Representation in Natural and Artificial Agents: An Embodied Cognitive Science Perspective*

Rolf Pfeifer and Christian Scheier

AILab, Computer Science Department, Universität Zürich,
Winterthurer Straße 190, CH-8057 Zürich, Switzerland

Z. Naturforsch. **53c,** 480–503 (1998); received May 18, 1998

Embodied Cognitive Science, Representation, Sensory-Motor Coordination, Self-Generated Data, Morphology

The goal of the present paper is to provide an embodied cognitive science view on representation. Using the fundamental task of category learning, we will demonstrate that this perspective enables us to shed new light on many pertinent issues and opens up new prospects for investigation. The main focus of this paper is on the prerequisites to acquire representations of objects in the real world. We suggest that the main prerequisite is embodiment which allows an agent – human, animal or robot – to manipulate its sensory input such that invariances are generated. These invariances, in turn, are the basis of representation formation. In other words, the paper does not focus on representations per se, but rather discusses the various processes involved in order to make learning and representation acquisition possible.

The argument structure is as follows. First we introduce two new perspectives on representation, namely frame-of-reference, and complete agent. Then we elaborate the complete agent perspective and focus in particular on embodiment and situatedness. We argue that embodiment has two main aspects, a dynamic and an information theoretic one. Focusing on the latter, there are a number of implications: Representation can only be understood if the embedding of the neural substrate in the physical agent is known, which includes morphology (shape), positioning and nature of sensors. Because an autonomous mobile agent in the real world is exposed to a continuously changing high-dimensional stream of sensory stimulation, if it is to learn category distinctions, it first needs a focus of attention mechanism, and then it must have a way to reduce the dimensionality of this high-dimensional sensory stream. Learning is very hard because the invariances are typically not found in the sensory data directly – the classical problem of object constancy: it is a so-called type 2 problem. Rather than trying to improve the learning algorithms – which is the standard approach – the embodied cognitive science view suggests a different approach which focuses on the nature of the data: the agent is not passively exposed to a given data distribution, but, by exploiting its body and through the interaction with the environment, it can actually *generate* the data. More specifically, it can generate correlated data that has the property that it can be easily learned. This learnability is due to redundancies resulting from the appropriate interactions with the environment. Through such interactions, the former type 2 problem is transformed into a type 1 problem, thus reducing the complexity of the learning task by orders of magnitude. By observing the frame-of-reference problem we will discuss to what extent these invariances are reflected – *represented* – in the "neural substrate", i.e. the internal mechanisms of the agent. It is concluded, that representation is not a concept that can be studied in the abstract, but should be elaborated in the context of concrete agent-environment interactions. These ideas are all illustrated with examples of natural agents and artificial agents. In particular, we will present a suite of experiments on simulated and real-world artificial agents instantiating the main arguments.

* This communication is a contribution to the workshop on "Natural Organisms, Artificial Organisms, and Their Brains" at the Zentrum für interdisziplinäre Forschung (ZiF) in Bielefeld (Germany) on March 8–12, 1998.

Reprint requests to Prof. R. Pfeifer.
Fax: 1-6356809.
E-mail: pfeifer@ifi.unizh.ch.

1. Introduction

In recent years classical ideas of representation as endorsed by logical paradigms, by classical artificial intelligence, cognitive science, and cognitive psychology, have been shown not to be without problems. The epitome of the classical approach is the notion of expert systems. The main underlying

0939–5075/98/0700–0480 $ 06.00

idea is that the knowledge of the expert can be extracted from the expert (a process called knowledge acquisition), formalized and represented in a computer system. This leads to the "knowledge" of the expert being represented in the symbol structures of a computer program. The symbols or symbol structures in the computer program are set into correspondence with the outside world. This position is sometimes called cognitivistic.

Criticisms of the classical view of representation abound and we have no intention of reproducing them here (e.g. Brooks, 1991; Bursen, 1980; Clancey, 1997; Franklin, 1995; Hendriks-Jansen, 1996). Rather than trying to demonstrate why the classical view of representation is problematic, we intend to propose a view of representation that promises to shed new light on some of the hard problems: embodied cognitive science (Pfeifer and Scheier, in press). When pursuing the approach of embodied cognitive science, it turns out that representation is no longer a stored structure that can be investigated separately, but something that is part of a complete agent interacting with its ecological niche.

Before we start, a note on the term representation as it will be used it in this paper is in place. According to Newell (1990, p. 59) the essence of representation is "... to be able to go from something to something else by a different path when the originals are not available." He casts this as a general law, the representation law $decode[encode(T)(encode(X))] = T(X)$ where X is the original external situation and T is the external transformation. This view largely corresponds to the common-sense understanding of representation. If you are trying to understand and generalize some phenomenon, you use a particular type of representation. Meteorologists use numbers to represent the amount of rain expected, data base designers use linguistic labels like first and last names, street names, department names to represent the customers, and numbers to designate the monthly income of an employee, electrical engineers use schematics to represent the essential characteristics of electronic circuits, composers use musical notes to represent particular sound patterns, and truck drivers use maps to represent the network of roads and streets. Within cognitive science, psychologists and ethologists use statistical models and differential equations, brain researchers also use differential equations and they strongly depend on visualization and imaging techniques: representing the to-be-explained phenomenon using methods from computer graphics. Note that all of these representations make some form of abstraction. If there is no abstraction, it is the same thing, not a representation. In this sense, representation is much like a model. This is all unproblematic unless a representation is taken to be a *mental representation*, i.e. as part of the mechanisms that generate behavior.

In this paper we will not discuss mental representation in the traditional sense in detail because it is, by definition, not linked to a particular physical medium such as a body but can be studied independently at the level of computational (algorithmic) processes. Since one of our main points will be embodiment, we cannot sensibly discuss the classical concept of mental representation. But we will briefly illustrate the basic ideas of classical representation using the example of a connectionist model of category learning, ALCOVE (see below). Also, we will not discuss symbolic representations. We follow Clancey (1997) in stating that symbolic representations are externally created by humans – as in the examples given above – and have to be distinguished from the mechanisms that actually generate behavior. According to Clancey, confounding the two would constitute a category error. A large part of the literature in cognitive science is concerned with symbolic representations. We will mostly talk about neural correlates of behavior or more generally about internal processes that – in the interaction with the real world – lead to certain behaviors. To use Cummins's (1989) term, the one aspect of representation that we are mostly interested in here is the one of *covariance*. This view is very natural and widely held in neurobiology. "How do we decide, for example, that a certain neural structure in the visual cortex of a frog is a motion detector? Roughly, we notice that a certain characteristic activity in the structure covaries with the presence of moving objects in the frog's field of vision. Given this fact, it seems natural to suppose that what makes that structure a motion detector is just the fact that it fires when there is motion in the frog's visual field." (Cummins, 1989, p. 9). Thus, the neural structure in question can be said to represent the occurrence of motion in the frog's

visual field, or – to use biological jargon – it codes for motion[1]. The main goal of this paper is to explore ways of understanding the interrelationship between behavior and internal processes in an agent.

We start by outlining the basic assumptions of embodied cognitive science. This includes a description of the frame-of-reference problem, the principle of complete agents, and a discussion of the importance of morphology for understanding representation. Then we discuss the implications of embodiment, that can be either dynamic or information theoretic, focusing on the information theoretic aspect. We then introduce some of the hard problems for representation, object constancy and the problem of a continuously changing stream of sensory stimulation. We then discuss the potential solution suggested by a complete agent perspective: interaction with the environment, in particular sensory-motor coordination. The consequences of this perspective are fundamental. They will be discussed using a suite of experiments on real and simulated robots. It is concluded that the complete agents perspective provides new insights into the thorny problems of representation and explanation of behavior in general.

2. Embodied Cognitive Science

Two key aspects of embodied cognitive science are (a) the frame-of-reference problem, and (b) the complete agent perspective. We first briefly introduce the frame-of-reference problem and then immediately move on to the complete agent perspective.

2.1 The frame-of-reference problem

When building a model of a natural system there are always a number of "participants" involved: the subject of investigation (e.g. the human infant in a project on category learning), the observer (typically the scientist conducting the investigation), the model designer (often identical with the observer), the artifact (the computer simulation program or the robot), and the environment. The frame-of-reference problem conceptualizes the relation between these "participants". There are three main aspects of the frame-of-reference problem:

(1) *Perspective issue:* We have to distinguish between the perspective of an observer looking at an agent and the perspective of the agent itself. In particular, descriptions of behavior from an observer's perspective must not be taken as the internal mechanisms underlying the described behavior. But the observer can try to understand what the world looks like from the agent's perspective.

(2) *Behavior-vs.-mechanism issue:* The behavior of an agent is always the result of a system-environment interaction. It cannot be explained on the basis of internal mechanisms only. In other words, it cannot be reduced to internal representation.

(3) *Complexity issue:* The complexity we observe in a particular behavior does not reflect the complexity of the underlying mechanisms.

These points are illustrated in focus box 1. Further illustrations of this principle will be provided as we go along. Note that all three aspects include the distinction between description of external behavior and the internal mechanism underlying that behavior. There is an unresolved dispute of where the description ends and the mechanism begins. After all, the only thing we ever have is a description, at least in the analytical disciplines. When we describe the mechanisms by which an ant generates its behavior, we provide a description of a rule that might govern its behavior (in the case of Simon's ant on the beach: if there is an obstacle on the right, turn left (and vice versa), see focus box). In the area of autonomous robots, the situation is different. There, we not only have a description, but we have built a mechanism that actually underlies the behavior of the robot. This is one of the essential features of the synthetic approach of embodied cognitive science which makes it potentially extremely powerful.

[1] This aspect strongly contrasts, for example, with the similarity view, which is typical for analogical, image-based, and pictorial representations, where the representation itself bears some similarity to the thing in the real world it is supposed to represent. The concept of similarity requires a kind of metric. Often, properties of the perceptual system are implicitly assumed to provide the metric (street maps and cities "look" similar to humans, especially if viewed from a bird's eye perspective). We will not elaborate these aspects any further. For details, see Cummins (1989).

Box 1: Simon's ant on the beach

Simon has used the metaphor of an ant to illustrate some basic principles of behavior (Simon, 1969). We use his metaphor to illustrate the three aspects of the frame-of-reference problem outlined in the main text. Let us assume that an ant starts on the right and its nest is somewhere on the left. So, roughly the direction it travels is from right to left. Figure 1 shows a typical trajectory the ant might take. It is highly complicated because the beach is full of pebbles, rocks, puddles, and other obstacles. But this complexity is, in fact, only an apparent one. It would be a mistake to conclude from the – apparent – complexity of the trajectory that the internal mechanisms which are responsible for generating the behavior of the ant are also complex. This is the complexity issue. The mechanisms which are driving the ant's behavior may be very simple, implementing "rules" that we could describe as follows: "if obstacle sensor on left is activated, turn right (and vice versa)". These "rules" are, of course, implemented in the neural structures of the ant. It is important to realize that the rules alone are not sufficient to explain the ant's behavior. This is the behavior-vs.-mechanism issue.

