2015

An embodied approach to evolving robust visual classifiers

Karol Zieba
University of Vermont, kzieba@uvm.edu

An embodied approach to evolving robust visual classifiers

A Thesis Presented

by

Karol Zieba

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Master of Science
Specializing in Computer Science

October, 2015

Defense Date: June 5, 2015
Thesis Examination Committee:
Josh Bongard, Ph.D., Advisor
Matthew Price, Ph.D., Chairperson
Margaret J. Eppstein, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

From the very creation of the term by Czech writer Karel Čapek in 1921, a “robot” has been synonymous with an artificial agent possessing a powerful body and cogitating mind. While the fields of Artificial Intelligence (AI) and Robotics have made progress into the creation of such an android, the goal of a cogitating robot remains firmly outside the reach of our technological capabilities. Cognition has proved to be far more complex than early AI practitioners envisioned. Current methods in Machine Learning have achieved remarkable successes in image categorization through the use of deep learning. However, when presented with novel or adversarial input, these methods can fail spectacularly. I postulate that a robot that is free to interact with objects should be capable of reducing spurious difference between objects of the same class. This thesis demonstrates and analyzes a robot that achieves more robust visual categorization when it first evolves to use proprioceptive sensors and is then trained to increasingly rely on vision, when compared to a robot that evolves with only visual sensors. My results suggest that embodied methods can scaffold the eventual achievement of robust visual classification.

Acknowledgements

I would like to thank my adviser, Josh Bongard, for his tremendous guidance and insight. I also thank my additional committee members, Dr. Price and Dr. Eppstein, for their feedback and advice.

This thesis is dedicated to my parents: Elżbieta and Jerzy Zięba.

Table of Contents

List of Figures

Material from this Thesis has been published in the following form:

Zieba, K., Bongard, J. (2015, July). An embodied approach for evolving robust visual classifiers. In Proceedings of the 2015 on Genetic and Evolutionary Computation Conference (pp. 201-208). ACM.

1

Introduction

1.1 Overview

This thesis begins by reviewing previous work on cognition, embodiment, robotics, and simulation upon which this thesis is based. Following this review, the methodologies used to design and construct the robotic experiment will be described. The results are then provided. Along with the results there will be a discussion of the various insights gleaned. Finally, I consider potential avenues for future work that can continue from the work presented in this thesis.

1.2 Cognition

Cognition can be broadly defined as some mental process that acquires and utilizes knowledge to develop understanding. There are many different processes that have been considered the building blocks of cognition: memory, association, pattern recognition, language, problem solving, etc (Coren, 2003).

Within the educational community one attempt at distinguishing the fundamental skills of cognition is known as Bloom's taxonomy (Anderson et al., 2001; Krathwohl, 2002).

Bloom's taxonomy, intending to aid educators, categorized cognitive skills into a multi-layered pyramid. Of these various skills, this thesis concerns itself particularly with the ability to teach machines the ability to categorize. In Bloom's taxonomy, classification is considered a combination of the pyramid's lowest two levels as shown in Figure 1.1: remember and understand. Once data is obtained, through the cognitive task of memory, an agent can begin to develop an understanding of the supplied data by classifying, or grouping, concepts into categories.

The classical, or Aristotelian, view of categorization is based on the work of Plato and refined by Aristotle in his work *Categories*. Categories in classical view are discrete entities whose members share a particular set of properties. Categories are viewed as mutually exclusive and collectively exhaustive (Cohen and Lefebvre, 2005). Because of these properties, objects to be categorized can be thought of as falling neatly into their proper categories. These principles can be found in the objects being categorized in this thesis: the objects vary only in size and each object is precisely one of two sizes.

However, categories can also be defined in ways that do not follow the neat principles of being mutually exclusive and collectively exhaustive. Researchers and philosophers have developed other approaches to categorization (Ashby and Maddox, 1993) that they believe are more similar to the embodied approach that human and animal brains internally categorize. One of these views is know as Prototype theory (Rosch, 1975; Rosch, 1999). Objects fit in a graded set of categories based on their similarity to a hypothetical or real prototype of that category. A common example used to elucidate prototype theory is to consider the mind's usual definition of a bird. A robin, which has feathers, a beak, and the ability to fly, is considered more prototypically birdlike than a penguin, which swims instead of flies. A similar categorization theory is Exemplar Theory (Ashby and Maddox, 1993), which compares unseen objects against the categories of objects previously seen, labeling the new object according to how many objects in existing categories it compares favorably against.

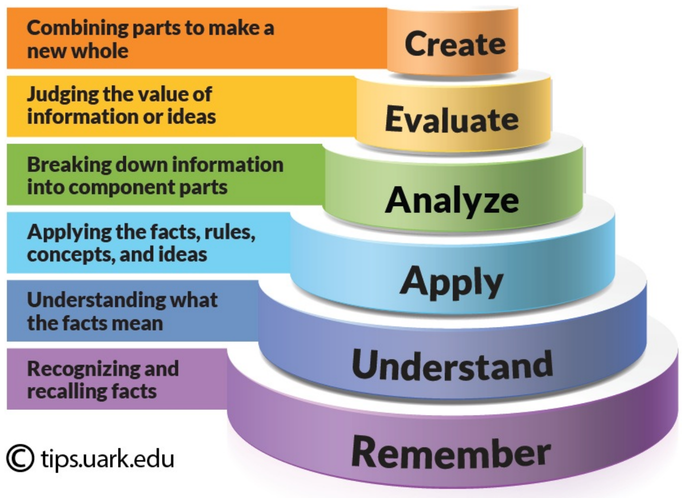

Figure 1.1: Bloom's Taxonomy. Used with permission from http://tips.uark.edu/using-blooms-taxonomy

Categories with one object can consider that object as prototypical. In both Prototype Theory and Exemplar Theory, categorization is seen as a comparison of objects in categories, rather than a comparison of the abstract definitions of a category. In this sense, it can be considered an embodied approach to categorization. The body must utilize experience to judge and label objects. Although these two views are not directly applicable to the experiment of this thesis, they will be relevant for future work. Categorizers in supervised learning approaches (Bengio, 2009; Bishop et al., 2006; Mitchell, 1997) attempt to extract features from pre-labeled examples to categorize unseen objects.

1.2.1 Disembodied Cognition

Traditionally cognition has been viewed as a process that our mind performs. One historical distinction in the field of epistemology has been that of internalism versus externalism (Fumerton, 1988). Internalism is the view that our mind, cognition, and justification is solely determined by factors within us. Externalism is the view that our mind and justification exist past our nervous system. There are various scales at which externalism is seen to operate.

The discussion of what constitutes AI can be pointed back as beginning with Alan Turing's 1950 quote "I propose to consider the question: can a machine think?" (Turing, 1950) In his article Turing outlines his now famous Turing Test through the lens of an imitation game. A machine would be considered intelligent if it could convince at least 70% of interrogators, through 5 minutes of questioning that it is a human. As it is done entirely through typewritten text, the Turing Test was primarily a feat of symbolic manipulation. Throughout AI's history, many have questioned whether a computer can have a mind and consciousness. One prominent example was John Searle's Chinese room thought experiment (Searle, 1980), which stirred up many questions. Chiefly relevant to this thesis is the concept of symbol grounding (Harnad, 1990). The symbol grounding problem is a question of how words obtain their meaning. In symbol manipulation the meaning of symbols is derived from their shape or label rather then their intrinsic meaning. As described by Searle, computation ultimately is the manipulation of its binary symbols. In the years following Turing's article, the dominant early paradigm of AI research was symbolic (Haugeland, 1989). The symbolic paradigm of AI assumed that symbolic representation and manipulation can be sufficient to perform cognitive tasks at the same level as human cognition. Although some of the earliest AI research involved early forms of artificial neural networks, broadly labeled"connectionism," these efforts were largely dropped by 1969 in favor of symbolic reasoning (Minsky and Papert, 1987). One field to successfully emerge from the symbolic paradigm in the 1980's was Expert Systems (ES) (Jackson, 1986). The combination of an inference engine and knowledge database was successfully utilized by both government and commercial entities. However, ES remained inadequate at addressing many aspects of cognition. Primary weaknesses of ES were its lack of adaptability to new problem domains and inability to achieve knowledge acquisition through automated means.

One approach to addressing the inadequacies of ES has been Machine Learning (ML).

Tom Mitchell has defined ML as: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 1997) As described by Mitchell, ML approaches are often statistical in nature. One recent approach to ML is known as Deep Learning (DL). DL attempts to model high-level abstractions in data through multiple layers of non-linear transformations, with each layer learning some combination of the data's underlying statistical patterns. Current DL methods have had success in categorizing (Bengio, 2009; Hinton, 2007), particularly in automatic speech recognition, image recognition, and natural language processing. In the case of image recognition, DL methods are able to learn from MNIST, a data set of hand written digits, with test errors of a quarter of a percent (Ciresan et al., 2012). While ML has accomplished a variety of successes, other work has shown that even state-of-the-art DL algorithms can fail (Szegedy et al., 2014; Nguyen et al., 2015). These categorizers are likely optimized to extract features in a very different way than the human mind. Perhaps providing a learning agent with experiences more akin to those of an adolescent person could be beneficial?

1.2.2 Embodied Cognition

One proposed method of addressing some of the problems in current AI and ML is through the lens of externalism. Particularly through Enactivism, which is the belief that cognition develops from dynamic interaction between an organism, or agent, and its environment (Thompson, 2007). This ties to the concept of embodied cognition; cognition is not just what is contained within our minds, but necessarily and intrinsically tied to our bodies' interactions with the environment. The theory of Embodied Cognition asserts that our bodies are not merely a tool of our mind, but that cooperating systems that can offload a mind's cognitive tasks (Rosch et al., 1992). One theorized advantage of the embodied approach to cognition, when combined with social interaction, is its potential in addressing the symbol grounding problem (Steels, 2008). Interaction with the environment and other agents is theorized to form a solid basis upon which to non-recursively define further symbols.

Human cognition has been shown to be tied to embodiment. One way of looking at the effects of embodiment on our cognition is to correlate the tangible with intangible. One such study showed that subjects who squeezed a soft ball were more likely to perceive gender-neutral faces as female, whereas subjects who squeezed a hard ball perceived gender-neutral faces as male (Slepian et al., 2010). The controlled hardness of the manipulated object primed the participants' reasoning. Another form of priming can be conferred through our inherent perception that weight signifies value. A study found that the bodily experience of weight, particularly in holding money on light or heavy clipboards, grounds our importance of its monetary value (Jostmann et al., 2009). Embodiment also extends to process of learning. The act of gesturing was found to increase students' ability to learn the abstract concepts of mathematics in school (Singer and Goldin-Meadow, 2005). While prototype theory (Rosch, 1975; Rosch, 1999) is thought to be similar to how our brains internally classify objects, disagreement exists on how an artificial agent should learn to do so. Ultimately, it is unknown how biological organisms learn to do so (McNerney, 2011). However, as every human being experiences every stage of development in an embodied state, our bodies might be necessary to jump-start higher cognitive function.

Many computers do not have a body in the traditional sense. While computers are situated in the world, they cannot generally act upon the world. Perhaps the lack of embodiment may lead to deception and overfitting that has been found in DL approaches (Szegedy et al., 2014; Nguyen et al., 2015). Overfitting in the cases we describe in this thesis refers to a modeling error through which the analysis function too closely fits the training data and therefore fails to properly classify novel or adversarial examples. In these examples, no rational human would misclassify the objects in the provided pictures. An imperceptibly doctored picture of a school bus would convince no one that the image is of an ostrich. When these classification methods do not receive adversarial examples, they tend to work wonderfully. However, their striking failures indicate that DL implementations doesn't reason the same way we do. Embodied approaches are theorized to be more resistant to the form of overfitting that non-embodied approaches experience.

Embodied cognition can also be used to solve a variety of tasks that are difficult to inapplicable for disembodied agents. As shown in research into minimal cognition (Slocum et al., 2000; Beer, 2003), embodied agents utilizing artificial neural networks (ANN) can be trained to perceive affordances, discriminate between self and non-self, exhibit short-term memory, and perform selective attention. In these cases, the design of the body is paramount to the tasks it can achieve. Nonetheless, the existence of the body allows for completion of certain cognitive tasks. Another task that embodied cognition can solve is determining which objects in a robot's environment are physically coherent (Fitzpatrick and Metta, 2003). Through the physical manipulation of objects in its visual sight, the Cog robot could determined which components of its environment are connected and which part of its vision is itself (Fitzpatrick et al., 2003). This experiment hypothesizes that vision and action are interconnected in the goal of visual interpretation of a scene.

