From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot

by

Paul Michael Fitzpatrick

B.Eng., University of Limerick (1995)
M.Eng., University of Limerick (1997)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2003

Author . . . .
Department of Electrical Engineering and Computer Science
May 22, 2003

Certified by . . . .
Rodney A. Brooks
Fujitsu Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot

by
Paul Michael Fitzpatrick

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2003, in partial fulfillment of the
requirements for the degree of
Doctor of Philosopy in Computer Science and Engineering

Abstract

This thesis presents a perceptual system for a humanoid robot that integrates abilities such as object localization and recognition with the deeper developmental machinery required to forge those competences out of raw physical experiences. It shows that a robotic platform can build up and maintain a system for object localization, segmentation, and recognition, starting from very little. What the robot starts with is a direct solution to achieving figure/ground separation: it simply 'pokes around' in a region of visual ambiguity and watches what happens. If the arm passes through an area, that area is recognized as free space. If the arm collides with an object, causing it to move, the robot can use that motion to segment the object from the background. Once the robot can acquire reliable segmented views of objects, it learns from them, and from then on recognizes and segments those objects without further contact. Both low-level and high-level visual features can also be learned in this way, and examples are presented for both: orientation detection and affordance recognition, respectively.

The motivation for this work is simple. Training on large corpora of annotated real-world data has proven crucial for creating robust solutions to perceptual problems such as speech recognition and face detection. But the powerful tools used during training of such systems are typically stripped away at deployment. Ideally they should remain, particularly for unstable tasks such as object detection, where the set of objects needed in a task tomorrow might be different from the set of objects needed today. The key limiting factor is access to training data, but as this thesis shows, that need not be a problem on a robotic platform that can actively probe its environment, and carry out experiments to resolve ambiguity. This work is an instance of a general approach to learning a new perceptual judgment: find special situations in which the perceptual judgment is easy and study these situations to find correlated features that can be observed more generally.

Thesis Supervisor: Rodney A. Brooks
Title: Fujitsu Professor of Computer Science and Engineering

Acknowledgments

My thanks to Rod Brooks for his support and advice throughout my time at MIT, and for running such a diverse and interesting research group – and letting me be a part of it. I am also grateful to Deb Roy and Trevor Darrell, the other members of my committee, for their helpful comments on the thesis.

The robots used in this thesis are the fruit of the combined labor of many students. I am grateful to them all, and particularly indebted to Brian Scassellati, Matthew Marjanović, Cynthia Breazeal, Matthew Williamson, and Bryan Adams.

I would like to take this opportunity to deny the existence of the Tea Conspiracy, a group which – if it existed – would no doubt have been a very pleasant break between working-all-day and working-all-night at the AI lab. Eduardo Torres-Jara strikes me as the sort who might ring-lead such a conspiracy, if it existed, which it emphatically does not (and even if it did, there is absolutely no evidence that its members are plotting a lab coup d'état). Thank you Eduardo for your indefatigable good-humor. My thanks to Giorgio Metta for being a great friend and colleague. To both Charlie and Clarke Kemp, for innumerable interesting and challenging conversations. To Artur Arsenio, Lijin Aryananda and Hideki Kozima, for making office 936 a fun place to be. And to everyone else in the Living Machines group – Martin Martin, Una-May O'Reilly, Paulina Varshavskaya, Aaron Edsinger, Jessica Banks, Jeff Weber, Jessica Howe, Annika Pfluger, and the mysterious Juan Velasquez.

My journey from herding goats in rural Ireland to doing robotics research in Cambridge was a long one. Boundless thanks to my parents, Ann and William, who made it all possible. And to my siblings, Deirdre, Liam, Mary, and John for their support. To Aunt Josephine and Granny F for writing to me so often while receiving so little in return. And to various groups in the governments in Ireland and the US for paying for my education – I might still be herding goats otherwise (not that there is anything wrong with that). I am grateful to the books of Terry Pratchett for providing endless amusement and inspiration. They are quoted at the start of each chapter of the thesis, and bear only the flimsiest of connection with the main text.

To Noémi Giszpenc for giving me the confidence that, if I could learn to pronounce her name, I could accomplish anything.

To my grandfathers,
Tom and Jerry.