The point is that the complexity of the ant's trajectory emerges from the *interaction* of the ant with its environment, not from the internal mechanisms or the environment alone: it would be just as erroneous to claim that the complexity of the trajectory is due to the complexity of the environment. The complexity of the environment is only a *prerequisite*. If we would increase the size of the ant, say, by a factor of 100, and let it start in the same location with exactly the same behavioral rules as before, it would go more or less in a straight line. What appeared to the normal ant as obstacles would no longer be obstacles for the giant ant. Its sensors are not sufficiently fine-grained to even detect the irregularities of the beach. The opposite is, of course, also possible: a behavior that looks very simple may be the result of complex processing. An example is moving a hand in a straight line from A to B. By using the synthetic methodology, it becomes immediately clear that this cannot be achieved by a simple mechanism. In the case of complex behavior, the mechanisms could also be complex, but they might be – and often are – simple.

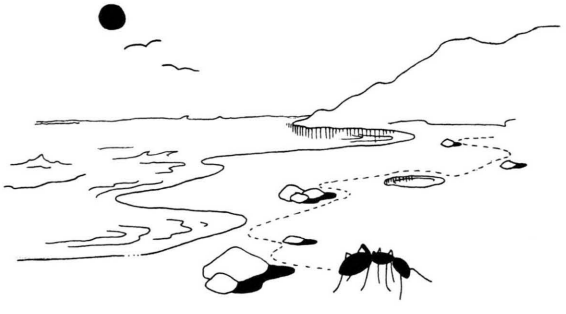

Fig. 1. Simon's ant on the beach. The ant walking on the beach was suggested by Herbert A. Simon as an illustration of the fact that behavior that looks complex to an outside observer may in fact come about by very simple mechanisms. The figure also illustrates the fact that behavior cannot be reduced to internal mechanism.

2.2 The complete agent perspective

The complete agent perspective states that we have to study agents that are self-sufficient, autonomous, embodied, and situated. Self-sufficient means that the agent is capable of sustaining itself over extended periods of time. For the purposes of this paper, this aspect is neglected. Autonomous means that the agent behaves in the environment independent of external control, in particular independent of human control. Note that this perspective differs fundamentally from standard simulation models, where the mediation between the simulation and the real world is always done by a human experimenter[2]. Embodied means that the agent has to be realized as a physical entity (which can also be simulated in the computer). Finally, the agent has to be situated meaning that it acquires all the information about the environment through its own sensory system. This enables the agent to obtain its own history which, in turn, increases its level of autonomy. The complete

[2] To be precise, a distinction between different types of simulation models has to be made, namely standard models and agent models. In agent models, the agent-environment interaction is simulated: the agents are equipped with sensors and a motor system, as well as a body, and there is an environment that has its own dynamics, i.e. a dynamics that is at least partially independent of the one of the agent.

agent perspective is closely related to the frame-of-reference problem. For example, when talking about situatedess, we are automatically trying to see the world from the perspective of the agent, rather than our own: what does the world look like to the agent? Let us now review some of the implications of complete agents with a focus on embodiment and representation.

3. Implications of the Complete Agent Perspective

There are two main implications of embodiment, one dynamic (in the physics sense), the other information theoretic.

Robots, animals, and humans are physical systems interacting with their environments. Nature exploits embodiment in beautiful and astonishing ways. An example is insect walking. Holk Cruse and his collaborators have found that in insects there is no central controller to coordinate all the six legs in walking (e.g., Cruse *et al.,* 1996). How is it possible that coordinated walking can come about without central control? Legs do communicate globally, but through the interaction with the environment, rather than through neural connections only. If an insect lifts one leg, the forces on all the others are instantaneously changed. This is due to the weight and stiffness of the insect and the surface on which it stands. There is in fact no need for a neural connection, or stated differently, the global communication between the legs is not *"represented"* within the agent (only some have connections), but all "communicate" with each other through the real world. Moreover, there is no internal global coordinating mechanism.

Considerations about dynamics are not only relevant for insects, but also for humans. To illustrate this point, let us look at a classical psychological experiment in which infants have to learn the location of various objects in space (Piaget, 1952). In this experiment infants acquire a bias to reach towards a particular location (this is known as the "AnotB error"). If a weight is attached to the infant's arm, this bias disappears (Diedrich *et al.,* 1997). It also disappears if the infants are in an upright position rather than sitting (Smith *et al.,* 1997). This experiment demonstrates that there does not seem to be an internal representation of a *decision process* or a *decision variable*, but that the error results from the reaching dynamics. This is further supported by the fact that the bias is stronger, the more the infants have reached for a particular location (Smith *et al.,* 1997). These results demonstrate the importance of embodiment when learning about objects in space. An example from engineering is Tad McGeer's "dynamic passive walker" (McGeer, 1990a, b). By exploiting the dynamics, walking behavior can be achieved that looks amazingly human-like but requires no internal representation or other kinds of internal processing. A version of such a walker with knees was built at Cornell University by Garcia *et al.* (in press) (Fig. 2). Note that walking in this machine is not internally represented in the agent: walking is the result of the physics of the interaction with the environment. If we only observe the behavior of the passive dynamic walker, we might be tempted to postulate an internal mechanism or an internal representation of walking, even though there clearly is none.

Let us now turn to the second implication of embodiment, information theoretic aspects. The kinds of sensors, their shape, and how they are positioned on the robot have important effects on the design of the control architecture. Consider the following experiment with simple robots, the Didabots. They are put into an arena with randomly distributed styrofoam cubes. After a while (about 20 min) the styrofoam cubes have ended up either along the wall or in a small number of clusters. The details are elaborated in focus box 2 (Didabots). First, a short note on the frame-of-

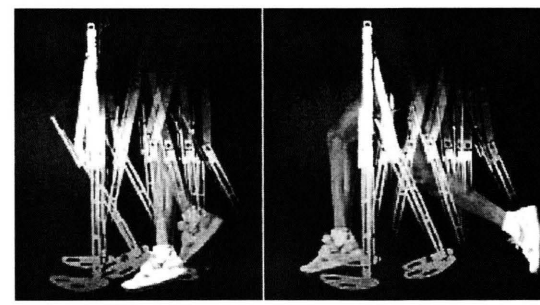

Fig. 2. The passive dynamic walker built at Cornell University, Department of Theoretical and Applied Mechanics. It walks down a shallow slope, driven solely by gravity. There is no electronic control whatsoever on the robot: the robot has no "brain". Still, the movements of the robot look very natural and human-like. This is a beautiful example of exploitation of dynamics. (from Garcia *et al.,* in press).

reference problem. The Didabots move the cubes into cluster, but there is no internal mechanism for clustering. There is also no internal representation of cubes, pushing, or clusters. All that is there internally is a simple artificial neural network, that implements the following rule: if there is sensory stimulation on the left, turn to the right (and vice versa). Why does it work? Again, the details are given in focus box 2. In essence, it depends on the exact geometric properties of the setup-up. The sensors (IR sensors in this case) have to have the right physical properties, namely a limited angle, through which they can measure reflected IR, and they have to be placed exactly in the right position. If their position is changed the behavior is changed entirely. For example, if the sensors are moved to the front of the robot, they will no longer clean up, but avoid the cubes most of the time. Thus, we see that we cannot interpret the neural signals, unless we are familiar with the physical properties of the robot, in particular its shape and the positioning of the sensors. Stated differently, the signals the robot's neural network gets strongly depend on its *morphology.*

Box 2: Didabots

Didabots (*Dida*ctical R*obots*) are simple robots equipped with infrared (IR) sensors, as shown in Fig. 3a. They are controlled by a very simple neural network that implements the following rule: if there is sensory stimulation on the left turn right (and vice versa), a rule intended for obstacle avoidance. If put into an arena with styrofoam cubes, they move the cubes into clusters, and some cubes end up along the wall (Fig. 3b). The reason is given in Fig. 3c. Normally the robots simply avoid obstacles. If they happen to encounter a cube head on, they push the cube. However, they are not searching for cubes and then pushing them: because of the particular geometry and the arrangement of the sensors, they push the cubes if they happen to encounter them in the appropriate direction. How far do they push it? Until there is another cube on the side that will provide sufficient sensory stimulation that the robot will turn away. But now there are already two cubes together and the probability that an additional cube will be deposited near them is higher. It is higher because the environment has changed, not because something has changed inside the robot: the Didabots are purely reactive. If now the position of the IR sensors is changed as shown in Fig. 3d, the Didabots will no longer move the cubes into clusters. For a more complete discussion of these experiments, see Maris and te Boekhorst (1996).

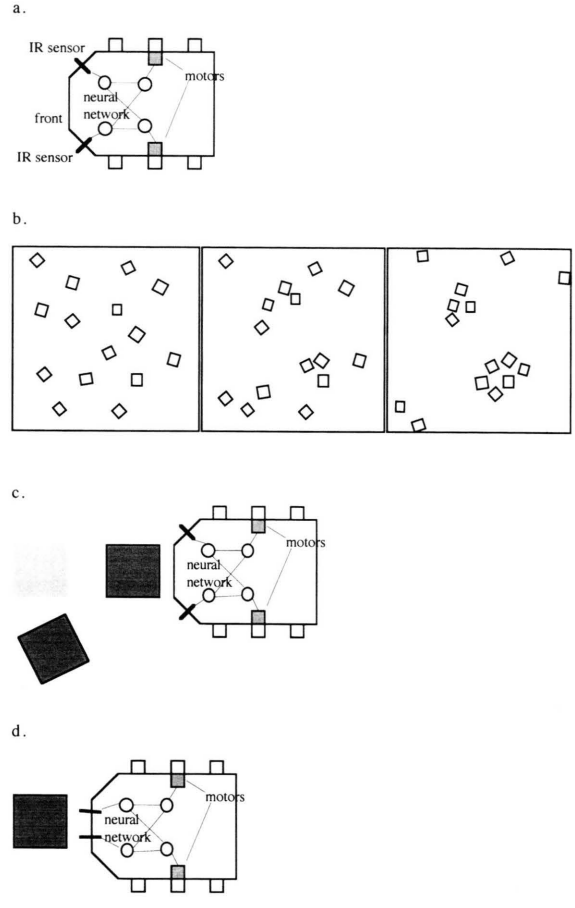

Fig. 3. Cluster formation with robots. This figure illustrates how simple robots, the Didabots, form clusters. (a) The robot with the two IR sensors and the simple neural network that implements the avoidance reflex. (b) The clustering process (overall duration ca. 20m). (c) Explanation of cluster formation. (d) Didabot with changed positions of the sensors: it's behavior will be entirely different even though the neural network is identical.