One theory of cognition that applies to EC is known as Categorical Perception (CP). CP states that the variation of a variable along a continuum is perceived as tending to fall into discrete categories rather than being a mixture of neighboring categories. Using embodiment to study EC is known as Active Categorical Perception (ACP) (Tuci et al., 2010; Beer, 2003). In Beer's paper (Beer, 2003) the task was to distinguish between two dimensional circles and diamonds. In Tuci's paper (Tuci et al., 2010) the task was to distinguish whether a palmed object was spherical or cubic in shape. One particularly useful concept to develop within EC is the study of distinctiveness of categories. CP is said to occur when perceived differences between objects in the same category (intra-category distances) are minimized and/or differences between objects of different categories (inter-category distances) are exaggerated, relative to a baseline (Harnad, 2003). If cognition can find a means to maximize the difference between intra-category distances and inter-category distances then it can be thought of applying EC, and therefore successful categorization. The advantage of ACP is thought to rest in the manipulating objects to reduce spurious differences between classes. As shown in some works, this manipulation can evolve as a side effect (Scheier and Pfeifer, 1995). The experiment laid out in this thesis attempts to exploit ACP's potential of reducing spurious differences between classes of objects.

1.3 Developmental Robotics

Developmental Robotics studies the intersection of robotics and developmental sciences. The focus of Developmental Robotics is on the development of one agent's control systems through experience and time. Therefore, research focuses on the ontological scale, that is the lifetime of a single agent or organism. There are two primary research drivers (Lungarella et al., 2003) in the field of Developmental Robotics. First, engineers seek novel methodologies to advance robotics using nature-inspired approaches of cognitive and morphological development. Second, robots are tools that can be utilized to investigate embodied models of development. As the focus is on development, most Developmental Robotics projects begin with a simple representation of an agent and seek to develop it. Generally active exploration of an organism's surroundings can be used to develop it. However, a couple developmental forces can be outlined: intrinsic motivation, social learning, and body morphology and growth.

One common approach to accelerate development is known as scaffolding. Scaffolding is the support given to an agent during a learning process designed to promote a deeper level of learning. Scaffolding can exist in a variety of means. One common form of scaffolding found in human development are the training wheels of a bike. Aspirational bikers first learn to sit and pedal before the scaffolding (the training wheels) is taken off. This type of scaffolding is also referred to as ontogenetic scaffolding. Scaffolding can also be implemented in robotics (Asada et al., 2009). Reil's and Husband's bipedal walking robot (Reil and Husbands, 2002) was developed to first move down a slope surface utilizing just its own momentum. Then it was evolved using progressively more powerful motors and more gradual slopes until it could walk on a flat surface.

One major goal for developmental psychology, and therefore developmental robotics, is to understand Information Self Structuring (Lungarella and Sporns, 2005). There is an intrinsic relationship between the various senses of a body and its environment. When we see an object, such as a banana, we tend to also know how it feels, smells, and tastes. We also understand its affordances (Gibson, 1977). Affordances can be thought of as the actions an organism is afforded, or permitted, in relation to its environment. However, while we have all these conceptions on what thinking is, the underlying foundations of cognition in humans and AI remains unknown. When DL and other methods learn to recognize an image, they do not associate seen objects with potential affordances. A person might recognize a tree stump as similar to a metal chair due to their similar use for sitting upon. However, current learning algorithms do not structure their knowledge in an embodied way and are therefore unlikely to associate these types of similarities.

1.4 Evolutionary Robotics

While Developmental Robotics focuses on the ontological development of a robot, Evolutionary Robotics studies evolving a population through evaluating members of that population with a fitness function (Nolfi and Floreano, 2000). This process is often highly automated. Most commonly an evolutionary algorithm is used to develop controllers for autonomous robots. Historically, morphological evolution of robotics has been limited due to its practical difficulty.

A fitness function is an objective function that evaluates some sort of input by returning a single metric of that input's fitness. Fitness functions can also be described through the lens of minimizing loss, or error. The goal of a fitness function in ER is to evaluate candidate solutions on the fitness landscape. The landscape is often multimodal in nature and difficult to traverse. Therefore, evolution plays an integral part in discovering solutions whose fitnesses are better than would be discovered through randomness alone. Fitness functions have been created for a variety of tasks in the field of ER, including locomotion (walking and flying), gait evolution, and agent herding (Nelson et al., 2009).

An evolutionary algorithm is a population-based metaheuristic approach to optimizing a given problem. Evolutionary Algorithms are inspired by biological evolution, often incorporating the concepts of reproduction, mutation, recombination, and selection. Members of a population of solutions are evaluated using a fitness function to determine their suitability to either remain in the population or whether their underlying genetic material is propagated to the next generation of solutions.

One example of the utility of evolutionary algorithms was in the development of the Evolved Antenna (Hornby et al., 2006). An antenna was required for satellites that had a series of nuanced design requirements. The cost of utilizing evolutionary computation for this task was found be a 40% reduction in man-hours over manual design.

Various forms of scaffolding can be applied to an ER experiment. Also referred to as Robot Shaping (Perkins and Hayes, 1996), the scaffolding in ER often occurs in phylogenetic time. The population of candidate robot controllers evolves as some attribute of them or their environment is scaffolded. Generally, this scaffolding can occur in three different ways:

Environmental scaffolding As shown by the bipedal walker (Reil and Husbands, 2002), some attribute of the environment can be scaffolded to ease the learning process. While most approaches tend to have the same fitness function regardless of the scaffolding (Bongard, 2011a), this requirement is not necessary.

Morphological Scaffolding This scaffolding works by progressively evolving the physical (or simulated) bodies of organisms or robots. The goal can be either to achieve high fitness on a target morphology, or to discover a morphology that is best suited to the task (Bongard, 2011a).

Sensorial Scaffolding Much how morphological scaffolding modifies the robot's morphology through evolution, sensorial scaffolding modifies the sensors of a robot through evolution. This scaffold seeks to develop robust behaviors in one or multiple sensor by beginning with a different or simplified versions of existing sensors. To date, no other paper has utilized this scaffold.

This thesis explores evolving robots that change over both phylogenetic and ontogenetic time. To date, no other paper was found to utilize both timescales in scaffolding. The experiment will focus on robots that progressively gain the ability to passively see their environment rather than actively feel it.

1.5 Simulations

There are two primary ways of running robotics experiments: physically building robots or designing them in a physics simulation. While research on a physical robot would be ideal, a simulator can achieve meaningful results. While the problem of the "reality gap" certainly exists, in that there are differences between the behaviors of simulated and physical robots, studies (Koos et al., 2010; Koos et al., 2013; Lipson et al., 2006; Bongard et al., 2006) have shown that both designed and evolved morphologies can transfer to reality and therefore minimize the reality gap.

Some of the principle benefits of conducting research using simulations include (Floreano et al., 2008):

- Evolution takes a long time (many generations and evaluations). It is expensive to parallelize with multiple robots.

- Physical robots can be damaged or defective.
- Resetting initial conditions often requires custom robot designs or human intervention.
- It is Hard/Impossible to modify morphology of a robot through evolutionary or ontogenetic time with current hardware, even with human intervention

Due to the first three listed benefits the robots in this thesis will be entirely simulated.

1.6 Motion Supports Categorization

Following birth, infants' initially very limited vision rapidly develops. From a few weeks of age, increasingly complex patterns of objects are visually recognizable, such as those formed by the outlines of disjoint shapes (Haith, 1993). Studies of infant grasping behavior have shown that it eventually becomes visually controlled (von Hofsten and Rönnqvist, 1988). Infants anticipate and attempt to grasp objects coming into close proximity to them (McCarty et al., 2001). While grasping motion has certain practical benefits, as assisting in feeding, it can be reasonably concluded that grasping serves as a way to scaffold an infant's vision. Grasping behaviors in infants share similar benefits with ACP in robotics. By bringing remote objects closer, grasping behaviors reduce spurious differences between objects, particularly position, while enhancing visual acuity to better distinguish objects.

From the human perspective of an infant's grasping behaviors we hypothesize that grasping can be a useful skill in the categorization of objects. From the robotic perspective, the experiments on the robot Cog (Fitzpatrick and Metta, 2003) have shown that a body that interacts with its environment can be utilized to develop an understanding of its visual scene. My review into existing literature has shown that gaps remaining in our knowledge of the development of robust visual classifiers using an embodied approach. This thesis investigates the effects of utilizing scaffolded sensors in an embodied agent to develop its ability to visually categorize. Does motion increase our ability to visually categorize?

2

Methodology

As part of testing this hypothesis Dr. Josh Bongard and I designed an evolutionary robotics simulation. This chapter will outline the various components and reasoning of our methodology.

The code for reproducing this experiment can be found in the following Git repository: http://git.io/vfYYP

2.1 Task

For a given neural controller, a series of cylinders were placed one at a time at predetermined locations. There were only two types of cylinders and they varied only in size: the larger cylinder's radius was 50% longer than the smaller cylinder's. The robot evaluated was tasked with classifying the radius of each cylinder placed within its grasp. The relative size of the cylinder categories was chosen for two reasons. First, we desired a large difference between categories so that the proprioceptive sensors could distinguish between categories. Second, we desired the difference between categories to be small enough so that large cylinders placed far on the Z axis would appear smaller than small cylinders closer on the Z axis. The first goal allows proprioception-based controllers to learn, and therefore generalize, off of the training set. The second goal offers vision-based controllers an incentive to manipulate the cylinders so that behaviors other than memorization can be discovered.

2.2 Robot Morphology

The robot's morphology (Figure 2.1) is planar and is comprised of five body segments connected together with four, one-degree-of-freedom hinge joints. The bulk of the robot's body is comprised of its chassis, which is locked in place. The two arms are each connected to the chassis at slightly different heights to allow them to slide past each other if their grip flexes sufficiently far inward. Each arm is composed of an upper and lower segment. These segments are attached with a hinge joint that rotates the two arm segments through the horizontal plane, with a range of motion constrained to $[-90^o, +90^o]$. The upper segment is attached to the chassis with a second hinge joint that rotates the entire arm relative to the chassis through the range $[-90^o, +90^o]$. The initial pose of the robot, as shown at the top of Figure 2.1, is considered to set the four joint angles to default values of 0°.

Each of the four joints are equipped with a motor that applies a torque to the joint proportional to the difference between the joint's current angle and the desired angle output by the robot's controller. The robot is equipped with four proprioceptive sensors, which report the current angle of each joint. As each of the joints offers 90% of motion in each direction of motion, the robot is free to evolve many strategies for object manipulation. It is also free to evolve motion strategies that do not manipulate the cylinders.

Vision can be described as, in the most fundamental sense, an instantaneous perception of remote cylinders. For this experiment we chose not to simulate vision, but rather to simulate a simpler set of distance sensors. Distance sensors operate much like visual ones, but instead of detecting variations in colors they detect variations in distance. Furthermore, like vision, distal sensors can be high resolution. Vision here is thus approximated using four sets of 'eyes', which point at -67.5^o, -22.5^o, $+22.5^o$, and $+67.5^o$ relative to the forward

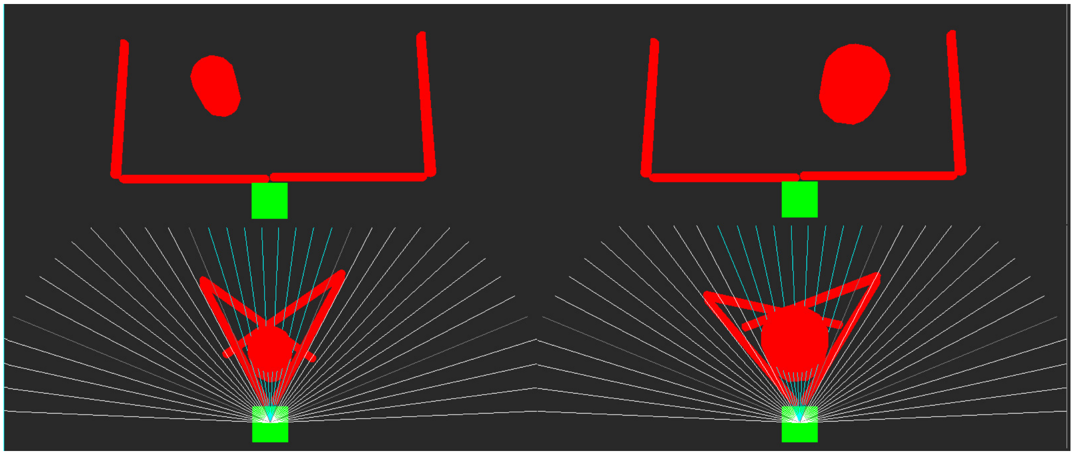

Figure 2.1: Each of the four frames show the robot under different environments. The top frames depict the start of simulations with a small cylinder and a large cylinder respectively. The bottom frames exhibit the rays the robot uses to see cylinders after the robot has gripped the cylinders during its simulation, with the slightly darker ray depicting the center of each eye.

facing direction, arbitrarily considered to be 0^o.