Contents

List of Figures

CHAPTER 1

Introduction

> *Everything starts somewhere, although many physicists disagree. But people have always been dimly aware of the problems with the start of things. They wonder aloud how the snowplough driver gets to work, or how the makers of dictionaries look up the spellings of words.* (Pratchett, 1996)

The goal of this work is to build a perceptual system for a robot that integrates useful "mature" abilities, such as object localization and recognition, with the deeper developmental machinery required to forge those competences out of raw physical experiences. The motivation for doing so is simple. Training on large corpora of real-world data has proven crucial for creating robust solutions to perceptual problems such as speech recognition and face detection. But the powerful tools used during training of such systems are typically stripped away at deployment. For problems that are more or less stable over time, such as face detection in benign conditions, this is acceptable. But for problems where conditions or requirements can change, then the line between training and deployment cannot reasonably be drawn. The resources used during training should ideally remain available as a support structure surrounding and maintaining the current perceptual competences. There are barriers to doing this. In particular, annotated data is typically needed for training, and this is difficult to acquire online. But that is the challenge this thesis addresses. It will show that a robotic platform can build up and maintain a quite sophisticated object localization, segmentation, and recognition system, starting from very little.

1.1 The place of perception in AI

If the human brain were a car, this message would be overlaid on all our mental reflections: "caution, perceptual judgements may be subtler then they appear". Time and time again, the difficulty of implementing analogues of human perception has been underestimated by AI researchers. For example, the Summer Vision Project of 1966 at the MIT AI Lab apparently expected to implement figure/ground separation and object recognition on a limited set of objects such as balls and cylinders in the month of July, and then extend that to cigarette packs, batteries, tools and cups in August (Papert, 1966). That "blind spot" continues to the current day – for example, the proposal for the thesis you are reading blithely assumed the existence of perceptual abilities that now consume entire chapters. But there has been progress. Results in neuroscience continue to drive home the sophistication of the perceptual machinery in humans and other animals. Computer vision and speech recognition have become blossoming fields in their own right. Advances in consumer electronics have led to a growing drive towards advanced human/computer interfaces, which bring machine perception to the forefront. What does all this mean for AI, and its traditional focus on representation, search, planning, and plan execution? For devices that need to operate in rich, unconstrained environments, the emphasis on planning may have been premature:

> "I suspect that this field will exist only so long as it is considered acceptable to test these schemes without a realistic perceptual interface. Workers who have confronted perception have found that on the one hand it is a much harder problem than action selection and that on the other hand once it has been squarely faced most of the difficulties of action selection are eliminated because they arise from inadequate perceptual access in the first place." (Chapman, 1990)

It is undeniable that planning and search are crucial for applications with complex logistics, such as shipping and chess. But for robotics in particular, simply projecting from the real world onto some form where planning and search can be applied seems to be the key research problem: "This abstraction process is the essence of intelligence and the hard part of the problem being solved" (Brooks, 1991b). Early approaches to machine perception in AI focused on building and maintaining detailed, integrated models of the world that were as complete as possible given the sensor data available. This proved extremely difficult, and over time more practical approaches were developed. Here are cartoon-caricatures of some of them:

- ▷ **Stay physical:** Stay as close to the raw sensor data as possible. In simple cases, it may be possible to use the world as its own model and avoid the difficulties involved in creating and maintaining a representation of a noisily- and partially-observed world (Brooks, 1991b). Tasks such as obstacle avoidance can be achieved reactively, and Connell (1989) gives a good example of how a task with temporal structure can be performed by maintaining state in the world and the robot's body rather than within its control system. This work clearly demonstrates that the structure of a task is logically distinct from the structures required to perform it. Activity that is sensitive to some external structure in the world does not imply a control system that directly mirrors that structure in its organization.
- ▷ **Stay focused:** Adopt a point of view from which to describe the world that is sufficient for your task and which simplifies the kind of references that need to be made, hopefully to the point where they can be easily and accurately maintained. Good examples include deictic representations like those used in Pengi (Chapman and Agre, 1987), or Toto's representations of space (Mataric, 1990).
- ▷ **Stay open:** Use multiple representations, and be flexible about switching between representations as each run into trouble (Minsky, 1985). This idea overlaps with the notion of encoding common sense (Lenat, 1995), and using multiple partial theories rather than searching – perhaps vainly – for single unified representations.