Let us now look at an example from nature, housefly navigation, in particular the phenomenon of motion parallax. The details are given in focus box 3. One of the main implications for represen-

tation is as follows. The elementary motion detectors (EMDs) can be said to represent motion to an outside observer. Although the uniform nature of the EMDs suggests linear motion, the linear motion of a point through the visual field follows a sine law. The essential conclusions are that we can only understand representation if we know the morphology and the embedding of the neural network in the complete agent.

Box 3: Motion parallax

What is shown in Fig. 4 is called the principle of motion parallax. In our discussion, we largely follow Franceschini *et al.* (1992). Motion parallax is the general term that refers to the fact that as the observer moves there are systematic movements in the visual field. The specific case we discuss here concerns the visual system of the fly. It is illustrated in Fig. 4. In the eye of the fly there is a non-uniform layout of the visual axes such that sampling of the visual space is finer towards the front than laterally. Figure 4a shows a point traveling across densely spaced vision segments in the retina, 4b a point traveling across more widely spaced vision segments in the retina. Given the same speed, a point will move slowly across the retina if it is near the front of the agent, it will move fast, if it is on the side. Because the vision segments are spaced more densely in the front than on the side a point at a given distance from the agent will require the same amount of time to traverse one vision segment. The speed at which a point traverses the visual field follows in fact a sine law: it is small at small angles (near the front), has maximum value at 90 degrees and then decreases again.

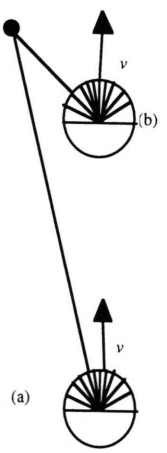

Fig. 4. Exploiting motion parallax. (a) point traveling across densely spaced vision segments in retina, (b) point traveling across widely spaced vision segments in retina. Given the same speed, a point will move slowly across the retina if it is near the front of the agent, it will move fast, if it is on the side. Because the vision segments are spaced more densely in the front than on the side a point at a given distance from the agent will require the same amount of time to traverse one vision segment.

This unequal spacing of the vision segments, the gradient, compensates for the sine law inherent in the optic flow field. The introduction of the sine gradient allows the underlying motion detection system to be built uniformly by elements each displaying the same temporal properties as its neighbors. There is no need for neural circuitry to compensate for the sine law. This illustrates a number of points. First, if we want to understand the behavior of the agent it is not sufficient to look at the control architecture. By looking at the neural architecture one could not say anything about the phonemon involved: we must take the embedding of the neural architecture into the morphology – the uneven spacing of the facets – into account. The homogeneous arrangements of EMDs in fact reflects a non-linear phenomenon which is "linearized" by the morphology. As the fly is moving at constant speed, the speed of a point moving across the visual field changes according a sine law. The arrangement transforms this mapping into one of "constant speed of fly" → "constant speed of motion across visual field". Second, the non-uniform physical arrangement of the visual segments, the facets, makes additional circuitry for compensating the continuously changing angular speed unnecessary. This means that shape or morphology is traded for computation. Morphology exploits physical processes which are very rapid and delivers "good" sensory signals, "good" in the sense that they can be processed by a relatively simple neural architecture. Because the system has been designed through evolution, this arrangement of morphology (the uneven spacing of the facettes) and the neural substrate (the array of EMDs) leads to the appropriate obstacle avoidance behavior. The EMDs – to an outside observer – do not *represent* (in the covariance perspective) linear motion, but motion governed by a sine law.

4. Exploiting Embodiment to Solve Some of the Hard Problems of Representation

Agents in the real world have many hard problems to solve. We will focus on the following two that we introduced earlier on, the focus of attention problem and the problem of object constancy (or scaling problem). We illustrate these problems with a category learning task that will occupy us for the remainder of the paper. Category learning is one of the most fundamental abilities of an adaptive agent: if the agent is not capable of making distinctions in the real world, e.g. distinguishing food from non-food, predators from prey, the nest from the rest of the world, and con-specifics from others, it will not survive for very long. Similarly, robots that cannot make distinctions are not likely to be able to do very useful work. Moreover, categorization underlies much of learning, transfer of learning, induction, generalization and abstraction. Categorization enables an agent to transfer what it has learned previously to a new situation. Upon entering a room, people instantly recognize chairs, computers, flowers, tables, and just about everything else they perceive. Organisms are often born with some a priori knowledge about the categories relevant to their ecological niche. Many animals, for example, have innate tastes for certain foods. Or certain rodents are born with knowledge of how to categorize shadows of predatory birds flying above them. Although such categorical knowledge is largely innate, it is always shaped to a certain degree by experience, i.e. by interactions of the organisms with its environment. Often, animals learn to use additional sensory modalities to make categorization more efficient, as when searching for food. In higher organisms, most categories are learned (Barsalou, 1992).

The behavior of interest is thus how agents are able to acquire and use categorical knowledge. It would be beyond the scope of this paper to review the very large literature on categorization and category learning in natural agents (see e.g. Barsalou (1992) for review). In what follows, we will focus on research conducted in two disciplines which have particularly contributed to the current understanding of categorization in humans: cognitive psychology and developmental psychology. Cognitive psychology has approached categorization from an information processing perspective (Fig. 5).

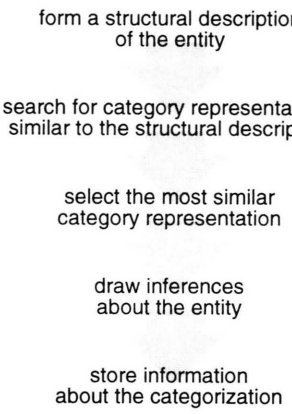

Fig. 5. The information processing approach to categorization. According to this view, categorization proceeds through the following steps: First, a structural description of the entity to be categorized is formed. Next, category representations with similar structural descriptions are searched, and the most similar category representation is selected. Based on the selected representation, inferences about the entity are drawn, and finally information about the categorization is stored in memory (adapted from Barsalou, 1992, p. 48).

According to this view, categorization in humans and higher animals involves the following steps (Barsalou, 1992). First, a structural description of the object to be categorized is formed. This description provides information about the object's primitive perceptual features such as horizontal and vertical lines as well as the relations between these basic features. Next, category representations with similar structural descriptions are searched in memory, and the most similar category representation is selected. Based on the selected representation, inferences about the object are drawn, and finally information about the categorization is stored in memory. If the object is a chair, for example, the categories chair, sofa, stool, or table might be considered, given their structural similarity. The selection process would then choose the category chair because it is the most similar. Inferences about the chosen category chair would then be drawn, for instance that it can be sat upon.

Much of the research on categorization in cognitive psychology has focused on the search and selection stages[3]. In particular, most recent work has

[3] A notable exception is a framework suggested by Lakoff (1987) according to which categories are *embodied*, that is grounded in perception, body movement, and experience of a physical and social character.

investigated the relative merits of two types of models[4]: (1) in *exemplar* models, the learner stores mental representations of exemplars, grouped by category, then categorizes new instances on the basis of their similarity to the stored exemplars. That is, according to the exemplar view people do not form abstract category knowledge but rather store collections of exemplars; (2) in *prototype* models, the learner stores a single, centralized, category representation, i.e. a prototype, rather than ensembles of exemplars.

4.1 Category learning in connnectionist models

In what follows, we will focus on a connectionist model that incorporates most of the currently accepted concepts of categorization. Later on we will contrast this model with the complete agent perspective. Most connectionist models of categorization consist of an input layer that codes object features and an output layer that represents the categories (e.g. Gluck and Bower, 1988). Typically, the goal is to learn – via supervised learning schemes such as the delta rule – an association or mapping between activations in the input layer and the corresponding activations in the category layer. One of the currently most popular models of categorization in psychology is ALCOVE (see focus box).

Box 4: ALCOVE

ALCOVE (e.g. Kruschke (1992); ALCOVE = attentional learning covering map) is a connectionist model of category learning. ALCOVE is a feedforward network with three layers of nodes. There is an input layer, a hidden layer, and an output layer. The input nodes encode the stimulus, one node per feature dimension or feature. Each input node of ALCOVE is connected to the hidden layer via so-called attention weights. These weights grow to reflect the relevance of a dimension for the categorical distinction being learned. A large weight between an input and a hidden node indicates that the network pays more attention to the feature encoded by that input node. The hidden layer consists of nodes that represent training exemplars (i.e. they light up as the feature vector representing the exemplar is presented at the input layer). The activation of a hidden node reflects the similarity of the stimulus to the exemplar represented by that node. These "exemplar" nodes compute their activation in two steps: First, they compute the distance between the stimulus and the exemplar they represent, then they compute their activation as a monotonically decreasing function of distance. The closer the input to the stored exemplar, the larger the activation of the exemplar node. Finally, the output layer consists of one node per category, with each node's activation computed as a sum of weighted activations from the exemplar nodes. The activations in the output nodes are interpreted as indicating the probability with which the network chooses a particular category in response to a particular input. Large activations lead to a high probability, and vice versa.

Learning is supervised and involves a learning and a test phase. In the learning phase, input vectors (typically binary feature vectors) are presented to the network. The network processes this input and activates one or several category nodes. The difference between the network output and the correct output (the "categorization error") is then propagated back to the hidden layer where the weights are adjusted in order to minimize the error. Note that this is, in essence, the standard backpropagation algorithm (for a discussion of the

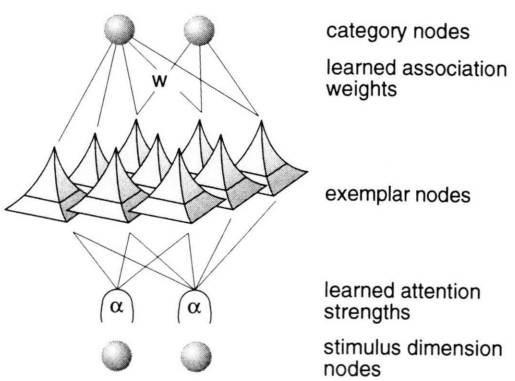

Fig. 6. Overview of the ALCOVE network. It is a three layer feedforward network. There is an input, a hidden and an output layer. The weights between the input and the hidden layer are called attentional strengths. The hidden neurons represent the exemplars the network has learned. The output nodes represent the categories.

[4] We do not discuss the more classical approaches to categorization that view categorization as being based on predicate logic. Although these are of historical interest, they are not considered to be valid models anymore.

similarities and differences between ALCOVE and standard backpropagation, see Kruschke, 1992). In the test phase, the output or category nodes are activated by summing across all the hidden nodes, weighted by the association weights between the hidden and the output layer. The output of the network is then compared to data obtained from human subjects.

In summary, the ALCOVE instantiates the view that category learning is supervised, and that categorization consists in finding a mapping from stimuli onto category representations.