Each eye is compromised of a concentrically-originating horizontal fan of nine equally spaced ($5\,^{\circ}$) apart rays. At each simulation time step a cast ray returns a value linearly proportional to the distance between the source of the ray and the first point of collision. A maximum value is returned if the ray is unobstructed. The nine rays' values are then averaged to provide a visual input value to each input neuron. A visual input value of -1 indicates a total occlusion by a cylinder right in front of the sensors. A visual input value of $+1$ indicates there is no cylinder within viewing range of that eye. A higher resolution of rays was not used due to the linearly increasing computational cost of casting rays. Its unlikely that further resolution would offer the robot more insight as cylinders at the end of the robot's grasp were visible to at least two rays. This indicates that at all material distances there was a gradient of feedback to the visual controllers for both cylinder sizes.

The following equation shows the setup of each of they vision sensors. The term N is the number of rays. The term R is the length of each of the rays. The subscript o refers to the origin of the ray and the subscript c refers to the point of first collision.

$$
\begin{aligned}
d_r &= \sqrt{(x_{r,o} - x_{r,c})^2 + (y_{r,o} - y_{r,c})^2 + (z_{r,o} - z_{r,c})^2} \\
v &= \frac{1}{N} \sum_{r=1}^{N} \begin{cases} 2\frac{d_r}{R} - 1 & \text{if ray } r \text{ collides} \\ 1 & \text{otherwise} \end{cases}
\end{aligned}
\tag{2.1}
$$

2.3 Controller

The robot's controller is a synchronous, deterministic, and real-valued neural network. Figure 2.2 reports its architecture, where each layer is fully connected to the succeeding layer. The middle (hidden) layer is also fully recurrent, obtaining inputs from all five input and five hidden neurons. The output layer's five neurons feed directly from the hidden layer. Four of the five input neurons were designated as sensor inputs. The fifth input neuron was a bias neuron permanently set to the maximum neuron value of one. Four of the five output neurons were used to control the joint motors. The final output neuron is the guess neuron, which had the role of performing the categorization, but did not influence the motion of the robot. At each time step the input neurons were encoded with the current sensor values. Each hidden neuron was then updated using:

$$
h_i^{(t+1)} = \operatorname{erf}\left(\sum_{j=1}^{5} n_j^{(t+1)} w_{j,i} + \sum_{j=1}^{5} h_j^{(t)} w_{j,i} \right) \tag{2.2}
$$

where n_j and h_j are the j^{th} input and hidden neurons, respectively, $w_{j,i}$ is the synaptic weight connecting neuron j to neuron i, and this weighted sum is normalized to a value in $[-1, +1]$ using the Gauss error function. Synaptic weights were restricted to the range $[-1, +1]$.

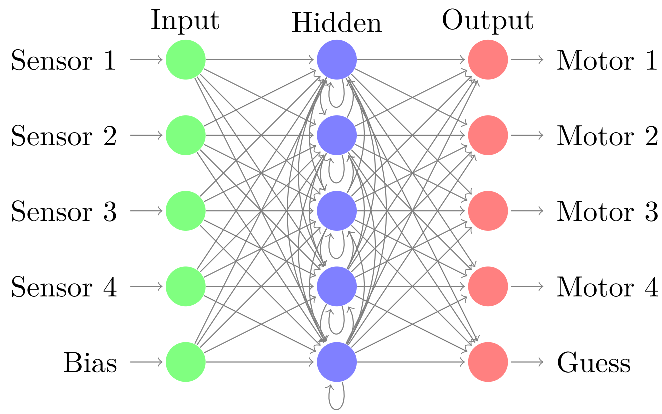

Figure 2.2: Neural Network of Controller

The output neurons were updated using:

$$o_i^{(t+1)} = \operatorname{erf}\left(\sum_{j=1}^{5} h_j^{(t+1)} w_{j,i}\right) \tag{2.3}$$

After the network was updated, the values of the four motor neurons were scaled to values in $[-90^o, +90^o]$ and then translated into torques by the motors, proportional to how far the current angle was from the desired angle. In essence, the motor neurons specify the particular position of each joint. During the evolutionary runs in which the robot is weaned off proprioception and on to vision, some mixture of proprioception and vision is supplied to the sensor neurons, rather than feeding increasingly less proprioception to four sensor neurons and increasingly more vision to an additional four sensor neurons. In this way evolution does not need to learn to ignore or value sets of weights over the evolutionary run.

The primary motivation for this controller was that it facilitated the discovery of motion and categorization behaviors strictly based on our sensory inputs over the course of a simulation. Rules of thumb for the number of hidden neurons in the three layer model indicate that the arithmetic mean, and no more than twice the number of inputs, should be used. Five hidden neurons allow for a variety of behaviors to evolve while limiting the number of synapses that evolution must optimize.

2.4 Evolutionary Algorithm

The Covariance Matrix Adaptation Evolution Strategy (Hansen, 2014) (CMA-ES) was chosen as the real-valued optimization method. In all evolutionary trials, only the synaptic weights in the robot's controller were evolved. All aspects of the robot's cognitive architecture and morphology remained fixed. CMA-ES evolved 75 synaptic weights, each constrained to a range of $[-1, 1]$. The initial synapses vector was initialized with uniformly random weights in the bounded range. CMA-ES specific parameters included initializing each run with a σ of 0.25 and a default of 16 fitness evaluations per generation. Each evaluation was composed of multiple simulations in which the robot was confronted with different cylinders placed at different positions. All other unmentioned parameters were kept at default settings. Since we utilized CMA-ES as a function minimizer, our experiment attempted to minimize the error of the robot's guess as to which class the cylinder currently in front of it belonged to. We shall use the term evolutionary run to refer to the process of evolving (training) our controllers for a given set of environments and sensor modality. Each evolutionary run was seeded randomly for each of the fifty training runs for each environment.

2.5 Environments

The environment of the robots differ primarily through the position and size of the cylinder presented in each simulation. Each robot's controller for a given sensor modality was simulated a specific number of times, which we define as an evaluation. During training the cylinders were placed as described below and shown in Figure 2.3 for each 6-simulation evaluation. The horizontal and vertical environments were chosen because they constrained the training data to one dimension. The alternating environment was chosen because it did not place both a large and a small cylinder at the same positions. Additionally, we also investigated how controllers evolved when exposed to fewer (4) and more (8) simulations. The evaluation types include:

Alternating (A4, A6, A8) The cylinders were placed in a two-cylinder deep rectangle, alternating large and small cylinders, each cylinder with their own unique position. The purpose of this positioning scheme was to investigate the robot's ability to generalize when position can be meaningfully and more readily be inferred to be the task during the training phase.

Horizontal (H4, H6, H8) The cylinders were placed across the X axis such that both sizes were tried at each unique position. This orientation of training positions was chosen to offer the visual system clearly different input for each position.

Vertical (V4, V6, V8) The cylinders were placed across the Z axis such that both sizes were tried at each unique position. This orientation of training positions was chosen to mask, as to minimize, the differences in initial positions.

Testing The cylinders were placed on a Cartesian plane over 78 positions for a total of 156 simulations. This range of cylinders can be considered the reasonable grip-able space.

As an examination into the potential generalizing capabilities of our system we also evolved 100 runs of 16 evenly spaced cylinder positions, including the four corners of the testing data set.

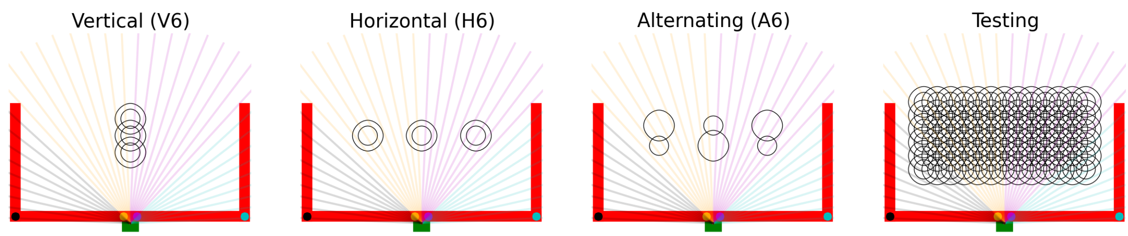

Figure 2.3: This figure depicts the initial positions for the cylinders under various environments. Additionally, the thick red lines indicate the initial position of the robot's limbs. The filled circles represent the locations of the joints. The circle and ray colors correspond to the ordered pairs of sensors combined during the ontological scaffolding: the teal outer joint on the right arm corresponds to the rightmost eye's rays.

2.6 Sensor Modalities and Scaffolding

2.6.1 Proprioception (P)

Robots evaluated under this sensor modality only utilized their proprioceptive sensors (joint angles) as inputs to their controller for the entirety of training and testing. The output motor neurons, which controlled the desired angle of the joints, only indirectly affected the value of proprioceptive input neurons: there was no feedback loop from the output neurons to the input neurons in the controller. Because of this, the proprioceptive controllers primarily relied on colliding with the target cylinders to affect the rate of change and final values of its sensors. Without contact these controllers had no capability of classifying.

2.6.2 Vision (V)

Robots evaluated under this sensor modality only utilized their vision sensors (four eyes composed of distal rays) as inputs to their controller for the entirety of training and testing. Unlike proprioception, visual sensors could process sensor data relevant to classifying the cylinder from the very first simulation time step. Furthermore, the output neurons could evolve to manipulate the cylinders to obtain a different perspective during a simulation.

2.6.3 Scaffolding

Although scaffolding is a common method employed in robotics (Dorigo and Colombetti, 1994; Perkins and Hayes, 1996; Saksida et al., 1997; Bongard, 2011b; Bongard, 2011a), we employed it here in a novel way. During the evolutionary process the robot is forced to rely progressively less on proprioception and progressively more on vision to perform categorization. Four different types of scaffolds were attempted and reported here. For each of scaffolding types, except for the none scaffold (N), a single parameter linearly descends from one to zero over the course of an evolutionary run. This parameter dictates how much proprioceptive input the robot has access to (blue line in Figure 2.4). A second parameter climbs from zero to one over the course of a scaffolding portion evolutionary run and dictates how much visual input the robot has access to (green line in Figure 2.4). The novelty in this approach is that each input neuron is experiencing sensorial scaffolding. Each neuron initially provides the controller proprioceptive input. But, as that input becomes vision, the synapses of the network must adapt to handle the new sensory modality. The hypothesis is that by automatically retaining the behaviors of ACP without any other non-sensory constraints on the controller we can achieve more robust visual controllers.

During testing, the controllers evolved using scaffolding were tested identically to the controllers evolved using the Vision (V) sensor modality. As the controllers are tested on the same input, and they are evolved for the same duration, it is reasonable to compare their capabilities.

During scaffolded evolutionary runs the robot could rely only on proprioception during the initial 30% of training. The next 60% of training time caused a constant linear decrease in the scaffold. During the final 10% of training, the robot could only rely on vision. Each robot evaluation was provided with a fraction that was zero during the first 10% of training, some value in $[0, 1]$ during the next 60% of training, and one for the last 10% of training. This value was used to tune the three scaffolding schedules described next.

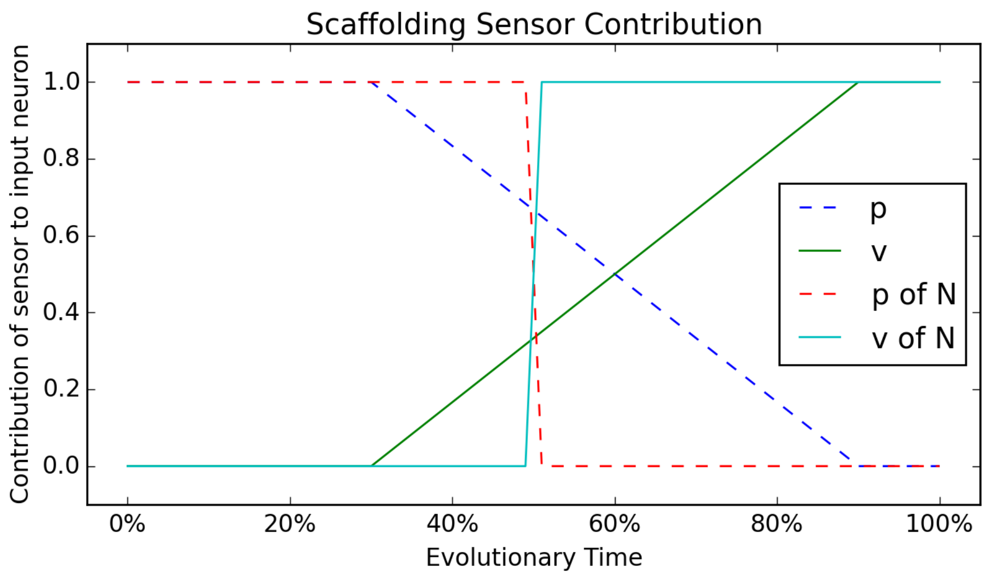

Figure 2.4: The relative contribution of proprioception (P) and vision (V) to a robot's input over the course of an evolutionary run that is scaffolded. This parameter is then used in each simulation as seen in Figure 2.5.