While there are some real conflicts in the various approaches that have been adopted, they also have a common thread of pragmatism running through them. Some ask "what is the minimal representation possible", others "what choice of representation will allow me to develop my system most rapidly?" (Lenat, 1995). They are also all steps away from an all-singing, all-dancing monolithic representation of the external world. Perhaps they can be summarized (no doubt kicking and screaming) with the motto "robustness from perspective" – if you look at a problem the right way, it may be relatively easy. This idea was present from the very beginning of AI, with the emphasis on finding the right representations for problems, but it seemed to get lost once division of labor set in and the problems (in some cases) got redefined to match the representations.

There is another approach to robust perception that has developed, and that can perhaps be described as "robustness from experience". Drawing on tools from machine learning, just about any module operating on sensor input can be improved. At a minimum, its performance can be characterized empirically, to determine when it can be relied upon and when it fails, so that its output can be appropriately weighed against other sources. The same process can be applied at finer granularity to any parameters within the module that affect its performance in a traceable way. Taking statistical learning of this kind seriously leads to architectures that seem to contradict the above approaches, in that they derive benefit from representations that are as integrated as possible. For example, when training a speech recognition system, it is useful to be able to combine acoustic, phonological, language models so that optimization occurs over the largest scope possible (Mou and Zue, 2001).

The success of statistical, corpus-based methods suggests the following additional organizing principle to the ones already enunciated :-

- ▷ **Stay connected:** Statistical training creates an empirical connection between parameters in the system and experience in the world that leads to robustness. If we can *maintain* that connection as the environment changes, then we can maintain robustness. This will require integrating the tools typically used during training with the deployed system itself, and engineering opportunities to replace the role that annotation plays.

This thesis argues that robots must be given not just particular perceptual competences, but the tools to forge those competences out of raw physical experiences. Three important tools for extending a robot's perceptual abilities whose importance have been recognized individually are related and brought together. The first is active perception, where the robot employs motor action to reliably perceive properties of the world that it otherwise could not. The second is development, where experience is used to improve perception. The third is interpersonal influences, where the robot's percepts are guided by those of an external agent. Examples are given for object segmentation, object recognition, and orientation sensitivity; initial work on action understanding is also described.

1.2 Why use a robot?

The fact that vision can be aided by action has been noted by many researchers (Aloimonos et al., 1987; Bajcsy, 1988; Ballard, 1991; Gibson, 1977). Work in this area focuses almost uniformly on the advantages afforded by moving cameras. For example, Klarquist and Bovik (1998) use a pair of cameras mounted on a track to achieve precise stereoscopic vision. The track acts as a variable baseline, with the system *physically* interpolating between the case where the cameras are close – and therefore images from them are easy to put into correspondence – and the case where the cameras are separated by a large baseline – where the images are different enough for correspondences to be hard to make. Tracking correspondences from the first to the second case allows accurate depth estimates to be made on a wider baseline than could otherwise be supported.

In this thesis, the work described in Chapter 3 extends the basic idea of action-aided vision to include simple manipulation, rather than just moving cameras. Just as conventional active vision provides alternate approaches to classic problems such as stereo vision and object tracking, the approach developed here addresses the classic problem of object segmentation, giving the visual system the power to recruit arm movements to probe physical connectivity. This thesis is a

Figure 1-1: Training data is worth its weight in ice cream in the speech recognition research community (certificate created by Kate Saenko).

step towards visual monitoring of robot action, and specifically manipulation, for the purposes of correction. If the robot makes a clumsy grasp due to an object being incorrectly segmented by its visual system, and ends up just brushing against an object, then this thesis shows how to exploit that motion to correctly segment the object – which is exactly what the robot needs to get the grasp right the next time around. If an object is awkwardly shaped and tends to slip away if grasped in a certain manner, then the affordance recognition approach is what is needed to learn about this and combat it. The ability to learn from clumsy motion will be an important tool in any real, general-purpose manipulation system.