The ALCOVE model incorporates a number of fundamental assumptions with respect to categorization which are representative for many cognitive models of categorization, in particular the exemplar view discussed above. The main points to be noted with respect to representation are that (a) the model is formulated on a very high-level of abstraction, (b) categories are represented explicitly as category nodes in the output layer of the network, and (c) there is no connection to the outside world: input and output are represented as feature vectors. Moreover, the output of the model has to be interpreted by the modeler. These characteristics are in sharp contrast to the complete agents approach to categorization. We will use the ALCOVE model to work out these differences. Note that the goal is not to criticize the ALCOVE model, but rather to illustrate where the standard cognitive model of categorization differs from the complete agents approach.

The fundamental question with respect to representation is the relation between the category representation and the real world. According to cognitive psychology, categories are structures in the cognitive system which represent perceived items in the environment. For example, the category *chair* represents certain visual and tactile stimuli. Thus, in this view the mind represents the external world by structuring a set of symbols so that the symbols are in correspondence with the structure of the external world. Consequently, most of the current debates in cognitive psychology, for example, concern the structure of internal representations (e.g. exemplar vs. prototype). Moreover, categorization is separated from perception. That is, categorization uses the structural description provided by the perceptual system and places the entity into a category (Barsalou, 1992). The creation of the structural description is not part of the model, but given by the designer of the model. This view on categorization is clearly *disembodied* in that categories are seen as being separate from perception or action. This is also reflected in the ALCOVE model where the categories are nodes in a layer of a neural network that can only be interpreted by attaching a symbolic label to them. There is no connection to the real-world, the input stimuli are highly abstract designer-defined descriptions of entities in the real world. The main task of the network is to activate a category node, given some input feature vector.

In what follows, we will challenge this view on categorization using results from embodied cognitive science.

4.2 Category learning based on static sensory data

The cognitivistic paradigm focuses on internal structures onto which external stimuli have to be mapped. In typical simulation models, only static stimuli are used. This is, however, not realistic when considering a complete agent interacting with the real world (as discussed earlier). Moreover, there is evidence suggesting that it is very hard to learn categories from static stimuli only. In a recent study, Nolfi (1996) has studied a robot – a Khepera™[5] robot, a popular research tool in the field of embodied cognitive science – whose task was to distinguish between walls and target objects (small cylinders). In other words, the robot faced a category learning problem: it had to learn to distinguish between walls and targets. The walls and target objects were sampled by placing the robot in front of them, and storing the activations of the six IR sensors for 180 different orientations and for 20 different distances. These data were then used to train a backpropagation network to categorize the two types of objects. Three types of network architectures were used: a two-layer network with six input neurons (one for each IR sensor), and one output neuron (coding walls by responding with a 0, and targets by responding with a 1), and two architectures where an additional layer of

---

[5] Khepera™ is a circular miniature robot with about 5cm diameter and is, depending on how it is equipped, 5 to 10 cm high. Its infrared (IR) sensors have a maximum range of about 5cm.

four and eight hidden neurons have been added, respectively.

The networks received the sampled sensory data at their input layer, and their task was to learn to respond appropriately by activating the output node for sensory data originating from targets, and by being silent when data from the walls were presented. Note that this essentially corresponds to the approach taken by most connectionist models of categorization according to which categorization involves mapping sensory patterns onto category representations (the output node). There is no motor component. Also note that the problem seems to be simple because walls are very distinct from the small target objects used in the study. Thus, it seems at first sight that in this trivial case, the mapping from inputs to category node should be learnable. The results of these experiments were as follows. Networks with no hidden units correctly categorized 22% of the patterns, while networks with hidden units, on average, were correct in 35% of the cases. Adding additional 4 hidden units did not improve performance. Thus, these networks showed a very poor categorization performance. The main reason for this is the ambiguity in the sensory data. This can be seen in Fig. 7 which depicts the sensory patterns that the networks categorized correctly, as a function of the distance and the angle of the robot relative to the objects.

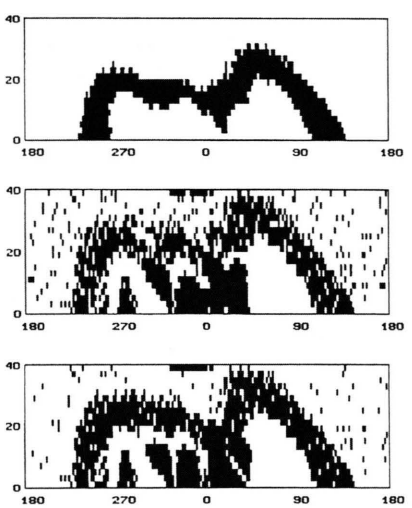

Fig. 7. Sensory patterns that the backpropagation networks categorized correctly, as a function of the angle (x-axis) and the distance (y-axis) of the robot relative to the objects. A black dot indicates correct categorization at a particular distance and angle. The top panel shows the results for the network without hidden units, the center and bottom panels depict the results for networks with a 4-neuron and a 8-neuron hidden layer, respectively (from Nolfi, 1996; Fig. 3).

Sensory patterns could only be correctly categorized in a rather narrow range of angles and distances. More specifically, objects could only be categorized when they were not more than 120° on the left or right-hand side, and no more than 32 mm away from the robot. In all other cases, the sensory data were ambiguous and the network could not categorize them appropriately. The white regions on the sides are obvious because there are no sensors on the back side of the robot. Also, the distant white areas are natural because the range of the sensors is limited. However, there are some white areas corresponding to locations very close to the robot. In other words, from most perspectives, the agent could in fact not learn the distinction. In summary, backpropagation networks similar to the ALCOVE model performed very poorly for the two categories. This is a truly surprising result: Why would such a trivial distinction not be learnable by the agent? The problem is to be seen in the general context of the object constancy problem. As discussed earlier, the problem is hard because one and the same object can lead to a very large number of different input patterns depending on the viewing angle relative to the object, the lighting conditions, noise in the sensors etc. Let us inspect this issue more closely.

In essence, we suggest, the core of the problem lies in the large input space and the ambiguities due to the object constancy problem. In a recent paper, Clark and Thornton (1997) have introduced the concept of type-2 problems to denote high-dimensional spaces in which regularities appear only "hidden" or are only "marginal". They distinguish type-2 (intractable) problems from type-1 problems where the regularities are apparent in the input data. Type-1 problems can be learned by an appropriate learning mechanism, i.e., one that is able to pick up regularities in the input space (e.g., backpropagation or even simple Hebbian learning). This is, however, not the case for type-2 problems where these regularities are not obvious and can only be recovered by means of appropriate transformations – "recodings" as Clark and Thornton call it – of the input data. In this case,

type-2 problems become type-1 problems and can be learned (for an illustration of type-1 and type-2 problems, see focus box 5). We can now reformulate the core problem of categorization as follows: the main problem in category learning of real-world objects is to turn type-2 problems into type-1 problems. The above mentioned study by Nolfi (1996) suggests that even a universal learning device such as a backpropagation network cannot learn the distinction between walls and small objects from static stimuli. In other words, these objects cannot be distinguished on the basis of the raw input stimuli alone. This indicates that additional mechanisms are needed that impose certain constraints on the sensory stimulation. Put differently, the main problem lies in the fact that real world data such as the ones collected in Nolfi's study constitute type-2 learning problems and hence first have to be transformed (or they may not be learnable in the first place, i.e. they may not even be of type 2). In the black region of Fig. 7 the data are of type 1, i.e. the invariances show up in the data which is why they can be learned with the backpropagation algorithm, whereas in the white regions, they are not. In the latter case, the data have first to be transformed into data of type 1.

Box 5: Type 1 and type 2 problems

Roughly speaking, a problem is of type 1 if the regularities can be found in the data set itself. Regularity means that the input data can be mapped onto categories (the output). In other words, in type 1 data the output can be predicted with high probability from the input by statistical analysis of the data. If the data are of type 2 they have to be subjected to a transformation before the outputs can be predicted from the inputs with high probability. This idea is illustrated in Tables I and II taken from Clark and Thornton (1997). For example, the conditional probability that the output is one, given that *x1* is 1 is 0.5, or the conditional probability that the output is 1 given that the input *x2* is 2 is 0.67. Most of the conditional probabilities are close to chance. However if the data is transformed, i.e. if we calculate *x4* as the difference between the values of *x1* and *x2* the conditional probabilities are 1. For example, the conditional probability $p(y1 = 0|x4 = 0) = 1$. If we interpret *x1*, and *x2* as sensory stimuli and *y1* as a category representation, we can say that in the first case, the mapping is not visible in the data (Table I), whereas the in the transformed case it is easily discernible (Table II). Note that the fact that a data set is not of type 1 does not automatically imply that it is of type 2 – there may be no regularities in the data set whatsoever, not even "hidden".

Table I.

x1	*x2*		*y1*
1	2	⇒	1
2	2	⇒	0
3	2	⇒	1
3	1	⇒	0
2	1	⇒	1
1	1	⇒	0

Table II.

x4		*y1*
1	⇒	1
0	⇒	0
1	⇒	1
2	⇒	0
1	⇒	1
0	⇒	0

There are two main strategies how this can be achieved. First, the internal processing of the input can be improved. Clark and Thornton (1997), for example, suggest that already learned representations can be used to reduce the complexity of learning subsequent patterns. In other words, the early knowledge is a kind of "lens" through which input data are recoded (Clark and Thornton, 1997, p. 63). The idea is that instead of learning the whole input space at once, the problem can be broken down and made more tractable by "starting small", i.e., learning some basic properties of the input, and then use the already learned knowledge to "filter" the subsequent input. In this account we have to be aware of the fact that the features acquired early on will bias the system and the processing of subsequent input (see Clark and Thornton, 1997, p. 65f for a discussion of this problem). Another possibility to increase the power of internal processing is to incorporate simple constraints such as symmetry in the network structure (e.g., Abu-Mostafa, 1992). This effectively means that the number of free parameters is reduced which leads to better generalization[6] (Abu-Mo-

[6] Stated differently, the learning machine's VC dimension is reduced. Thus, less training data are required to achieve the same level of generalization error. The term VC dimension stems from the fact that Vapnik and Chervonenkis are responsible or the pertinent theory development (which they originally published in Russian in 1974).

stafa, 1992; Vapnik and Chervonenkis, 1989; Wapnik and Tscherwonenkis, 1979). The constraints reflect a priori assumptions about the data distributions in the environment (e.g. that they are symmetric). Another example is a continuity constraint, which in this case represents an assumption about the type of objects in the environment. By exploiting such constraints, learning complexity can often be significantly reduced. For a systematic account of these issues, see, for example Meier (1997).

The second approach to transform type-2 into type-1 problems is directly derived from the basic tenets of embodied cognitive science. It consists in exploiting processes of system-environment interaction. In other words, the idea is to exploit the fact that a mobile agent can actively structure its input by manipulating the world. We call this way of interacting with the world *sensory-motor coordination* (e.g. Pfeifer and Scheier, 1997; Pfeifer and Scheier, in press)[7]. We suggest that this manipulation can serve to transform a problem that would be of type 2 (without sensory-motor coordination) into one of type-1. We will return to this issue in the discussion section. In order to give a comprehensive account we have to look at the additional case studies first.