None (N) For the first half of evolutionary time the robot's controller solely received input from its proprioceptive sensors. For the second half of evolutionary time the robot's controller solely received input from its vision sensors. This scaffold operated solely in phylogenetic time.

Melding (X) During the evaluation of an individual robot, the values arriving at the sensor neurons were an admixture of the four proprioceptive and the four visual sensors (Fig. 2.5a). The proportions of both sensor modalities gradually changed over evolutionary time: robots in the first generation obtained 100% proprioceptive input and 0% visual input, robots halfway through an evolutionary run received roughly 50% proprioceptive input and 50% visual input, and robots in the final generation received 100% visual input. This scaffolding operated solely in ontogenetic time.

Swapping (S) Partway through the evaluation of a single robot, its input would switch from proprioception to vision (Fig. 2.5b). The point at which this swap would occur

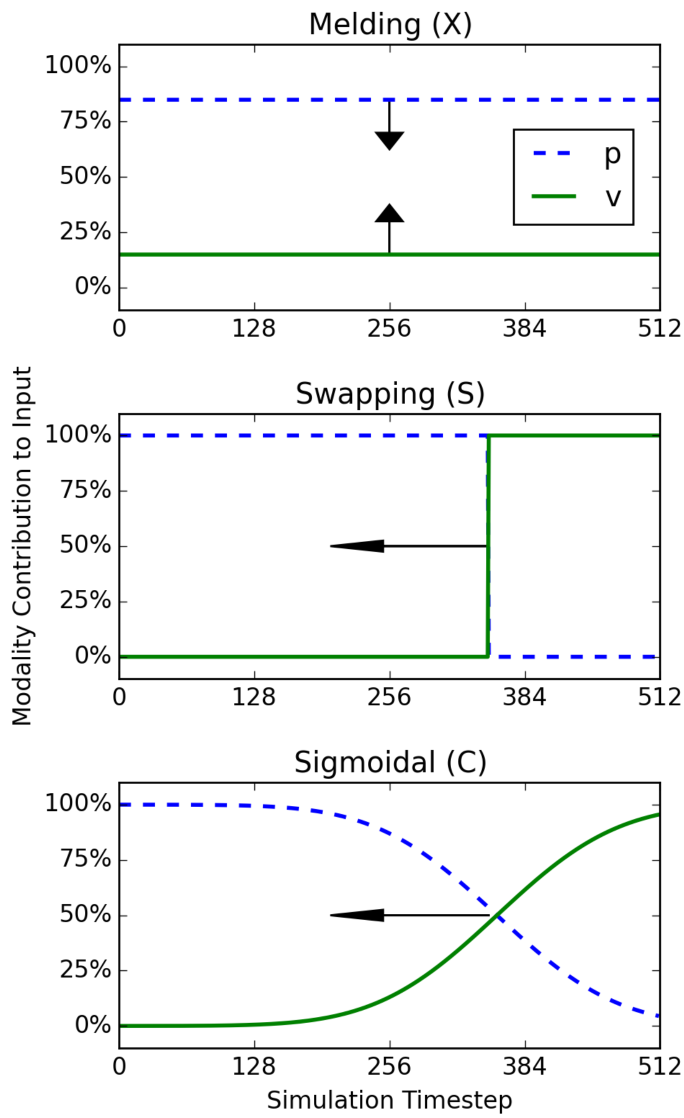

Figure 2.5: Changes in contribution of proprioception (P) and vision (V) during the evaluation of a single controller. The lines represent ontological scaffolding, or the scaffolding that occurs over one simulation of the robot. The arrows represent how the relative contribution of proprioception and vision change as the evolutionary run proceeds. The movement in the direction of the arrows, as described in Figure 2.4, represents evolutionary scaffolding.

changed over evolutionary time: robots in the first generation received only proprioceptive input, robots halfway through an evolutionary run received proprioceptive input for the first 256 time steps and visual input for the last 256 time steps, and robots in the last generation received only visual input. This scaffold operated on both phylogenetic and ontogenetic time.

Sigmoidal (C) A sigmoidal smoothing function was used to determine the amount of contribution of vision and proprioception to the input layer during any single time step of the evaluation (Fig. 2.5c). The shape of this sigmoid was altered over the course of an evolutionary run such that the contribution of proprioception dropped more precipitously—and the amount of visual input increased more precipitously—later during the evolutionary run. Essentially, this scaffold is a combination of the other two scaffolds. This scaffold operated both on phylogenetic and ontogenetic time.

In the case of the sigmoidal smoothing function the contribution of vision to the value of the input neurons is shown in equation 4. The term g represents the current generation in the evolutionary run out of G generations. The term t represents the time step in the simulation out of T time steps.

$$c_v = \frac{\operatorname{erf}\left[4\left(\frac{g}{G} + 2\frac{t}{T} - 1\right) - 2\right] + 1}{2} \tag{2.4}$$

2.7 Fitness

Each simulation lasted 512 time steps in the Bullet Physics Engine (Coumans et al., 2014). The final 10% of values of the controller's guess neuron were recorded and used to compute the controller's fitness. The guess neuron's values were compared against the cylinder's class label (-0.5 for small and 0.5 for large) to obtain a difference. This difference is averaged over the time steps to become our error:

$$e = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{T}\sum_{t=0.9T}^{T}|g_{c,t} - r_c| \tag{2.5}$$

The term C represents the number of cylinders placed and T represents the total number of time steps for an evaluation. $g_{c,t}$ denotes the value of the guess neuron when the robot is simulated under environment c and r_c denotes the relative radius of the cylinder in environment c. ($r = -0.5$ for the small cylinder and $r = 0.5$ for the large cylinder.) An error of zero indicates that the controller is able to settle on a stable value over the final 10% of the robot's evaluation period. Importantly, the category values were not set to the extrema of the neuron's output range (-1 and +1) because these extrema would require coordination from multiple hidden neurons to guarantee that the extra could be reached.

The robot morphology and task were formulated such that there were at least four types of movement that could be used to manipulate cylinders. The robot could choose to not move cylinders by extending its joints outward. The robot could open one of its arms while closing the other to slide cylinders which come into contact with the closing arm away from it. The robot could close its inner joints while keeping its outer joints relatively open, leading to the cylinder becoming trapped in a diamond-like arm pattern. Finally, the robot could fully close both arms, leading to the cylinder becoming trapped in a triangle formed by the arms. This behavior is shown in Figure 2.1. We found that the controller rarely changed its motion strategy partway through a simulation.

2.8 Tests

After evolution, we assessed how robustly a robot could categorize when simulated in novel environments. To do so, we extracted the controller with the lowest training error obtained during the final 10% of the generations from each evolutionary run. This robot was denoted as that run's representative. The representative controllers were then presented with the Testing environment as shown in Fig. 2.3. In these test evaluations the robots were only allowed to use the visual sensors for categorization. The only exception were those runs in which only proprioception was allowed during training; these robots were allowed to use only proprioception during testing. As during training, testing error was calculated using Equation (5), but averaged over 156 simulations instead of four, six, or eight simulations.

In the next section we investigate the intra-category and inter-category distances between cylinders caused by the robot's movement. The following equations describe these values. In each case I and J represent the number of large and small cylinders, respectively.

$$\begin{aligned} D_{\text{intra}} &= \frac{2\sum_{i=1}^{I}\sum_{j=i+1}^{I}\sqrt{(x_i - x_j)^2 + (z_i - z_j)^2}}{I(I-1)} \\ &+ \frac{2\sum_{i=1}^{J}\sum_{j=i+1}^{J}\sqrt{(x_i - x_j)^2 + (z_i - z_j)^2}}{J(J-1)} \end{aligned} \tag{2.6}$$

$$D_{\text{inter}} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{(x_i - x_j)^2 + (z_i - z_j)^2} \tag{2.7}$$

3

Results

In this chapter we report on a total of 5400 evolutionary runs. We evolved the robot's controllers against every environment: the combination of cylinder position (horizontal, vertical, and alternating) and simulation count (4, 6, 8). Robots had six modalities: just proprioceptive input (P), just visual input (V), or evolved against one of the four scaffolding strategies (N, S, X, C). For each of the 54 combinations of cylinder positioning, simulation count, and scaffolding strategy, we performed 100 evolutionary runs. Every evolutionary run was subject to 10,000 fitness evaluations. For robots trained against four, six, and eight cylinders, they were evolved for 40,000, 60,000, and 80,000 robot simulations, respectively.

The average testing errors for the representative controllers are reported in Table 3.1. A robot whose strategy would be to randomly guess the size of its cylinders would have a test error of 0.5. When we refer to robots as memorizing we mean that their test error is high; these robots overfitted the training examples and therefore cannot perform well on the generalized test set.

In most cases, the robots trained with vision (column "V" in Table 1) memorized more than robots trained using one of the scaffolds (columns "N" through "C" in Table 1). However, robots trained with proprioception and then tested using proprioception also memorized on occasion: these robots obtained similarly high testing error as the robots

	P	V	N	S	X	C
32	0.028**	0.057	0.080*	0.094***	0.073	0.080*
A8	0.165***	0.247	0.227	0.226	0.207**	0.184***
H8	0.151***	0.231	0.216	0.246	0.218	0.219
V8	0.172***	0.279	0.265	0.306	0.245^	0.273
A6	0.183***	0.300	0.272	0.276	0.230***	0.239***
H6	0.181*	0.225	0.250	0.253	0.241	0.217
V6	0.222***	0.409	0.351***	0.356***	0.310***	0.339***
A4	0.355	0.372	0.359	0.364	0.364	0.386
H4	0.187***	0.362	0.271***	0.282***	0.258***	0.253***
V4	0.201***	0.412	0.347***	0.351***	0.350***	0.341***

Table 3.1: Test errors for 100 runs over the different cylinder simulations per evaluation, positions, and sensor modalities. The asterisks designate p values below 0.05, 0.01, and 0.001 for one through three asterisks respectively. The îndicates a p value of 0.056. p values were calculated by applying a t-test to the average test errors of vision when compared to other those of the other sensor modalities for each of the environments.

trained and tested with vision in environments H6 and A4. This result implies that, although the task may seem sufficiently simple in that categorization using proprioception will always result in robust categorization in unseen environments, there are movement strategies that evolve for which this is not the case. In this case, the P solutions evolved behaviors that would swing the arms asymmetrically, utilizing feedback from the cylinders' positions to complete the task of deciding their size.

A visual way of interpreting the characteristic behavior of the evolved controllers is through looking at how correctly each position on the testing data set is categorized. Figures 3.1, 3.2, 3.3 show these results. A further analysis of the test errors and movement of the evolved controllers is done through box-plots in Figures 3.4 and 3.5. The discussion of the results will heavily depend on these two sets of box-plots.

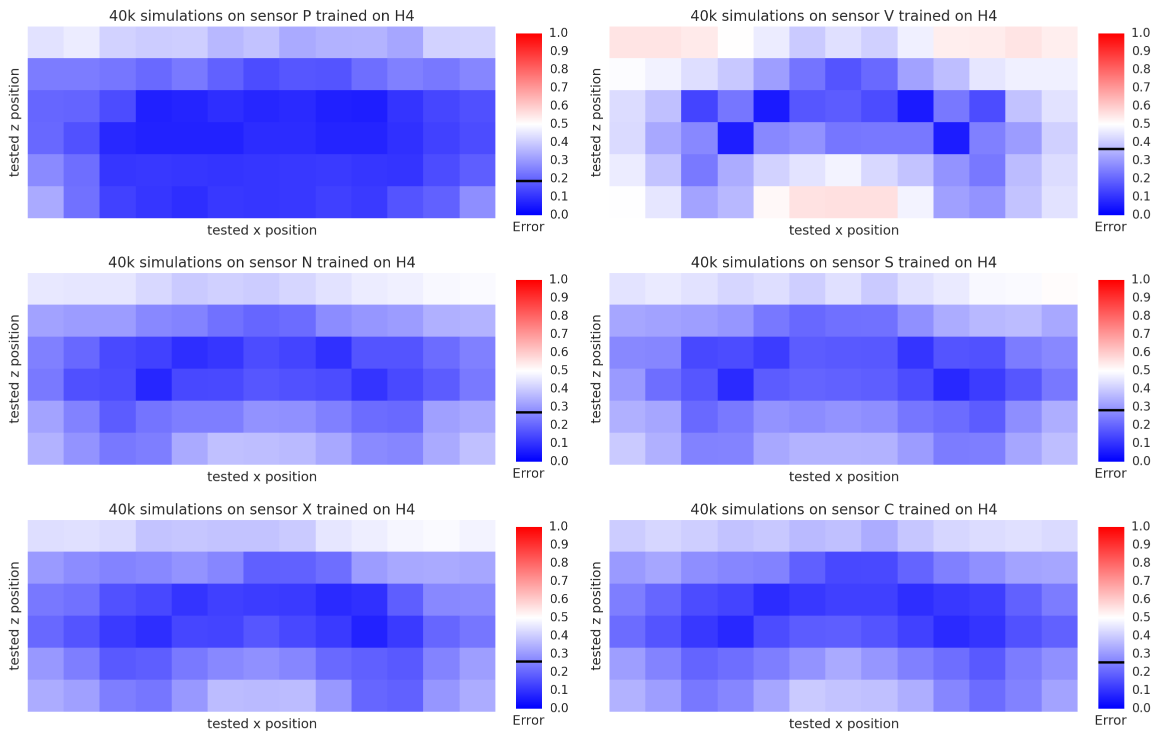

Figure 3.1: The average test fitness for each of the sensor modalities over the 78 test positions as trained on the H4 environment. The vision (V) sensor modality clearly shows the areas in which both cylinders are correctly classified, yet it fails to correctly classify other positions. The proprioceptive modality (P) and the scaffolds show that approaches which were afforded proprioception during training characteristically generalize to a higher extent.