Certain elements of this thesis could be abstracted from the robotic implementation and used in a passive system, such as the object recognition module described in Chapter 5. A protocol could be developed to allow a human teacher to present an object to the system and have it enrolled for object recognition without requiring physical action on the robot's part. For example the work of Nayar et al. (1996) detects when the scene before a camera changes, triggering segmentation and object enrollment. However, it relies on a very constrained environment – a dark background with no clutter, and no extraneous environmental motion. Another approach that uses human-generated motion for segmentation – waving, pointing, etc. – is described in Arsenio et al. (2003). The SAIL robot (Weng et al., 2000a) can be presented with an object by placing the object in its gripper, which it then rotates 360° in depth, recording views as it goes. But all these protocols that do not admit of autonomous exploration necessarily limit the types of applications to which a robot can be applied. This thesis serves as a proof of concept that this limitation is not essential. Other researchers working on autonomous development are motivated by appeals to biology and software complexity (Weng et al., 2000b). The main argument added here is that autonomy is simply unavoidable if we wish to achieve maximum robustness. In the absence of perfect visual algorithms, it is crucial to be able to adapt to local conditions. This is particularly clear in the case of object recognition. If a robot moves from one locale to another, it will meet objects that it has never seen before. If it can autonomously adapt to these, then it will have a greater range of applicability. For example, imaging a robot asked to "clear out the junk in this basement." The degree of resourcefulness required to deal with awkwardly shaped and situated objects make this a very challenging task, and experimental manipulation would be a very helpful technology for it.

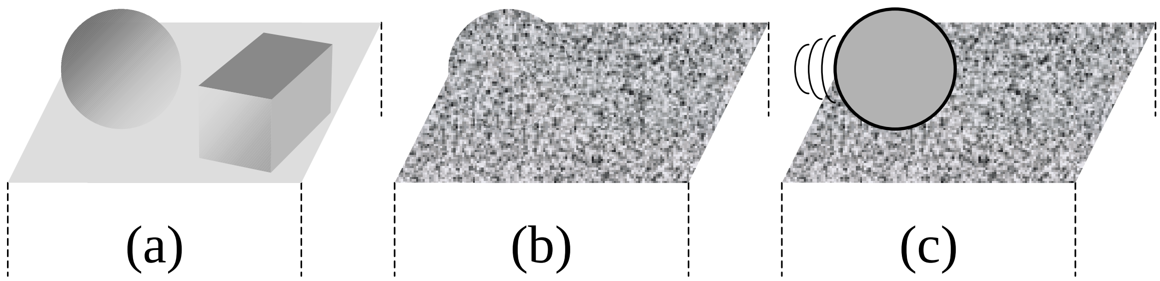

Figure 1-2: Cartoon motivation for active segmentation. Human vision is excellent at figure/ground separation (top left), but machine vision is not (center). Coherent motion is a powerful cue (right) and the robot can invoke it by simply reaching out and poking around.

1.3 Replacing annotation

Suppose there is some property P of the environment whose value the robot cannot usually determine. Further suppose that in some very special situations, the robot *can* reliably determine the property. Then there is the potential for the robot to collect training data from such special situations, and learn other more robust ways to determine the property P. This process will be referred to as "developmental perception" in this thesis.

Active and interpersonal perception are identified as good sources of these "special situations" that allow the robot to temporarily reach beyond its current perceptual abilities, giving the opportunity for development to occur. Active perception refers to the use of motor action to simplify perception (Ballard, 1991), and has proven its worth many times in the history of robotics. It allows the robot to experience percepts that it (initially) could not without the motor action. Interpersonal perception refers to mechanisms whereby the robot's perceptual abilities can be influenced by those around it, such as a human helper. For example, it may be necessary to correct category boundaries or communicate the structure of a complex activity.

By placing all of perception within a developmental framework, perceptual competence becomes the result of experience evoked by a set of behaviors and predispositions. If the machinery of development is sufficient to reliably lead to the perceptual competence in the first place, then it is likely to be able to regenerate it in somewhat changed circumstances, thus avoiding brittleness.

1.4 Active perception

The idea of using action to aid perception is the basis of the field of "active perception" in robotics and computer vision Ballard (1991); Sandini et al. (1993). The most well-known instance of active perception is active vision. The term "active vision" has become essentially synonymous with moving cameras, but it need not be. There is much to be gained by taking advantage of the fact that robots are actors in their environment, not simply passive observers. They have the opportunity to examine the world using causality, by performing probing actions and learning from the response. In conjunction with a developmental framework, this could allow the robot's experience to expand outward from its sensors into its environment, from its own arm to the objects it encounters, and from those objects both back to the robot itself and outwards to other actors that encounter those same objects.