4.3 Category learning based on sensory-motor coordination

We start our elaboration of the alternative approach to category learning and representation suggested by embodied cognitive science – sensory-motor coordination – with two examples from evolutionary robotics. The idea in evolutionary approaches to embodied cognitive science is to evolve the control architecture of an agent and reward individuals that achieve a high percentage of correct categorization by including them in the next generation of the evolutionary process. In this way, the following issue can be addressed: what strategy will prove to be the fittest, i.e., which approach will achieve the best categorization performance? Based on the arguments introduced above we predict that whatever strategy will be the fittest, it will have to include mechanisms of sensory-motor coordination. An alternative strategy that robots might evolve – derived from the information processing framework – would be the mapping of sensory patterns to an internal representation of the categories (as, for example, in the ALCOVE model).

4.3.1 Distinguishing between cylinders and walls

Let us first look at the Nolfi study. We have already referred to one part of this study above. The point there was that a backpropagation network could not learn to distinguish between target objects (small cylinders) and walls. Here we summarize the alternative approach that Nolfi has implemented to address the problem of learning to distinguish between the two categories. He used a genetic algorithm to evolve a neural controller able to perform the categorization task. Individuals were evolved in simulation, using the same sensory data as in the experiments with the backpropagation networks, i.e., real sensory data were used to drive a simulated robot. The evolved individuals were then downloaded onto the physical robot (a Khepera™ robot) in order to test their capability to operate in the real world. The process began with 100 randomly generated genotypes, each representing a network with a different set of randomly assigned connection weights from input to output layer. Each generation was allowed to operate for 5 epochs consisting of 500 actions each. At the beginning of each such epoch, the robot was randomly placed in the environment at some distance from the target object. After the fifth epoch, individuals were allowed to "reproduce" as follows. The networks of the 20 fittest individuals were copied five times, resulting in 100 (20×5) new individuals that constituted the next generation. Random mutations were introduced in this reproduction process. Overall, 100 generations were evolved. Fitness was computed by measuring the number of cycles an individual spent at a distance less than 8cm from the target object. Three network architectures were used: networks without hidden units, and with 4 or 8 hidden units, respectively. The first result was that on average networks without hidden units could solve this task better than the ones with hidden units (Nolfi (1996), Fig. 4). In other words, the network with

[7] We owe this term to John Dewey (Dewey, 1896) who pointed out the problems with the position of working from sensory stimulation, to internal processing, to output, over a century ago.

the least means of internal representation outperformed the other networks. As pointed out by Nolfi, this might relate to the fact that additional hidden neurons require longer genotypes and thus increase the search space of the genetic algorithm (which then would be an artifact of the genetic algorithm, not directly related to the problem of categorization). Another interpretation can be given, however, using the concept of sensory-motor coordination. Let us first summarize the behavior of the evolved robots. All individuals never stop once they are in front of the target. Rather, they start to move back and forth as well as slightly to the left and right, thereby keeping a fixed range of angles and distances with respect to the target. In other words, the evolutionary process has produced a mechanism of sensory-motor coordination in order to solve the categorization task.

This is a demonstration of how sensory-motor coordination cannot only be used to enable a robot to categorize objects in its environment, but how such a behavior actually evolves in robots faced with a categorization task. Using the terminology introduced above, we can describe the behavior of these agents as one that has turned the former type-2 problem into a type-1 problem. This can also be seen in Fig. 8 where the relative positions a typical agent has assumed with respect to a target (the black area in Fig. 8) are shown. We mentioned that the category learning problem can be simplified by such a reduction and this is nicely demonstrated by these experiments. In addition, however, the reduction must be such that *regularities* in this sensor space result. Although Nolfi did not analyze this aspect, we suspect that one would find clear signs of such regularities in the sensory data of the evolved agents. It is clear that there must be regularities because they can be seen in the agent's behavior. We will present such an analysis below in a different case study.

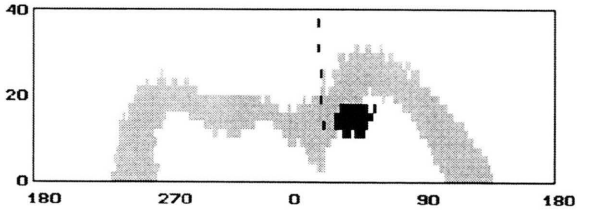

Fig. 8. The gray "wings" are the sensory patterns that the backpropagation networks categorized correctly (same data as shown in figure 7 top panel), as a function of the angle (x-axis) and the distance (y-axis) of the robot relative to the objects. The black area indicates the relative positions (angle and distance) that a typical evolved agent assumes when it reaches a target. Note that this area is very small, and also contains regions where the backpropagation networks did not categorize correctly (from Nolfi, 1996; Fig. 6).

4.3.2 Distinguishing between cylinders and diamonds

In another study by Beer (1996), very similar results were achieved. Beer also evolved agents that had to solve a categorization task. More specifically, the agents had to discriminate between circles and diamonds, catching circles while avoiding the diamonds. The study was conducted in simulation. The experimental setup is shown in Fig. 9.

The agents could move horizontally. Objects – diamonds and circles – were falling from above, starting from varying degrees of horizontal offsets. The neural network that controlled these agents

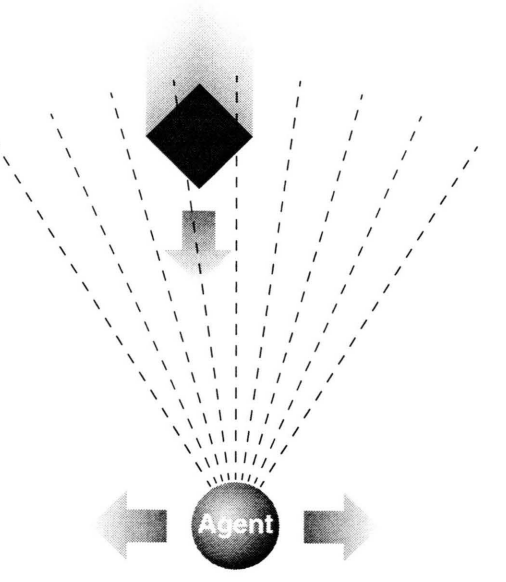

Fig. 9. The agent is equipped with a number of rays (broken lines) with which it can measure the distance from objects. Objects enter the environment from above and then move towards the agent. The agent has two motors with which it can move horizontally left and right. The task of the agent is to discriminate between circles and diamonds (a diamond is shown in the figure; adapted from Beer (1996), p. 425).

was evolved using a genetic algorithm (see Beer (1996), for details on this algorithm). The network consisted of 5 fully interconnected neurons that received input from 7 distance sensors on the agent, and that were connected to two motor neurons. Before evolving agents for this categorization task, Beer conducted another experiment where agents only had to orient to an object. It turned out that even in this simple, simulated environment, agents that could visually discriminate between objects were much more difficult to evolve than agents that could simply orient to an object (p. 425). This again indicates that learning about categories is indeed a very hard problem even when one tries to simplify the problem by using simulated environments. So, how did the agents solve the categorization task in these experiments? The results are depicted in Fig. 10.

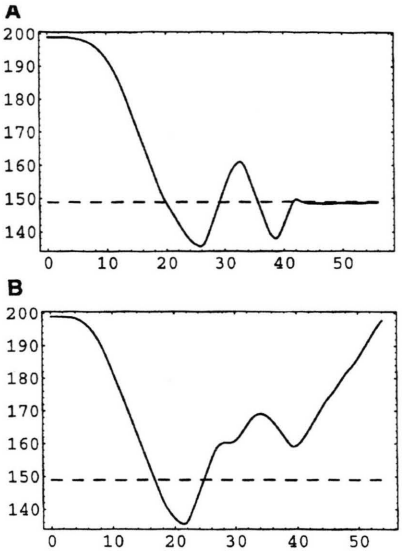

Fig. 10. Results from categorization experiments of Beer (1996). The two figures represent plots of the horizontal positions (y-axis) over time (x-axis) of an evolved agent that can categorize circles and diamonds. The path of the agent is indicated as a solid line, the one of the object as a dashed line. (a) Path of the agent catching a circle. The agent categorizes the object by keeping an invariant coupling with it. This is indicated by the overlap of the agent's and the object's path after about 43 time units. (b) Path of the agent avoiding a diamond.

The main points are as follows. The robot, when confronted with a circle or a diamond, first foveated the object (first 20 time units), then actively scanned the object in the next 20 time units, and finally either centered the object in the case of circles or avoided them in the case of diamonds. What is the reason for this "foveate-scan-decide" strategy (Beer, 1996)? Beer suggests that the foveating behavior has the advantage of placing the object in a standard position with respect to the agent. This is another example of how sensory-motor coordination (foveating, active scanning) can help to simplify categorization and reduce the amount of information that has to be stored internally. In this case, the agent reduces the sensor space by assuming a standard position with respect to the object it has to categorize. In other words, "this agent is not merely centering and then statically pattern-matching an object. Rather, its strategy seems to be a dynamic one, with active scanning apparently playing an essential role." (p. 426). Again, we suggest that sensory-motor coordination transforms the sensor space such that regularities become apparent and the objects can be learned. Based on this mechanism, all objects could be categorized correctly (Beer, 1996). Note also that as in the case of the Nolfi (1996) study summarized above, this solution has not been hand-crafted but rather emerged out of an evolutionary process.

In summary, the Nolfi and Beer experiments reveal that agents that were evolved to solve a category learning problem employed mechanisms of sensory-motor coordination to solve the task, and did not try to learn a mapping from static input patterns to internal representations. The underlying reason is that category learning in the real world is a type-2 problem and, as a consequence, regularities cannot be found without additional "tricks" that transform it to a type-1 problem. The best such "trick", we suggest, and the one that the evolutionary methods used in the experiments converged to, is sensory-motor coordination, i.e., the "trick" to interact with the objects so as to generate regularities in the input space. We suspect that very similar processes were at work in natural evolution, leading to the types of sensory-motor coordination we can observe in natural agents.

We now turn to an information theoretic analysis of the types of data generated by sensory-motor coordination.

4.3.3 Distinguishing between small and large objects

Experiment

In the following experiment, a task similar to the one used in the Nolfi experiments was used: the agent has to learn to collect some types of objects (e.g., small ones) while ignoring others (e.g., large ones) i.e., it must be able to categorize the objects in its environment. Instead of using evolutionary methods, the agent was designed based on the above considerations. The main goal was to design the agent that would be able to learn the categories based on mechanisms of sensory-motor coordination. We call this type of agent SMC (for *s*ensory-*m*otor *c*oordination) agent. The ecological niche is a flat environment with a home base to which the robot had to bring the objects it collected (Fig. 11) (for details, see Scheier and Pfeifer (1995); Pfeifer and Scheier (1997), in press).