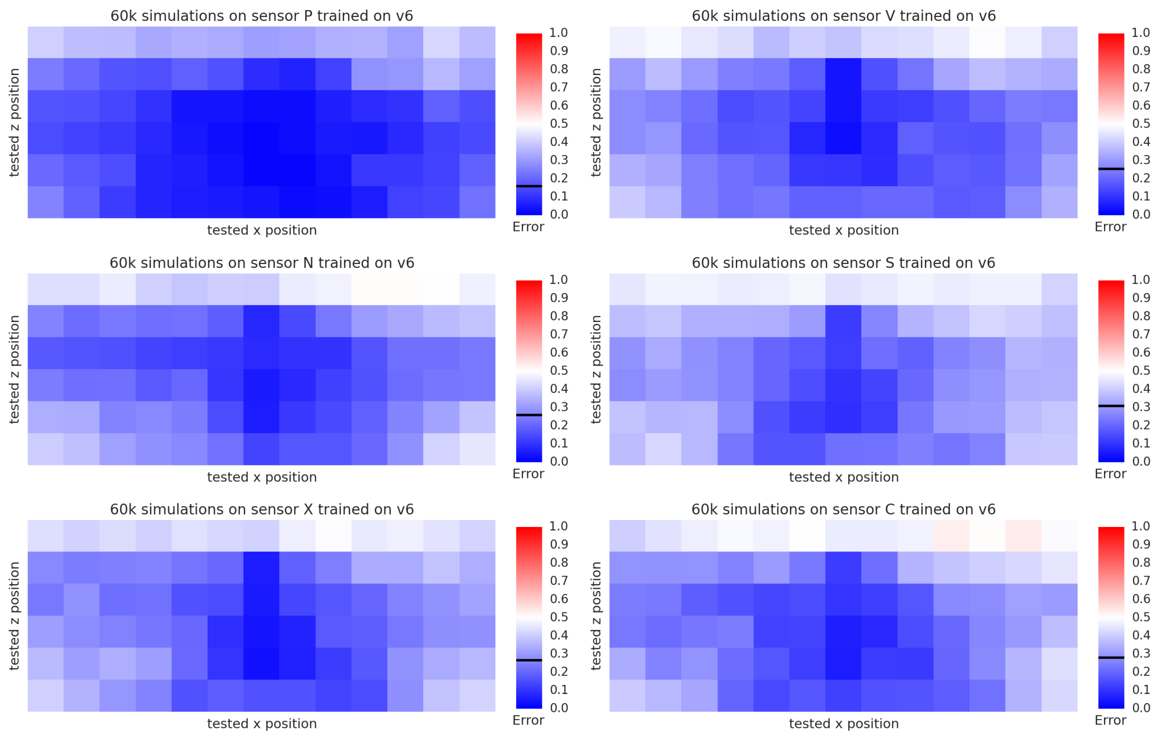

Figure 3.2: The average test fitness for each of the sensor modalities over the 78 test positions as trained on the V6 environment. The vision (V) sensor modality clearly generalizes along the range of training positions. Some of the vision modality's tested controllers can also classify cylinders along the semicircle represented by the relative distal range along the two eyes it was trained on. Any cylinder placed outside of this narrow arc is characteristically unclassifiable. For the scaffolded modalities, a horizontal band successful classification appears. Additionally, cylinders directly in front of the eyes are more readily classified. This indicates that the scaffolded approaches often development motion strategies that generalize along dimensions that the training set does not vary on.

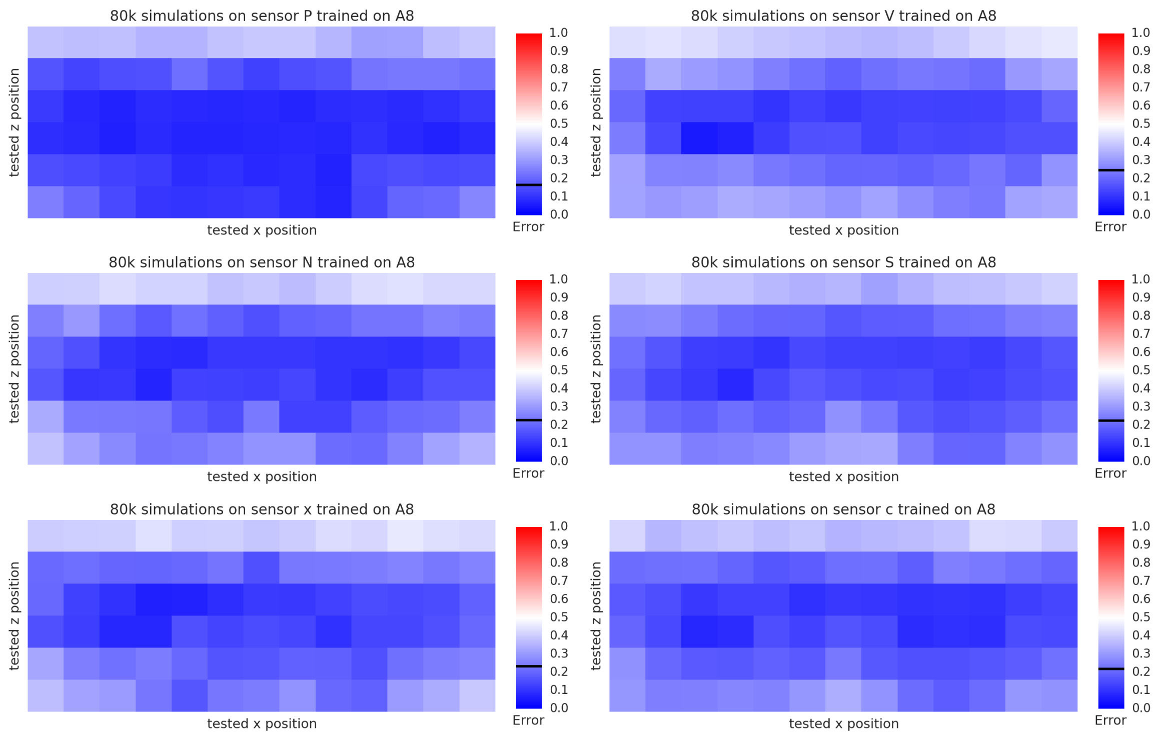

Figure 3.3: The average test fitness for each of the sensor modalities over the 78 test positions as trained on the A8 environment. As the training set was spread out in the graspable region of the robot, both vision and scaffolded modalities achieved higher rates of generalization. However, when we compare the melding (X) and sigmoidal (C) modalities against the visual controller, we see a horizontal band of successful categorizing. Similar to how we saw in the V6 environment, this indicates that motions evolved by the scaffolded modalities are more apt to generalize.

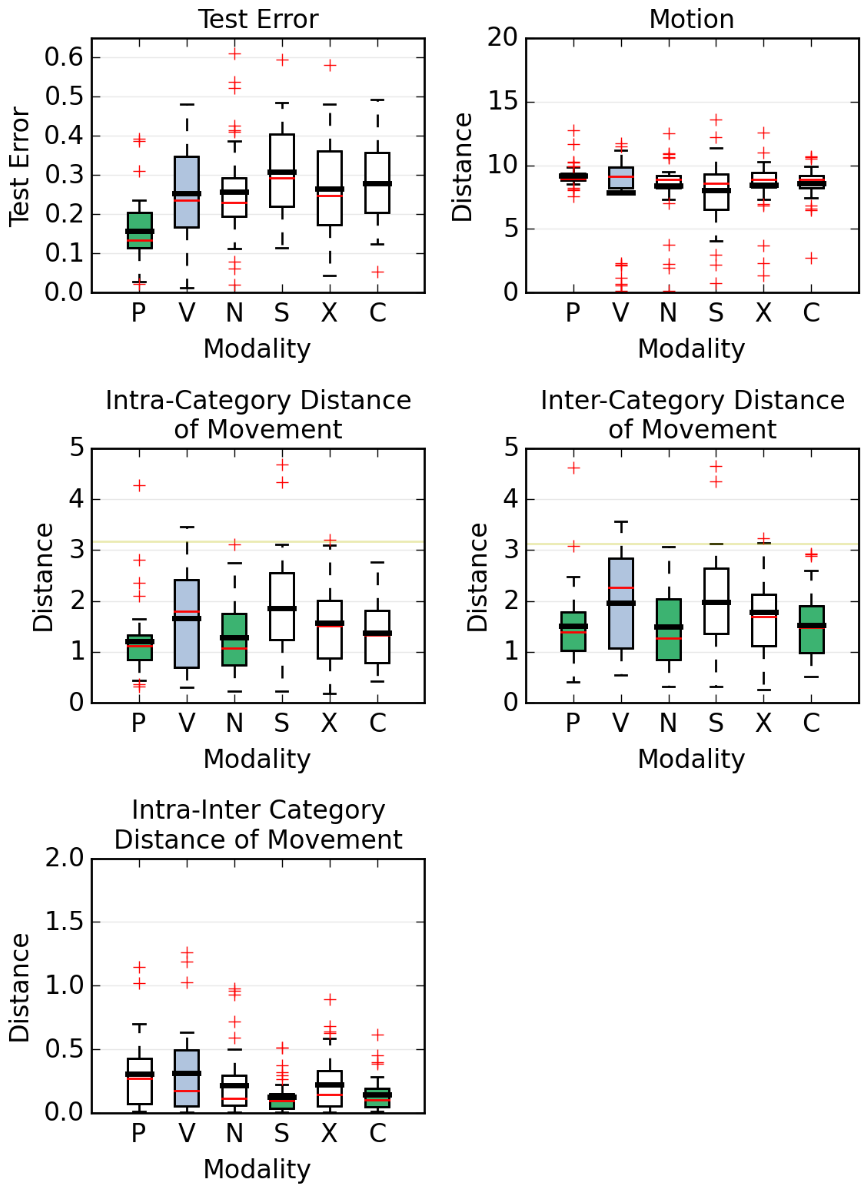

Figure 3.4: Statistics of V6-trained controllers over 60,000 simulations per run. We define motion as the average euclidean distance between the beginning and ending positions of the cylinders during testing simulations. The light blue box-plot represents vision. Green box-plots for each subplot are significantly different than vision at a p level of 0.05. The horizontal red lines designate medians and the thick horizontal black lines designate the mean. In the intra and inter-category graphs the horizontal yellow lines designate what the distances would be if the test cylinders were not perturbed. The box-plot's whiskers represent the 25th and 75th percentiles.

