are being modeled, different methods are better than others. In the future, it would be interesting to explore dynamically switching between these method based on error characteristics or the type of heterogeneity that is being modeled.

In this paper, we have proposed leveraging similarity via a physically shared context in order to model differences between heterogeneous robots. Understanding these differences is crucial for knowledge transfer among heterogeneous robots, which will become increasingly important as specialized robots performing specific tasks become ubiquitous. Although they each perform different tasks, there is a great deal of generalized skills that they can trade in order to speed up learning. This is especially important if a developmental robotics approach is taken, where learning can take a large amount of time and must be in the context of social interaction. Such interaction allows robots to learn skills for problems that they have not yet encountered or mastered.

Several extensions to this work are possible. The differences between the robots and the models we have used here are simple. It would be useful to perform experiments similar to those in this paper with real heterogeneous robots that are currently in the lab. Exploring what type of models can be used would be interesting, be it via machine learning where possible or symbolic representations. Ultimately, the goal is for the robots to use these models to share knowledge and adapt it based on their differences.

Acknowledgements

We would like to acknowledge the SAIC Scholars program for their generous support.

References

Alissandrakis A., K.D. (2002), Imitating with ALICE: Learning to Imitate Corresponding Actions across Dissimilar Embodiments, *in* 'IEEE Transactions on Systems, Man, and Cybernetics'.

Arkin, R. (1997), 'AuRA: principles and practice in review', *Journal of Experimental & Theoretical Artificial Intelligence* **9**(2), 175--189.

Baillie, J. & Nottale, M. (2005), 'Dynamic Evolution of Language Games between Two Autonomous Robots', in 'Proceedings of the IEEE International Conference on Development and Learning'.

Billard, A. & Dautenhahn, K. (1998), 'Grounding communication in autonomous robots: an experimental study', *Robotics and Autonomous Systems* **1-2**, 71-81.

Drescher, G. (1991). *Made-up minds: A constructivist approach to artificial intelligence*. MIT Press.

Gerkey, B.; Vaughan, R. & Howard, A. (2003), 'The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems', in 'Proceedings of the 11th International Conference on Advanced Robotics', 317--323.

Gerkey, B. & Mataric, M. (2004), 'A Formal Analysis and Taxonomy of Task Allocation in Multi-Robot Systems', *The International Journal of Robotics Research* **23**(9), 939.

Jung, D. & Zelinsky, A. (2000), 'Grounded Symbolic Communication between Heterogeneous Cooperating Robots', *Auton. Robots* **8**(3), 269--292.

Kaplan, F. & Hafner, V. (2006), 'The challenges of joint attention', *Interaction Studies* **7**(2), 135--169.

MacKenzie, D.; Arkin, R. & Cameron, J. (1997), 'Multiagent Mission Specification and Execution', *Autonomous Robots* **4**(1), 29--52.

Movellan, J. & Watson, J. (2002), 'The development of gaze following as a Bayesian systems identification problem', in 'Proceedings of the 2nd International Conference on Development and Learning', 34--40.

Scassellati, B. (2002), 'Theory of Mind for a Humanoid Robot', *Autonomous Robots* **12**(1), 13--24.

Steels, L. & Kaplan, F. (1999), 'Bootstrapping Grounded Word Semantics', in Briscoe, T., editor, *Linguistic evolution through language acquisition: formal and computational models*, pages 53-74, Cambridge University Press. Cambridge, UK.

Swain, M. & Ballard, D. (1990), 'Indexing via color histograms', in 'Proceedings of the Third International Conference on Computer Vision', 390--393.