Two types of objects were used: small and large cylinders with a height of 3 cm and 2 cm, and a diameter of 1.5 cm and 4 cm, respectively. The task of the agent is to learn to bring the small ones to a home base while avoiding the large ones. Whenever the agents picked up a small object, the sensory-motor sequence preceding this event was reinforced. Large objects were too heavy for the robot to be picked up in which case there was no reinforcement signal. In all experiments, there were 15 objects of each category randomly distributed over the whole arena.

The arm of the gripper can move through any angle from vertical to horizontal, whereas the gripper can assume only an open or a closed position. Arm and gripper positions can be sensed by position sensors coupled with the respective motors. The arm position sensor takes values from 0 (bottom back) to 255 (bottom forward), the gripper position sensor takes values between 0 (open) to 255 (closed). Objects inside the gripper can be detected by an optical barrier that is mounted on the gripper. The optical barrier takes values from 0 (no object) to 255 (object presence)(Fig. 12).

The sensory system of the SMC agent was identical to the one used by Nolfi (except that all eight sensors are used in SMC). From his analysis we know that through this sensory system, the objects cannot be learned by trying to identify a mapping from sensory stimulation onto internal representation. The SMC agent is equipped with eight IR sensors. Each IR sensor has 2^{10} (or roughly 1000) states which amounts to an input space consisting of roughly to 10^{24} different states. The categorization task is further complicated by the fact that there is a significant amount of noise in the sensor data, and that the sensor readings for the two types of objects overlap. This is illustrated in Fig. 13 where the activation of the two front sensors of the robot are shown as the robot approached a small and a large object.

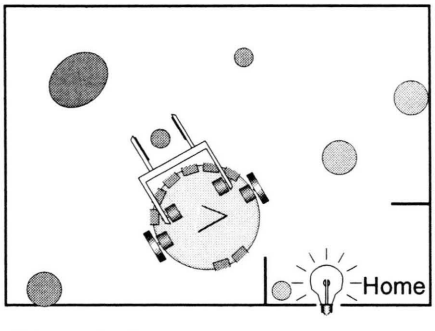

Fig. 11. The ecological niche of the SMC agent. There are small and large objects. The task of the agent is to bring the small objects to a predefined location ("Home"), and to learn to avoid exploring the large objects.

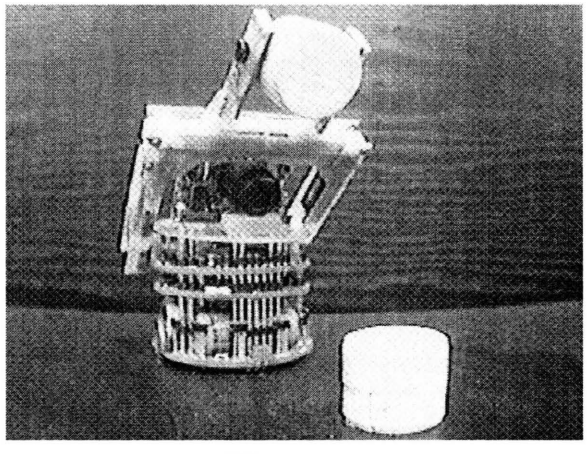

Fig. 12. The Khepera™ robot with an arm-gripper system used in the SMC experiments. The arm of the gripper can move through any angle from vertical to horizontal, whereas the gripper can assume only an open or closed position. Arm and gripper positions can be sensed with position sensors mounted in the respective motors. In addition, there is an optical barrier inside the gripper with which the presence of objects can be sensed.

The main idea of our approach is to reduce this high-dimensional space by appropriate interactions with the object. What does it mean to reduce a high-dimensional space? In essence, it means to impose some structure or constraints on the number of states the various sensors can occupy. One way to do this is to generate spatio-temporal correlations in the sensor space by interacting appropriately with the object. In other words, the strategy is to explore the object in such a way that (a) there exist correlations between different sensors and (b) the sensor readings are correlated in time. This is the approach we have chosen for the SMC agents. The sensory-motor coordination that leads to a reduction in sensor space is circling.

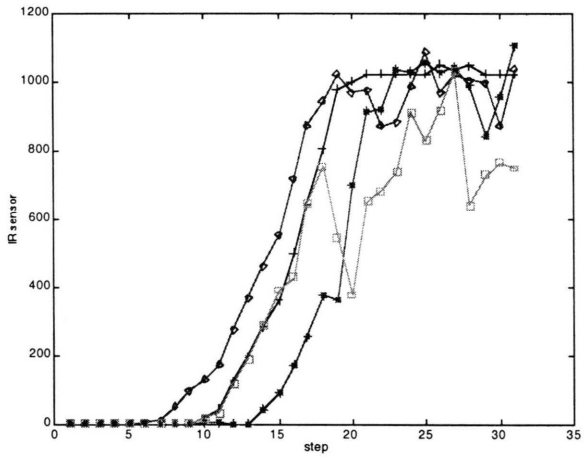

Fig. 13. IR sensor readings for small and large objects. The data were taken from the front two sensors of the robot. The robot approached a small and a large object. The data are not only noisy, but there is significant overlap between the data from the two types of objects. Asterisks and plusses denote activation of IR sensor at front left of the robot for small and large objects, respectively, and squares and diamonds denote activation of IR sensor at front right of the robot for small and large objects, respectively.

Instead of having the robot approach the object from different angles and try to learn a mapping of the resulting sensor activation and some category node, we have the robot circle around the objects. This circling behavior induces high spatio-temporal correlations in the sensor patterns (see below). It is equivalent to the object rotation behavior found in human infants (Ruff, 1984).

Before presenting results on the reduced sensor space we need to introduce the mechanisms underlying the circling behavior. This is achieved by three processes, move-forward, avoid-obstacle, and turn-towards-object. In the architecture used, the processes are all continuously and asynchronously active and influence the motor variables. The effect is as follows: Normally, the agent will move forward (move-forward process) and when encountering an obstacle, it will avoid it by turning away (avoid-obstacle process). At the same time, if there is stimulation in one of its lateral (left or right) sensors, it will turn slightly towards the object (turn-towards-object process). The interaction of these three processes leads to a behavior that we might want to call move-along-object. This is shown in Fig. 14.

This is a form of sensory-motor coordination: sensors and motor actions are coupled – they influence each other mutually. Let us look at the resulting sensor readings and motor speeds. Figure 15 shows the activation of the IR sensors as the robot circles around the objects.

Note that there is relatively little variation in these data, in particular for the large objects (indicated by asterisks in Fig. 15). For category learning, the motor speeds were used in addition to these sensor activations. The important point about these motor speeds is that they are different for the two types of objects. We can compute the angular velocity of the robot by subtracting the motor speed of

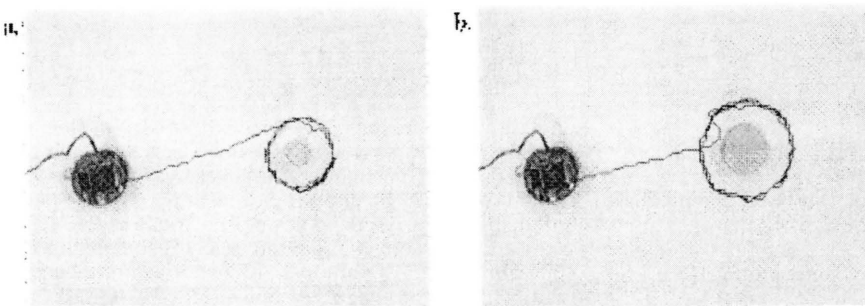

Fig. 14. Circling behavior of agent. (a) Agent circling around small object. (b) Agent circling around large object.

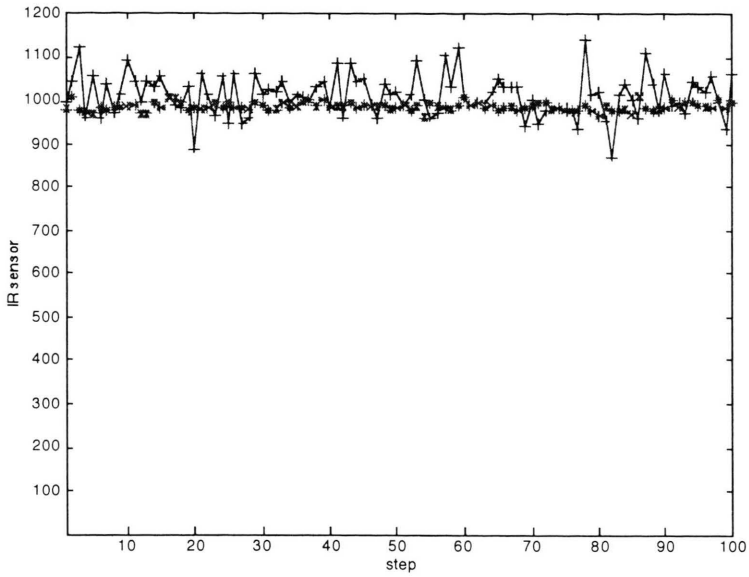

Fig. 15. Activation of lateral IR sensors as the robot circles around small (plusses) and large (asterisks) objects. The data were taken from the IR sensor on the left side of the robot, as it was circling around the object with the object on its left hand side.

the left from the motor speed from the right motor. This angular velocity is larger for small objects than for large ones. This is shown in Fig. 16.

In this example, the robot first moved in the open plane for about 40 steps. It then started to circle around a small object for about 80 steps. After it had left the object, the robot first moved in the open plane again, then avoided a large object (indicated by large fluctuations around 150 and 170 steps), and finally started again to circle around a small object at around 180 steps. The difference between the angular velocities can be exploited for category learning. In addition, however, we want the learning to be based on the sensory readings. We said at the beginning of this section that this learning is hard because of the high-dimensional state space. It turns out that the circling behavior just described, an example of sensory-motor coordination, significantly reduces the dimensionality of this space.

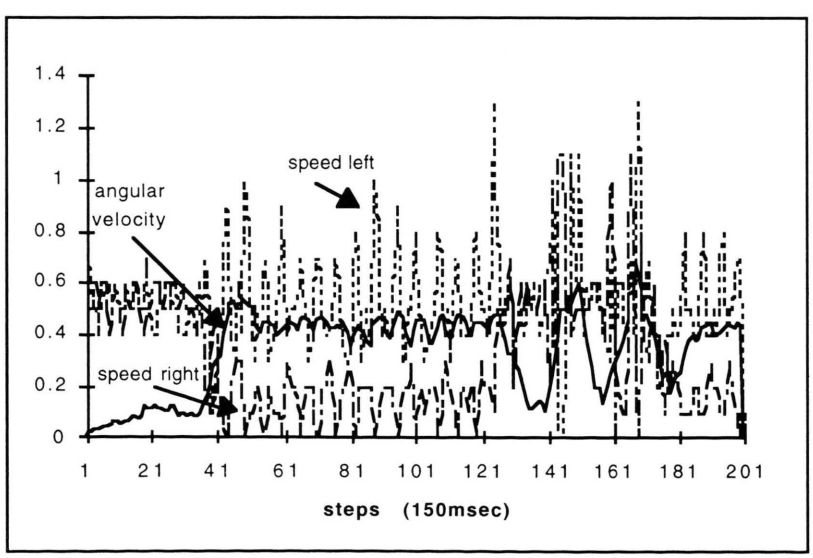

Fig. 16. Motor speeds and resulting angular velocities of the circling behavior. The robot first moved in the open plane for about 40 steps. It then started to circle around a small object for about 80 steps. After it had left the object, the robot first moved in the open plane again, then avoided a large object (indicated by large fluctuations around 150 and 170 steps), and finally started again to circle around a small object at around 180 steps.