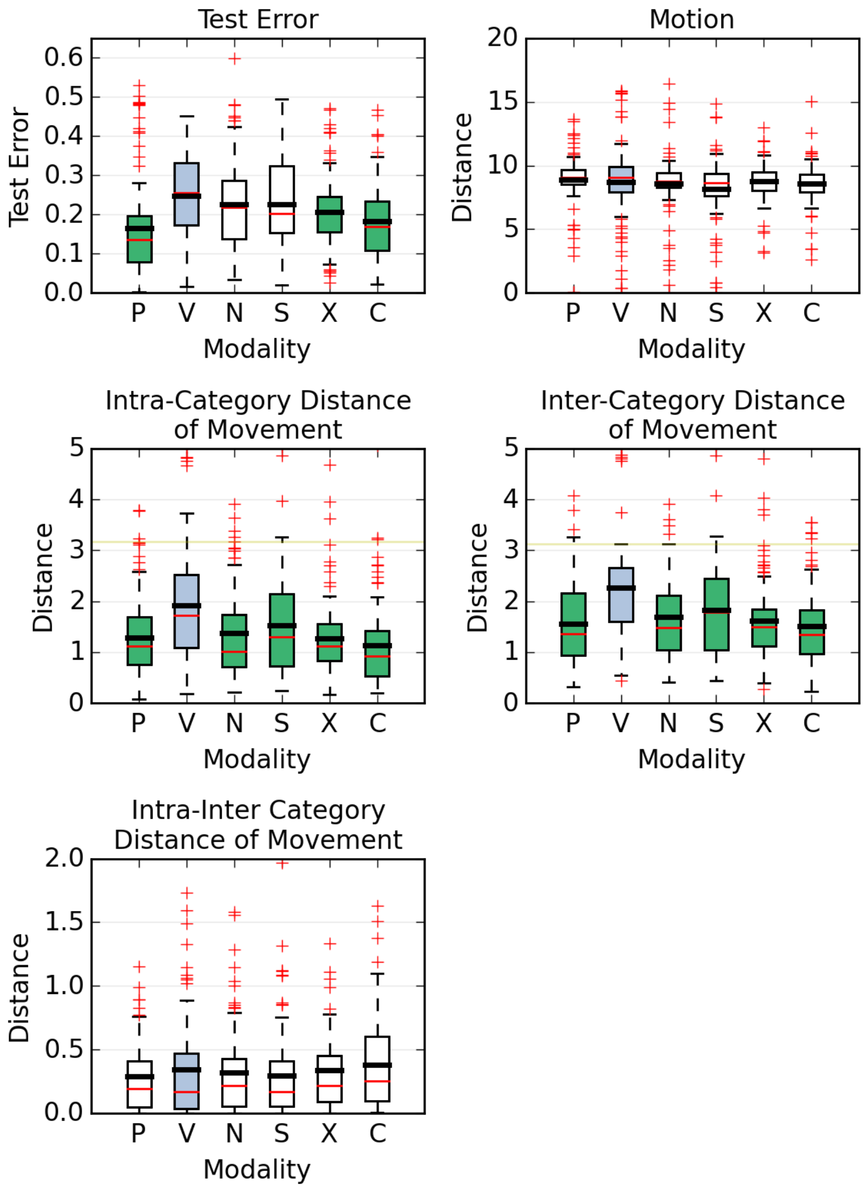

Figure 3.5: Statistics of A8-trained controllers over 80,000 simulations per run.

3.1 Results of the training against the 32-training objects environment

We also looked at training all of our sensor modalities on an environment whose layout of positions was representative of the full range of positions that the cylinders could be

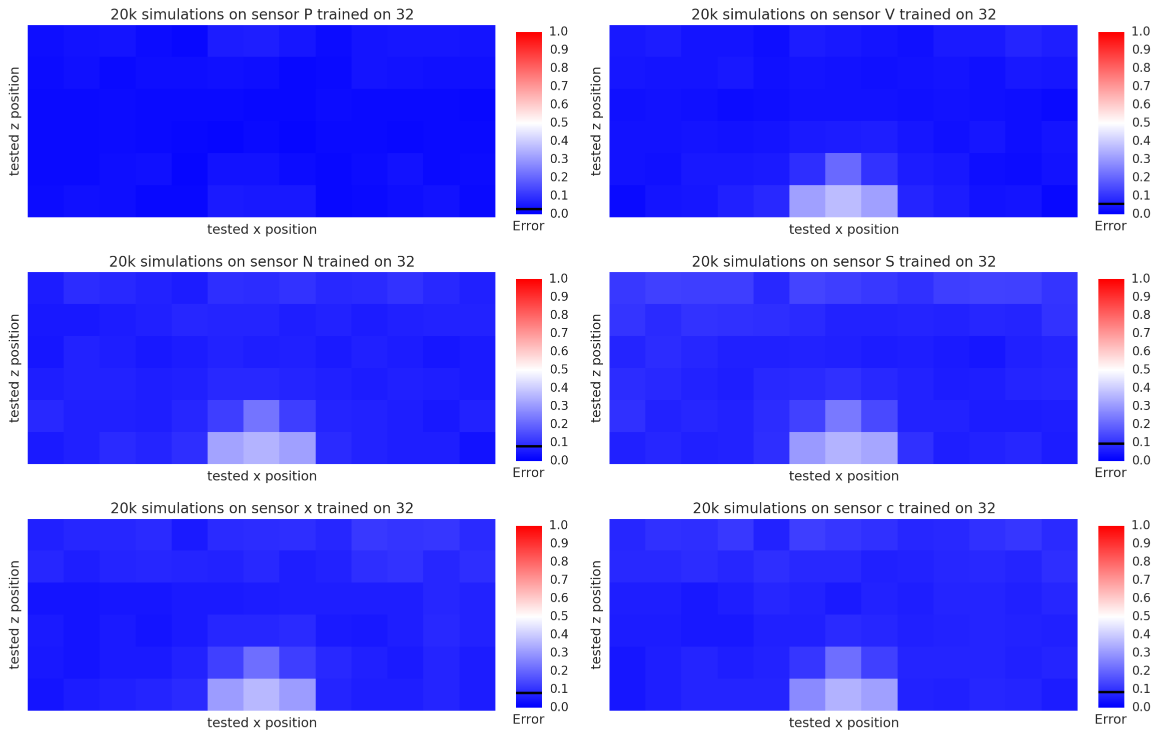

Figure 3.6: The average test fitness for each of the sensor modalities over the 78 test positions as trained on the 32 environment. The visual (V) modality and all the scaffolds characteristically evolved different strategies than proprioception (P) did. This is indicated through the region close to the chassis of the robot having relatively low fitness while the other areas have essentially perfect fitness. As the training set did not include a cylinder right in the middle, evolution has overfitted the training set and is categorically classifying small cylinders as large. These results are unusual as the misclassification appears to happen in over 90% of runs, signifying the ultimately the same behavior is often converged upon. This may imply that our morphology has a limited number of ideal behaviors.

placed at. As seen in Figures 3.6 and 3.7, the best evolved controller was close to perfectly generalizing the training data set in every evolutionary run. While the proprioceptive (P) modality had significantly lower test error than the visual (V) modality, the swapping (S) and sigmoidal (C) scaffolds both had significantly higher error than vision. This represents a departure from the environments with fewer environmental positions.

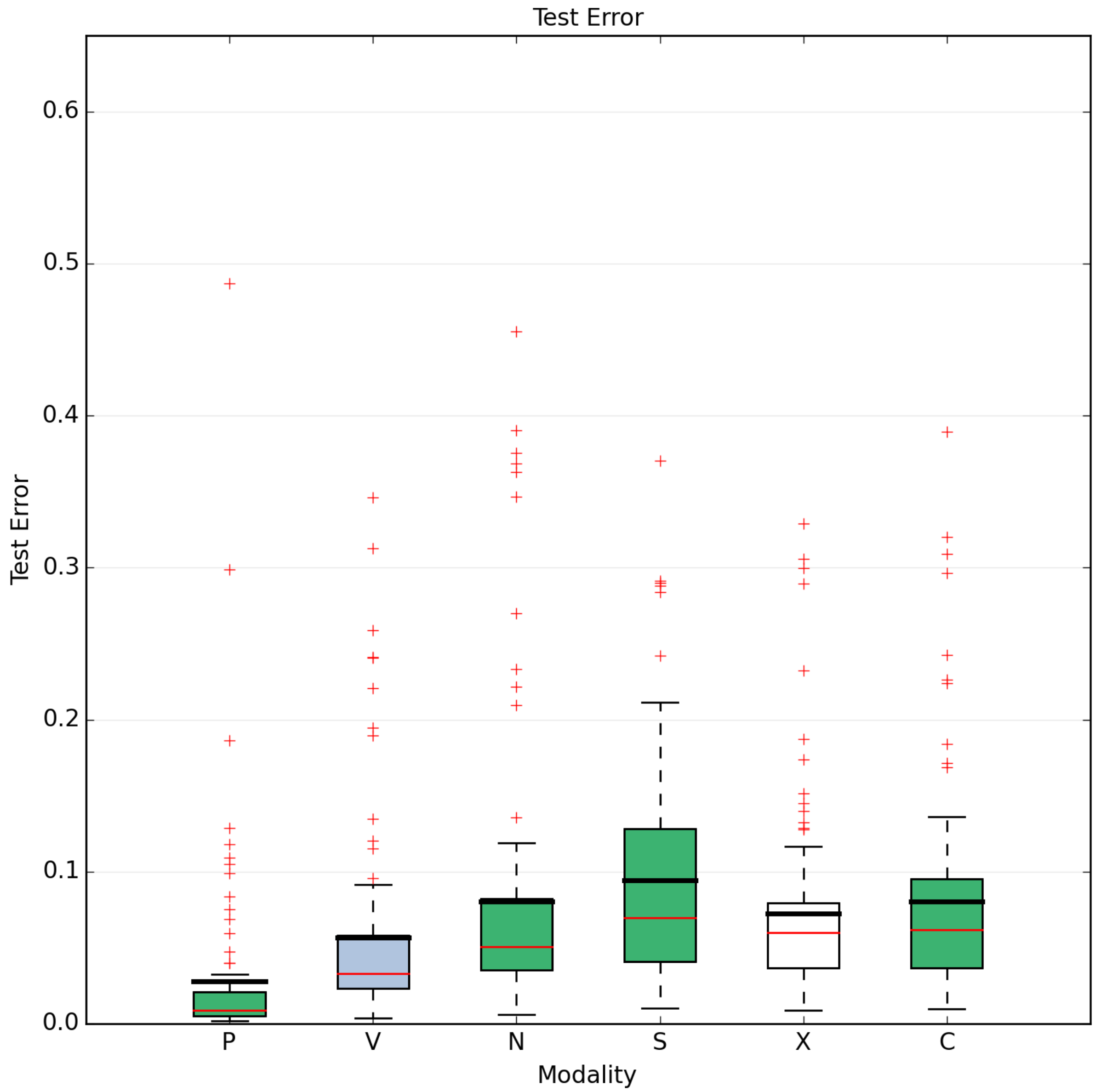

Figure 3.7: Statistics of 32 environment controllers over 320,000 simulations per run.

4

Discussion

4.0.1 Proprioception is a superior generalizer

Thee proprioception trained and tested modality experiences significantly lower average test error than the vision modality as shown in Table 3.1. This result indicates that proprioceptive ACP behaviors tend to generalize more readily than vision-based behaviors. To put the results in perspective, a robot which has a test error of zero is perfectly generalizing. A test error of 0.5 indicates random guessing or always guessing one of the two categories. A statistically significantly lower test mean for a series of evolved robots indicates that brains evolved under that modality tend to exhibit characteristically higher rates of generalizing behavior. As memorizing and generalizing behavior happens on a densely populated continuum for any set of run, the discussion shall not focus on any particular evolved controller, but the set of controllers evolved for the various modalities. To simplify some of our discussion, I shall consider that a test mean below 0.25 indicates characteristically generalizing behavior. Accordingly, higher test means will indicate characteristically memorizing behavior.

The two environments in which there are not significant differences between the non-scaffolded modalities are A4 and H6. The A4 environment is particularly unique in that position can be very readily inferred by evolution as the classification task. With only four positions, the cylinders are placed on the corners of a square. Upon manual examination it was revealed that many proprioceptive runs were classifying the cylinders using feedback from highly-specialized contracting motions. These memorizing behaviors ended up conferring no benefit in fitness for scaffolded runs. This lack of benefit was expected. As proprioception did not discover adequate features, the scaffolding of memorizing features would prime evolved controllers in a negative way. Even though manipulation of the blocks was evolved, scaffolding did not modify proprioception's maladapted movement to evolve generalizing behavior.