Analysis

We now analyze the sensor space resulting from the circling behavior. The most important point to be noted is that the IR sensors vary very little due to the sensory-motor coordination behavior. In other words, the previously large state space has been significantly reduced. Let us now turn to a more detailed analysis of the sensor space of this robot.

(a) Correlations: For the following analyses we had the robot move across the open plane, approach an object and circle around it. We recorded the sensory space of the robot by means of 10-dimensional vectors consisting of the current readings from the 8 IR sensors and the two motor speeds. The process of sensory-motor coordination – circling in this case – should lead to high correlations in these data. Two kinds of correlations were calculated. The first one is the correlation between the sensor readings over time: the 10-dimensional vectors of the sensor readings are taken over a time window of 10 steps and the correlations calculated. If the pattern is stable over time, the correlation will be high, if there is a lot of change, it will be low. This is shown in Fig. 17. The second correlation is the one between the different sensor activations: this correlation will be high if the covariation between the sensors is high. In other words, the correlation can be high even if there is a lot of change in the sensor readings. Of course, if the pattern is stable over time – which is the case when the agent is circling around a cylinder – the sensor readings will also be correlated.

If we inspect Fig. 17 we find that the correlation is at an intermediate level as the agent moves about in the open (due to noise – if there were no noise, the correlation would be maximal). As the agent approaches an object, the correlation drops because now there is rapid change in sensory activation. Once the agent is near the object, the dynamics of the reflexes begins to play and we have time-locked activity in the sensory space. This can be seen in the correlations which rapidly jump to the maximum level. Note that these correlations are induced by the agent's own movements, or, in other words, by the sensory-motor coordination. And the sensory-motor coordination requires embodiment. Stated differently, through its own behavior, the agent generates *redundancy* in the sensory signals. And, as is well-known, redundancy is the prerequisite for learning (Ashby, 1956). We have shown previously, that the sensory-motor coupling leads to a reduction of the degrees of freedom in the input space. More specifically, the different objects can be learned based on one single dimension (that can be mapped onto angular velocity) (Pfeifer and Scheier, 1997).

The fact that there are correlations, i.e. that there is a stable pattern over time, also yields a focus of attention mechanism: whenever there are correlations, it is a good idea not only to heed the current situation, but also to learn. Because the stream is no longer just changing in arbitrary ways,

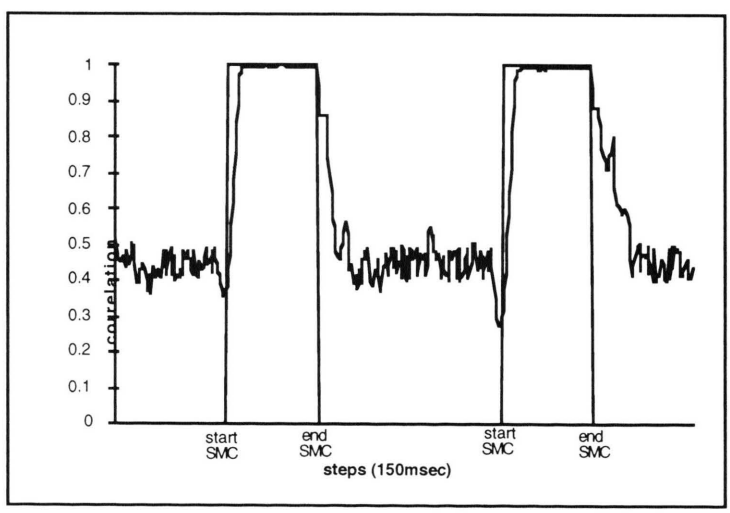

Fig. 17. Change in correlations between subsequent input vectors as the agent is approaching and then circling around a large object. Vectors are 10-dimensional: 8 IR sensors and the two motor speeds. "Start SMC" means that the agent starts circling around the object (analogously for "end SMC"). The arrow indicates the drop in correlation as the agent approaches an object.

there are very clear correlations that turn the category learning problem into a tractable (type 1) one.

As an aside, note that the correlations shown in Fig. 17, although they depend on the robot's individual characteristics, are "objective" and can be numerically calculated. Therefore they can, in principle, be picked up and exploited by the agent itself, thus providing the basis for what one might call "self-cognition". The agent can acquire a notion of its own emergent behavior.

(b) Bartlett's dimension: In order to further explore the issue of dimensionality of the sensory space we also calculated Bartlett's measure of dimensionality of a data matrix (Jackson, 1991). This measure yields the number of dimensions necessary to explain the non-random variation in the data matrix. The hypothesis is that the number of dimensions is equal to the number of the largest unequal Eigenvalues of the covariance matrix of the data matrix. A 6x30 matrix (6 IR sensors, 30 time steps, resulting in 30 successive values of each IR sensor) was used. The results are as follows:

- Approach: number of dimensions ND = 2 ($p < 0.05$)
- Circling: ND = 1 ($p < 0.05$)
- Static: ND = 4 ($p < 0.05$)

In the case of circling, the number of dimensions is 1 which is equivalent to one principal component. During approach the sensor readings change rapidly but they are still correlated which leads to a Bartlett dimension as low as 2. In the static case, the data were generated by randomly selecting a distance and a viewing angle, and then sampling the resulting sensory reading. In other words, the same procedure as applied by Nolfi in his backpropagation experiments (see above) was used. The resulting dimensionality is 4 which might explain the problems of the backpropagation algorithm to learn to classify these data.

In summary, the core mechanism of SMC is to reduce the sensor space by means of sensory-motor coordination, in this particular case circling. The results obtained generalize to agents that, unlike SMC, are equipped with several types of sensors or modalities. Sensory-motor coordination leads to redundancies across modalities, which in turn significantly simplifies cross-modal learning. This is a fundamental process. It is, for example, the basis of cognitive development in humans infants (Thelen and Smith, 1994). The main idea is illustrated in Fig. 18.

Collision sensors and proximity sensors are based on different physical processes: in the former case it is actual touch (which can be implemented, for example, as a micro-switch), in the latter measurement of reflected intensity of infrared light. Whenever there is a collision, there is high activation in the proximity sensors that are adjacent to the particular collision sensor. Thus, if there is a collision, the information from the IR sensors is largely redundant. In other words, the agent is getting similar information twice, once through every channel. This is an instance of so-called direct coupling. It only works if the sensors are positioned appropriately. This can immediately be seen by inspecting the robot in Fig. 18b. If this robot hits an obstacle, there is no overlap in the spatial information delivered by the two sensory channels and there is nothing to be learned. But the robot can behave in the real world. For example, it could simply continuously move back and forth, such that both types of sensors are in front of the object alternatively. Near an obstacle the sensory signals from the different channels will be correlated, just as for the properly designed robot in Fig. 18a. (An alternative would be to simply rotate the sensors, but that would be a somewhat different argument and we do not discuss it here). It is the more expensive solution than if the sensors are mixed, as in Fig. 18a. If the wiggling angle is preprogrammed, this is not a sensory-motor coordination, because the wiggling behavior of the robot is not influenced by the sensor signals. But if the wiggling behavior depends on the sensory

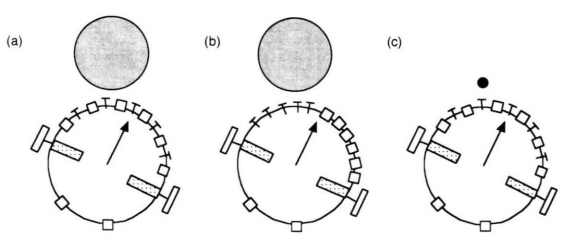

Fig. 18. Morphology, environment, and redundancy. The T-shaped sensors are collision detectors, the square ones proximity sensors (realized, e.g., by IR sensors). (a) Redundancy generated through direct coupling when hitting large object. (b) No redundancy generated through direct coupling, but may be generated through sensory-motor coordination. (c) No redundancy generated through direct coupling when hitting long thin object.

signals we have the general phenomenon of a sensory-motor coordination. It would be interesting to use artificial evolution to see whether such wiggling behavior is in fact generated when we use as a fitness function distance traveled, minus a reduction for the collisions.

If the different types of sensors are positioned appropriately in advance by the designers, correlations can be generated and learned through direct coupling – no sensory-motor coordination is required, a cheap solution. But even if the sensory channels are not a priori positioned to yield correlations, through its own behavior, through sensory-motor coordination, correlations can be achieved which can then be picked up by a learning mechanism. This way, one sensory channel becomes a predictor for the other. To put it differently, the information from one sensory channel is partly contained in the other. We see that sensory-motor coordination provides flexibility in how the potential redundancy in the sensory signals of different channels can be exploited. So, the design decision of where to put the sensors becomes highly involved: the potential overlaps in various types of sensory-motor coordinations have to be taken into account. What kinds of sensory-motor coordinations are possible or necessary depends, in turn on the environment. Imagine for a moment, that the robot in Fig. 18a operates in an environment with only narrow, vertical objects. The correlations could no longer be achieved through direct coupling because the objects are not large enough to stimulate both sensors simultaneously (Fig. 18c). The robot in Fig. 18c would have to engage in some kind of behavior (e.g. Wiggling back and forth) in order to generate the appropriate redundancy required for learning.

As of now, there is no general solution to this problem of how to optimally design sensory systems with different channels and where to position the sensors. This is because this strongly depends on the task environment. Thus, a good bet, is once again to draw inspiration from natural systems, hoping that evolution did in fact a good designer's job. The phenomenon of redundancy in the different sensory channels is a universal one. We have seen examples from robots, and developmental studies provide abundant evidence for it. A fun illustration is the so-called McGurk effect. It concerns the fact that the visual and auditory channels are both used in speech perception. McGurk and McDonald (1976) used videos of people uttering certain sounds and then they changed the sound, so that the physical sound was not compatible with the sound suggested by the video tape. For example when the video showed /pa–pa/ the physical sound was /na–na/. Their perceivers often reported hearing /ma–ma/. Another example is a /da/ on the video and a /ba/ as a physical sound. This was often reported as /va/ by perceivers. Because there is redundancy in the two channels, the perceivers noticed the incompatibility (not consciously). The result was a kind of compromise: they seemed to believe both channels to some extent.

5. Discussion and Summary

What we have tried to do in this paper is to elucidate the prerequisites that have to be met by an agent if it is to acquire representations of objects in the world. We introduced two fundamental concepts from embodied cognitive science, the complete agent perspective and the frame-of-reference problem. We have argued that often no clear distinction is made between behavior, internal mechanism and internal representation, and observer-based attributions. For example, we argued that the SMC agent had learnt distinctions between small and large cylinders. We can ask ourselves generally when it is appropriate to state that an agent has learned to categorize the environment in certain ways. In connectionist models like ALCOVE we say that this is the case if for certain feature vectors always a certain category node becomes active (and not others). In other words, in these kinds of models, the definition of category learning is based on internal representations. Whenever the correct node lights up, the correct categorization has been achieved. This kind of characterization relies on the *internal representation* of a category. In the SMC agent a different approach has been taken. We say that the SMC agent has learned to categorize its environment if it reacts differently to similar situations than it did earlier on, purely based on its behavior. For example, at the beginning of the experiment, the agent tried to pick up any object that it encountered, small or large. Over time, as it encountered a large object it would no longer pick it up. Thus, it makes

sense to say that the agent has acquired category knowledge. We can then ask the question how this knowledge is represented internally. The purpose of the paper has been to demonstrate that before we can discuss this question, we have to take embodiment into account, because real world sensory data are typically of type 2 which makes the formation of representation a demanding task.

In other words, we have been trying to work out some of the "preliminaries", some basic understanding of what is involved in, so to speak, forming an "internal world from an external one", which is one of the main purposes of representation. The main results are as follows.

(1) The interaction with the environment – through sensory-motor coordination – can provide the mechanisms of focus of attention (to deal with the continuously changing stream of sensory stimulation), and for dimensionality reduction. The various case studies have demonstrated that sensory-motor coordination leads to situations that enable the agent to learn and employ category distinctions. The implication is that perception is not a sensory problem only (i.e. a phenomenon that concerns the input only), but rather something involving the entire agent, the sensory and the motor system. Traditional models typically do not have to deal with this problem in the first place, because they are not tested with real world data. The latter are, as we have seen, normally of type 2, whereas the data used in traditional models often are of type 1. Of course, we have not provided a general solution to the hard problem of focus-of-attention. But we have shown how sensory-motor coordination can enable an agent to filter out the relevant stimulation from the continuous stream.

(2) The problem of object constancy is hard because in general the data delivered by the sensory systems are not of type 1. If behavioral regularities can be observed, it can be suspected that the data are of type 2, i.e. the regularities have to be present in the data one way or other, but the data have to be subjected to a transformation before the regularities become visible. We speculate that this transformation can – in some cases – be achieved by sensory-motor coordination (e.g, circling in the case of the SMC agent). The importance of the distinction between type 1 and type 2 data sets for learning problems can hardly be overestimated. It is theoretically fundamental and shows clearly why "perception" (as defined by mapping sensory stimulation onto internal representation) is hard. This is one of the most fundamental points to be made in this paper, i.e. the distinction between situations in which the regularities show up in the data (and can therefore be learned and represented), and situations where they don't, i.e. where an additional transformation is required before the regularities can be seen.

(3) The redundancy generated in the sensory signals through sensory-motor coordination forms the basis for category learning. The interesting aspect from a complete agent perspective is that the redundancy is not simply given but has to be generated. This provides an additional theoretical reason why it might be beneficial to engage in sensory-motor coordination. It furnishes the foundation for sensible learning processes to take place. It is another instance where we clearly see the importance of embodiment. More details, as well as quantitative measures are provided in Pfeifer and Scheier (in press; chapter 13).

(4) The sensory signals, and thus the internal representations (in whatever form they are encoded) are strongly dependent on the nature of the sensors and on where the sensors are positioned on the agent, i.e. on the morphology of the agent. This is one of the essential implications of embodiment. Systems in nature have a distribution of sensors suitable for their particular ecological niche. Recall the case study of motion parallax where there is an array of identical EMDs for motion detection. The interpretation of their activation requires an understanding of the morphology of the eye. An additional illustration is provided by the following simple thought experiment. Assume that the agent has learned to categorize small and large cylinders. Now change the position of the IR sensors by moving them, for example, all to one side of the robot and leaving the neural network exactly as it is: the agent will no longer be able to perform the categorization task.

In summary, as mentioned a number of times, we have not yet solved the problem of representation, nor have we been able to provide a general formalism. Rather, we have outlined a number of fundamental considerations to take into account in the discussion of representation. Much additional research will be required to arrive at a general "theory of embodied representation".

Acknowledgments

This research has been supported in part by grant number 11-50955.97 of the Swiss National Science Foundation. We would like to thank Gregor Schoener and Stefan Reimann for reviewing an earlier version of this paper, and Ralf Salomon and Akio Ishiguro of Nagoya University for many helpful comments and discussions.

Abu-Mostafa Y. S. (1992), Hints and the VC dimension. Neural Computation **5**, 278–288.

Ashby W. R. (1956), An Introduction to Cybernetics. Chapman and Hall, London.

Barsalou L. W. (1992), Cognitive Psychology. An Overview for Cognitive Scientists. Lawrence Erlbaum, Hillsdale.

Beer R. (1996), Toward the evolution of dynamical neural networks for minimally cognitive behavior. In: From Animals to Animats. Proc. of the 4th Int. Conf. on Simulation of Adaptive Behavior (P. Maes, M. Mataric, J.-A. Meyer, J. Pollack and S. W. Wilson, eds.). A Bradford Book, MIT Press, Cambridge, Mass., 421–429.

Brooks R. A. (1991), Intelligence without representation. Artificial Intelligence **47**, 139–160.

Bursen H. A. (1980) Dismantling the Memory Machine. A Philosophical Investigation of Machine Theories of Memory. D. Reidel Publishing, Dordrecht, NL.

Clancey W. J. (1997), Situated Cognition: On Human Knowledge and Computer Representations. Cambridge University Press, New York.

Clark A. and Thornton C. (1997), Trading spaces: computation, representation and the limits of uninformed learning. Behav. Brain Sci. **20**, 57–90

Cruse H., Bartling C., Dean J., Kindermann T., Schmitz J., Schumm M. and Wagner H. (1996), Coordination in a six-legged walking system. Simple solutions to complex problems by exploitation of physical properties. In: From Animals to Animats. Proc. of the 4th Int. Conf. on Simulation of Adaptive Behavior (P. Maes, M. Mataric, J.-A. Meyer, J. Pollack and S. W. Wilson, eds.). A Bradford Book, MIT Press, Cambridge, Mass., 84–93.

Cummins R. (1989), Meaning and Mental Representation. A Bradford Book, The MIT Press, Cambridge, Mass..

Dewey J. (1896), The reflex arc in psychology. Psychol. Review **3,** 1981, 357–370. Reprinted in J. J. McDermott (eds.): The Philosophy of John Dewey. The University of Chicago Press, Chicago, IL, 136–148.

Diedrich F. and Thelen E. (submitted). Reaching dynamics in the AnotB error. Manuscript under review.

Franceschini N., Pichon J. M. and Blanes C. (1992), From insect vision to robot vision. Phil. Trans. R. Soc. Lond. B **337**, 283–294.

Franklin S. (1995), Artificial minds. A Bradford Book, MIT Press, Cambridge, Mass .

Garcia M., Chatterjee A., Ruina A. and Coleman M. (in press), The simplest walking model: stability, complexity, and scaling. ASME Journal of Biomechanical Engineering.

Gluck M. A. and Bower G. H. (1988), From conditioning to category learning: an adaptive network model. J. Exp. Psychol.: General **117**, 227–247.

Harnad S. (1990), The Symbol Grounding Problem. Physica D **42**, 335–346.

Hendriks-Jansen H. (1996), Catching Ourselves in the Act. Situated Activity, Interactive Emergence, Evolution, and Human Tought. A Bradford Book, MIT Press, Cambridge, Mass..

Jackson E. J. (1991), A User's Guide to Principal Components. John Wiley & Sons, Inc.

Kruschke J. K. (1992), ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review **99**, 22–44.

Lakoff G. (1987), Women, Fire, and Dangerous Things. What Categories Reveal About the Mind. University of Chicago Press, Chicago.

Maris M. and te Boekhorst, R. (1996), Exploiting physical constraints: heap formation through behavioral error in a group of robots. In: Proc. IROS'96, IEEE/RSJ International Conference on Intelligent Robots and Systems.

McGurk H. and McDonald J. (1976), Hearing lips and seeing voices. Nature **264**, 746–748.

McGeer T. (1990a), Passive dynamic walking. Int. J. Robot. Res. **9**, 62–82.

McGeer T. (1990b), Passive walking with knees. Proc. of the IEEE Conference on Robotics and Automation **2**, 1640–1645.

Meier D. (1997), Generalization and constraints in learning machines. Unpublished Ph.D. Thesis, Department of Computer Science, University of Zürich.

Newell A. (1990), Unified Theories of Cognition. Harvard University Press, Cambridge, Mass.

Nolfi S. (1996), Adaptation as a more powerful tool than decomposition and integration. Technical Report 96-03, Department of Neural Systems and Artificial Life, Institute of Psychology, C. N. R., Rome, Italy.

Pfeifer R. and Scheier C. (1997), Sensory-motor coordination: the metaphor and beyond. In: Robotics and Autonomous Systems **20**, Special Issue on "Practice and Future of Autonomous Agents" (R. Pfeifer and R. Brooks, eds.), 157–178.

Pfeifer R. and Scheier C. (in press), Unterstanding Intelligence. MIT Press, Cambridge, Mass.

Piaget J. (1952), The Origins of Intelligence in Children. International University Press, New York.

Pylyshyn Z. W. (ed.) (1987), The Robot'ts Dilemma. The Frame Problem in Artificial Intelligence. Ablex, Norwood, N. J.

Ruff H. A. (1984), Infants' manipulative exploration of objects: Effects of age and object characteristics. Develop. Psychol. **20**, 9–20.

Scheier C. and Pfeifer R. (1995), Classification as sensory-motor coordination. In: Proc. European Conference on Artificial Life, ECAL-95, pp. 656–667.

Simon H. A. (1969), The Sciences of the Artificial. MIT Press, Cambridge, Mass. (2nd edition).

Smith L. B., Thelen E., Titzer R. and McLin D. (submitted), Knowing in the context of acting: the task dynamics of the AnotB error. Manuscript under review.

Thelen E. and Smith L. (1994), A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Bradford Books, Cambridge, Mass.

Vapnik V. N. and Chervonenkis A.(1989), On the uniform convergence of relative frequencies of events to their probabilities. Theory Prob. Appl. **16**, 246–280.

Wapnik W. N. and Tscherwonenkis A. J. (1979), Theorie der Zeichenerkennung. Akademie-Verlag, Berlin (original publication in Russian, 1974).