The H6 environment was unusual because the visual classifiers evolved to be surprisingly adept at generalizing the categorization task. Inspection of the evolutionary runs indicates that this result was due to five of the six simulations being relatively straightforward to classify without motion. However, the large cylinder placed in view of both middle eyes was likely mistaken for a small cylinder during the initial training generations. Such a fitness cliff would bias search in the direction of generalizing behavior. In Figure 4.1 some example behaviors of best evolved controllers and their respective run are shown. In all the presented examples ACP-style movement was selected for. The upper two evolved controllers' behavior is bring the cylinders to a central spot by closing its limbs. This behavior is similar to that of proprioception's for the same data set, as seen in Figure 4.2. Although the stability of the guess neuron varies, these approaches rapidly converge on their target value. In the lower two runs the behavior of moving the cylinders over a dominant eye is evolved. These behaviors take more time, and therefore the guess neuron doesn't update its values until later in the simulation. In the third chart, one can see attractor behavior on the hidden layer; a large cylinder occluding the eye causes the guess neuron to asymptotically arrive at the large cylinder's value. The attractor is a chain on the hidden layer such that it approaches a value based on its inputs and current value. When the

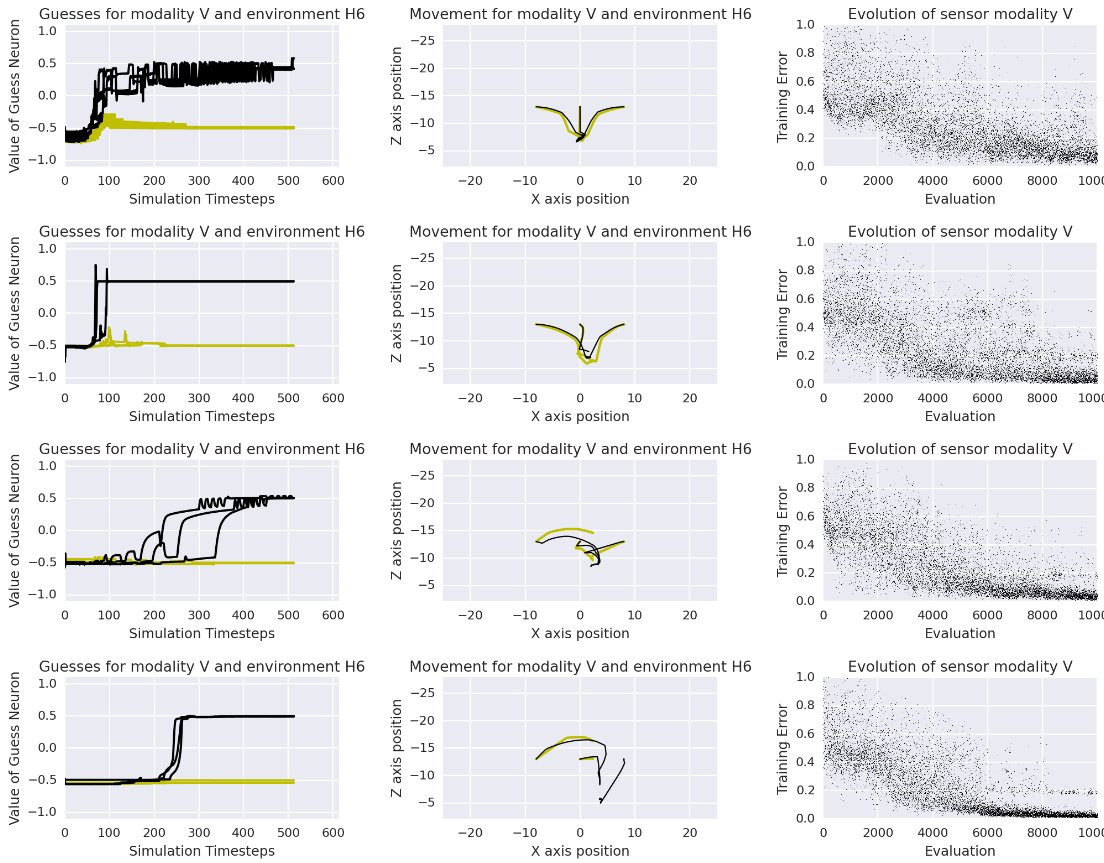

Figure 4.1: These results show the behaviors of some selected best runs of trained using the vision (V) modality on the H6 environment. The black lines represent large cylinders while the yellow lines represent small cylinders. The leftmost graphs show the changes in the values of the guess neurons through simulation. The middle graphs show the corresponding movement of the blocks. The rightmost graphs show the evolutionary algorithm's convergence towards its goal.

guess neuron updates virtually instantly, the guess neuron can be thought of as functionally being in a feed forward network. However, the slower asymptotic approach to a value is indicative of the utilization of memory on the recurrent hidden layer. The convergence rate of the evaluation fitnesses speak to the stability of the evolved behaviors. Large variations indicate the evolution easily disturbs the categorization process through small mutations.

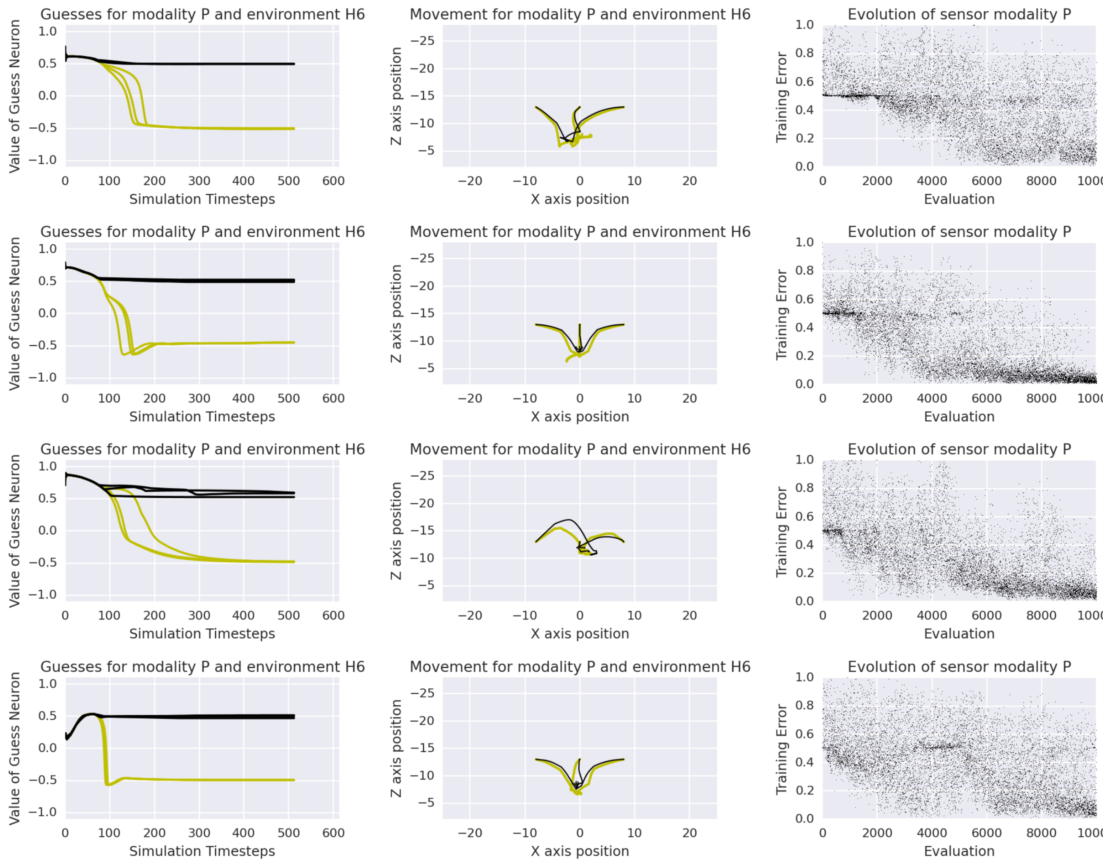

Figure 4.2: These results show the behaviors of some selected best runs of trained using the proprioceptive (P) modality on the H6 environment. The black lines represent large cylinders while the yellow lines represent small cylinders. The leftmost graphs show the changes in the values of the guess neurons through simulation. The middle graphs show the corresponding movement of the blocks. The rightmost graphs show the evolutionary algorithm's convergence towards its goal.

4.0.2 General Scaffolding Results

The first scaffold we looked into was the None, or naive, scaffold. In this scaffold we limited our ACP phase to the first half of generations and the vision phase to the second half. The results showed that due to the scaffolding step the ACP-evolved behaviors contributed to the evolution of controllers that exhibited robust visual categorization in novel environments. This is indicated by the significantly lower testing error obtained when compared to the runs in which only vision was available (V) in a third of the environments as shown in Table

3.1. The behavior exhibited by one of these robustly-categorizing robots is illustrated in Figure 2.1. As can be seen, this robot's evolved behavior of closing its arms together has the effect of moving cylinders at different positions to the same position directly in front of the robot. This has the result of reducing differences in irrelevant properties of the cylinder; here, such an irrelevant difference is the different positions of the cylinders.

In contrast, a robot that does not move will generate no difference in sensor signatures during different cylinder placements if it relies on proprioception for categorization, and very different sensor signatures if it relies on vision. Neither bode well for robust categorization in unseen environments. In the former case, the robot will not be able to successfully categorize even under training environments. In the latter case, there is a danger that the robot will memorize the training environments and fail to generalize to any unseen environments. This highlights the importance of motion for ACP and that proprioception is more likely to lead to active behaviors: a blind robot must move and contact cylinders in order to categorize them.

One limitation of the None scaffold is that it is a harsh change to the experience of the robot. Previously optimized neural weights will not perform well under the new sensory experience. To improve this, we generated a smoother transition schedule as shown in Figure 2.4. When looking at results of the scaffolds in Table 3.1, we see that all the scaffolds are significant improvements over the vision modality for V4, H4, and V6. However, even though the swapping (S) scaffold is more gradual in phylogenetic time, it's still harsh during ontogenetic time as the swapping affects the same latter time steps of each robot's lifetime that are used to decide the value of the guess neuron. The melding (X) and sigmoidal (C) scaffolds are attempts to further smooth the sensory transitions. Both X and C scaffolds exhibit significant improvements on the A6 and A8 environment. In certain environments (H4, V4, V6) the choice of scaffold does not appear to matter. This is due to vision's tendency to readily memorize from the training set. Only when controllers were primed with ACP-style behavior did they tend to evolve more generalizing behavior. When vision needed movement to readily evolve controllers, the X and C scaffolds continued to successfully influence evolution.

The results of the swapping (S) and None scaffold (N) indicate that jarring sensory transitions are not the optimal means to scaffold sensors. The necessary task of the experiment is to have the guess neurons stabilize at a value of either 0.5 or -0.5 for the final 10% of simulation time-steps. If there is a shock on the value of the input neurons, the evolutionary algorithm will find that it needs to relearn certain synapse weights. Unfortunately, evolution does not solely change the weights of one layer, but all of them. This jarring input change breaks evolved asymptotic behavior on any of the hidden neurons. The other output neurons are similarly affected, leading to different behaviors. The X and C scaffolds are not as jarring: the X scaffold's input neurons have small incremental changes in weight contributions over only phylogenetic time and the C scaffold gradually decreases the contribution of the proprioceptive sensors on the input neurons over both ontogenetic and phylogenetic time. As X and C were the most often significantly successful scaffolds, these results indicate that smoothing over ontogenetic time and phylogenetic time is needed. It also shows that changes in scaffolding over ontogenetic time are not necessary to achieve robust results.

4.0.3 Scaffolding success through motion

For the experiment set involving vertical arrangement of six cylinder positions (V6), we obtained some of our most successful results. Since the training set consisted of closely positioned cylinders, vision-evolved controllers had a natural tendency to memorize with little movement. In this sense, the problem was deceptively simple for the visual controllers. As shown in Figure 3.4 both the proprioception and all of the scaffolded runs resulted in significantly more motion during testing. This indicates that when vision favors passive behaviors that do not involve cylinder manipulation, then scaffolding can be a good way to bias search toward movement-based categorization. This movement-bias is retained while the robot transitions to vision, and results in increased robustness of the eventual visual classifier.

One of the primary indicators for whether a controller would generalize was the extent to which it manipulated the cylinder. As shown in Figure 3.4, the motion induced by the vision-based controllers is significantly lower than any observed in the scaffolded runs. The lack of motion in the categorization task is one of the major reasons that the visual classifier was only able to successfully categorize cylinders similarly positioned to its training positions. Without movement during training and testing it was difficult for the controller to extrapolate previously unseen sensory experiences.

The scaffolding process can therefore lead to robust visual classifiers. The efficacy of scaffolding indeed increased as the training set grew increasingly sparse (eight cylinders are reduced to six and then four in Table 3.1 and accordingly the amount of computational effort available was increasingly restricted (from 80,000 robot simulations to 60,000 to 40,000). When fully representative training is performed, such as when 32 objects were used in training and both the proprioception and vision modalities, the results indicate that if both modalities evolve characteristically generalizing behaviors, then the scaffolding step may produce worse results, particularly in the jarring scaffolds.

4.0.4 Scaffolding success in other cases

We also investigated the effect of scaffolding when the visual classifier's motion was not significantly different from robots that relied on proprioception. This was the case for the A8 training regimen, as shown in Figure 3.5. However, even in this case, the sigmoidal (C) scaffolding schedule achieved significantly lower test error than pure vision.

The reason for this is that motion is not a meaningful metric in and of itself. A robot may evolve to move its arms a great deal and push the cylinders away from it in ways that exaggerate the irrelevant feature of cylinder position. To distinguish between helpful and unhelpful motion, we looked at intra-category and inter-category distances. Intra-category distance, the average distance between the final position of a cylinder and every other cylinder in its category, would be low for the behavior shown in Figure 2.1 as the cylinders would be pulled to about the same location. Since cylinders are getting pulled close regardless of size, we would expect to see inter-category distance, the average distance between an cylinder and every other cylinder not in its category, to also decrease a similar amount.

Because the radii of the cylinders are different, we do not expect inter-category distances to be lower than intra-category distances as the centers of the two cylinder sizes would be in marginally different places (25% of the small cylinder's radius) when the cylinders are flush against the robot's chassis. For unhelpful movement, cylinders may be pushed away from a swinging arm or not moved at all: both intra-category and inter-category distances should thus remain high. The results in Figure 3.5 show that the scaffolds that were most successful have intra-category and inter-category differences that are low, like those for proprioception. The unsuccessful scaffold (S) characteristically evolved higher intra-category and inter-category behavior, which were more in line with the same metrics for the pure vision runs (V).

From this, it seems likely that the best predictor of whether a particular run will produce robust visual classifiers is whether the difference between intra-category and inter-category distances is magnified by motion induced by the robot's limbs. Indeed this is what is observed in the results from the V6 training regimen (Figure 3.4).

The types of movements that the scaffolds can help prime is therefore also an important component of whether they lead to robust visual categorization. One signature of whether motion is helpful is if it reduces the separation between intra-category and inter-category differences. Through the lenses of intra-category distances, inter-category distances, and movement levels we could see that characteristically-generalizing ACP-style behaviors were maintained during scaffolding. Even as the inputs were modified, the generalizing behavior held and led to the successful results. Whether generalizing behaviors are more likely to survive scaffolding is an interesting avenue to further explore.

4.0.5 Scaffolding Issues

As shown in Table 3.1, both the vertical and horizontal environments' scaffolds lead to relatively better generalizers as we provide fewer training positions, and therefore less computational power. This highlights vision's inclination towards memorization. In the case of the alternating cylinder positions, a different pattern emerges. In the case of A4, neither proprioception nor any of the scaffolds have significantly different means; proprioception becomes just as much a memorizer as vision. This explains the lack of success of the scaffolds; they do not have a robust categorization strategy from which to begin weaning the robot off proprioception. However, as we add computational power and complexity through A6 and A8, proprioception-based robots memorize less. Even as the environments exhibit greater variation and vision-based controllers memorize, proprioception based controllers resist memorization and are thus still able to be scaffolded. This indicates that even when problems are not constrained to a single dimension of position, there may be success through sensor scaffolding. The underlying pattern for success is whether proprioception can evolve and then pass these successful grasping behaviors to vision. The grasping behaviors that work are the ones that collapse the state space by reducing intra-category and inter-category distances. Scaffolding can be successful when proprioception tends to generalize a problem's state space when vision does not. A remaining unknown is the lack of improvement of the scaffolds in the H8 and V8 environments. When the training is too limited along one dimension, it may also limit the emergence or maintenance of generalizing behaviors during the scaffolding phase.

While the C and X scaffolds had an advantage over the other two scaffolds, this advantage was relatively nuanced in multiple environments. As designing scaffolding strategies is a manual process, the ability to succeed with a variety of scaffolds bodes well for this paper's method.

5

Conclusions and Future Work

The results have demonstrated that a robot may be gradually transitioned from ACP to visual classification. Furthermore, the scaffolds have shown to be improvements over a classifier not trained with ACP.

It is important to recognize that both the ACP and vision sensors of these robots convey relatively little information to the robots. While low-information sensors require fewer neural synapses to optimize, they do not represent the complexity of inputs fed into existing visual classifiers. One limitation of the current network and evolutionary method was its long training time. If we were do design a system to handle larger input spaces, we would need to adapt our neural network and training algorithm. The existing neural networks would otherwise be very difficult to evolve as dense and recurrent neural networks exponentially explode in parameters as they grow. Future experiments can potentially use scaffolding to train convolution neural networks' feature representation (Krizhevsky et al., 2012). Multimodal deep learning systems have in the past been trained to concurrently extract features from both video and audio input streams (Ngiam et al., 2011). It remains possible that we can similarly feed both ACP and vision during training and test these systems without ACP. On a simpler level, the current approach for distal vision operates through the casting of rays, which are computationally expensive on a CPU. Moving more of the computing logic to a GPU or devising what the distal sensors should see based on the position and shape of the presented objects may greatly reduce training and testing time.

The human brain can often project what certain senses should or will be feeling based on the experience of another. For instance, when one sees a ball, they can also project how the ball would feel and move if lifted. One way of scaffolding a visual system would be to train it to predict the feel of another lower-feature sensor. One such approach would be to predict the motion of an object. Algorithms have been developed that can rudimentarily convert two-dimensional images into three dimensional cylinders (Wei, 2005). These predicted three dimensional representations could be manipulated. The predicted behavior could be compared to that of a supervised training set to gauge how well the sensor essentially predicts what the other would feel. Once these predictions are adequately trained, a visual learner could utilize them to assist in categorization.

Another limitation is the systems' current ability to categorize many classes due to the large number of additional synapses, and eventually hidden neurons, that would need be added as we expanded the number of guess neurons. Assuming we were not using convolutional neural nets, we would need to limit the number of classifying neurons. Bounding boxes (Tuci et al., 2010) could potentially limit the number of classifying neurons by tracking their behavior and classifying the robot's experience ranges. Another approach could be to evolve novel behavior, such as the final position of the object or arms, in response to different objects. These behaviors could potentially also be evolved through the use of a fitness function that rewards categorization through the utilizing various affordances of supplied objects.

Another direction of future work would be to further explore the effects of the various general types of scaffolding. Both morphological and environmental scaffolding were unexplored in this thesis. Research of infants has shown that from an early age they can grasp cylinders moving at them. From the environmental scaffolding perspective, this can be utilized to create categorization agents that better understand affordances and are able to manipulate a larger area of objects. Another form of environmental scaffolding would be to modify the number of simulations per fitness evaluation over an evolutionary run. Infants also grow in size and strength as they age. From a morphological scaffolding perspective, arms akin to those of a human's might not be an ideal morphology for categorizing with ACP. Furthermore, they may be less than optimal for scaffolding. Previous work into morphological scaffolding has shown that evolving morphology can facilitate the acquisition of ACP in robots (Bongard, 2010). The evolution of morphology could therefore discover new useful combinations of controllers, morphology, sensors, and actions. The scaffolding approaches themselves could also be incorporated into the evolutionary algorithm. Evolving more efficient transitions from embodied to non-embodied categorization would confer evolutionary advantages along both of our studied timescales. By being able to evolve the scaffolding schedule we would not need to parameterize it, and therefore remove assumptions and potential limitations to scaffoldability.

The final and most aspirational example of future work would be the development of embodied language understanding (Fischer and Zwaan, 2008) or embodied symbol manipulation (Lakoff and Núñez, 2000). Such a system would be built upon a pedestal of sensor-motor signals. As this language would be directly grounded in sensor-motor signals, further language could be evolved or developed to then automatically and gradually transition to increasingly abstract reasoning with the grounded symbols. This grounding could proceed to develop a machine's understanding of affordances and potentially a human-like understanding of reality.

Bibliography

Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. (2009). Cognitive developmental robotics: a survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34.

Ashby, F. G. and Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37(3):372–400.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1):1–127.

Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 4. springer New York.

Bongard, J. (2010). The utility of evolving simulated robot morphology increases with task complexity for object manipulation. *Artificial Life*, 16(3):201–223.

Bongard, J. (2011a). Morphological and environmental scaffolding synergize when evolving robot controllers: artificial life/robotics/evolvable hardware. In *Proceedings of the 13th annual conference on Genetic and Evolutionary Computation*, pages 179–186. ACM.

Bongard, J. (2011b). Morphological change in machines accelerates the evolution of robust behavior. *Proceedings of the National Academy of Sciences*, 108(4):1234–1239.

Bongard, J., Zykov, V., and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121.

Ciresan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE.

Cohen, H. and Lefebvre, C. (2005). *Handbook of categorization in cognitive science*. Elsevier.

Coren, S. (2003). *Sensation and perception*. Wiley Online Library.

Coumans, E. et al. (2014). Bullet Physics Library, version 2.82. http://bulletphysics.org.

Dorigo, M. and Colombetti, M. (1994). Robot Shaping: Developing Autonomous Agents Through Learning. *Artificial Intelligence*, 71(2):321–370.

Fischer, M. H. and Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*, 61(6):825–850.

Fitzpatrick, P. and Metta, G. (2003). Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 361(1811):2165–2185.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning about objects through action: initial steps towards artificial cognition. In *IEEE Intl Conf on Robotics and Automation*, volume 3, pages 3140–3145. IEEE.

Floreano, D., Husbands, P., and Nolfi, S. (2008). Evolutionary robotics. In *Springer handbook of robotics*, pages 1423–1451. Springer.

Fumerton, R. (1988). The internalism/externalism controversy. *Philosophical perspectives*, pages 443–459.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*.

Haith, M. M. (1993). Preparing for the 21st century: Some goals and challenges for studies of infant sensory and perceptual development. *Developmental Review*, 13(4):354–371.

Hansen, N. (2014). Cma evolution strategy, 1.1.05. https://www.lri.fr/ hansen/cmaes_inmatlab.html.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Harnad, S. (2003). Categorical perception. *Encyclopedia of cognitive science*, 67(4).

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11(10):428–434.

Hornby, G. S., Globus, A., Linden, D. S., and Lohn, J. D. (2006). Automated antenna design with evolutionary algorithms. In *AIAA Space*, pages 19–21.

Jackson, P. (1986). *Introduction to expert systems*. Addison-Wesley Pub. Co., Reading, MA.

Jostmann, N. B., Lakens, D., and Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological science*, 20(9):1169–1174.

Koos, S., Mouret, J.-B., and Doncieux, S. (2010). Crossing the reality gap in evolutionary robotics by promoting transferable controllers. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 119–126. ACM.

Koos, S., Mouret, J.-B., and Doncieux, S. (2013). The transferability approach: Crossing the reality gap in evolutionary robotics. *Evolutionary Computation, IEEE Transactions on*, 17(1):122–145.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lakoff, G. and Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books.

Lipson, H., Bongard, J. C., Zykov, V., and Malone, E. (2006). Evolutionary robotics for legged machines: From simulation to physical reality. In *IAS*, pages 11–18.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental Robotics: a survey. *Connection Science*, 15(4):151–190.

Lungarella, M. and Sporns, O. (2005). Information self-structuring: Key principle for learning and development. In *Development and Learning, 2005. Proceedings. The 4th International Conference on*, pages 25–30. IEEE.

McCarty, M. E., Clifton, R. K., Ashmead, D. H., Lee, P., and Goubet, N. (2001). How infants use vision for grasping objects. *Child development*, 72(4):973–987.

McNerney, S. (2011). A brief guide to embodied cognition: why you are not your brain. *Scientific American, Nov*, 4(1).

Minsky, M. L. and Papert, S. A. (1987). *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*. MIT press Boston, MA:.

Mitchell, T. M. (1997). Machine learning. wcb.

Nelson, A. L., Barlow, G. J., and Doitsidis, L. (2009). Fitness functions in evolutionary robotics: A survey and analysis. *Robotics and Autonomous Systems*, 57(4):345–370.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *IEEE CVPR*.

Nolfi, S. and Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. MIT press.

Perkins, S. and Hayes, G. (1996). *Robot shaping: Principles, methods and architectures*. University of Edinburgh, Department of Artificial Intelligence.

Reil, T. and Husbands, P. (2002). Evolution of central pattern generators for bipedal walking in a real-time physics environment. *IEEE transactions on Evolutionary Computation*, 6:159–168.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Rosch, E. (1999). Principles of categorization. *Concepts: core readings*, pages 189–206.

Rosch, E., Thompson, E., and Varela, F. J. (1992). *The embodied mind: Cognitive science and human experience*. MIT press.

Saksida, L. M., Raymond, S. M., and Touretzky, D. S. (1997). Shaping robot behavior using principles from instrumental conditioning. *Robotics and Autonomous Systems*, 22(3):231–249.

Scheier, C. and Pfeifer, R. (1995). Classification as sensory-motor coordination. In *Advances in Artificial Life*, pages 657–667. Springer.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(03):417–424.

Singer, M. A. and Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, 16(2):85–89.

Slepian, M. L., Weisbuch, M., Rule, N. O., and Ambady, N. (2010). Tough and tender embodied categorization of gender. *Psychological Science*.

Slocum, A. C., Downey, D. C., and Beer, R. D. (2000). Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In *From animals to animats 6: Proceedings of the sixth international conference on simulation of adaptive behavior*, pages 430–439.

Steels, L. (2008). The symbol grounding problem has been solved. so whatś next. *Symbols and embodiment: Debates on meaning and cognition*, pages 223–244.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Tuci, E., Massera, G., and Nolfi, S. (2010). Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. *IEEE Transactions on Evolutionary Computation*, 14(6):885–899.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, pages 433–460.

von Hofsten, C. and Rönnqvist, L. (1988). Preparation for grasping an object: a developmental study. *Journal of Experimental Psychology: Human Perception and Performance*, 14(4):610.

Wei, Q. (2005). Converting 2d to 3d: A survey. In *International Conference, Page (s)*, volume 7, page 14.