Active vision work on the humanoid robot Cog is oriented towards opening up the potentially

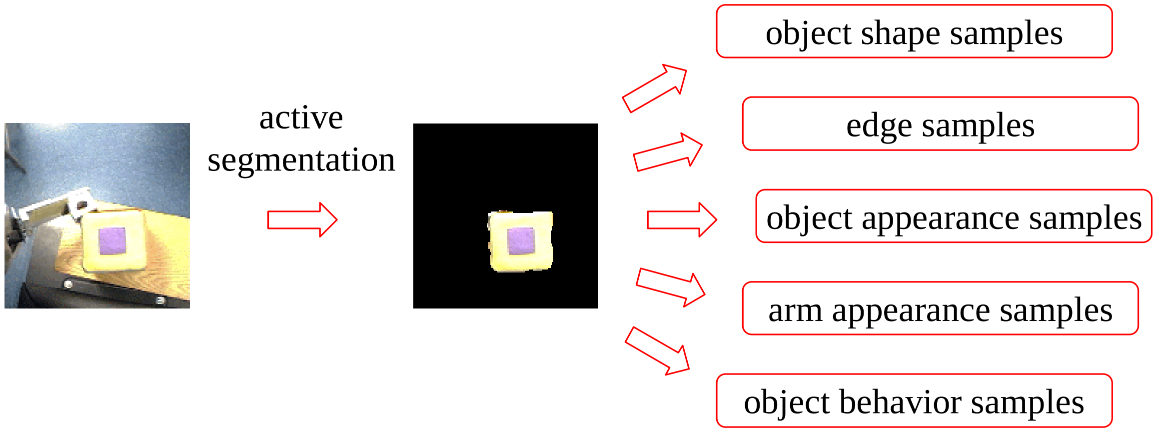

Figure 1-3: The benefits of active segmentation using poking. The robot can accumulate training data on the shape and appearance of objects. It can also locate the arm as it strikes objects, and record its appearance. At a lower level, the robot can sample edge fragments along the segmented boundaries and annotate them with their orientation, facilitating an empirical approach to orientation detection. Finally, tracking the motion of the object after poking is straightforward since there is a segmentation to initialize the tracker – hence the robot can record the motion that poking causes in different objects.

rich area of manipulation-aided vision, which is still largely unexplored. Object segmentation is an important first step. Chapter 3 develops the idea of *active segmentation*, where a robot is given a “poking” behavior that prompts it to select locations in its environment, and sweep through them with its arm. If an object is within the area swept, then the motion generated by the impact of the arm can be used to segment that object from its background, and obtaining a reasonable estimate of its boundary (see Figure 1-3). The image processing involved relies only on the ability to fixate the robot’s gaze in the direction of its arm. This coordination can be achieved either as a hard-wired primitive or through learning. Within this context, it is possible to collect good views of the objects the robot pokes, and the robot’s own arm. Giving the robot this behavior has several benefits. *(i)* The motion generated by the impact of the arm with an object greatly simplifies segmenting that object from its background, and obtaining a reasonable estimate of its boundary. This will prove to be key to automatically acquiring training data of sufficient quality to support the forms of learning described in the remainder of this thesis. *(ii)* The poking activity also leads to object-specific consequences, since different objects respond to poking in different ways. For example, a toy car will tend to roll forward, while a bottle will roll along its side. *(iii)* The basic operation involved, striking objects, can be performed by either the robot or its human companion, creating a controlled point of comparison between robot and human action.

Figure/ground separation is a long-standing problem in computer vision, due to the fundamental ambiguities involved in interpreting the 2D projection of a 3D world. No matter how good a passive system is at segmentation, there will be times when only an active approach will work, since visual appearance can be arbitrarily deceptive. Of course, there will be plenty of limitations on active segmentation as well. Segmentation through poking will not work on objects the robot cannot move, either because they are too small or too large. This is a constraint, but it means we are well matched to the space of manipulable objects, which is an important class for robotics.

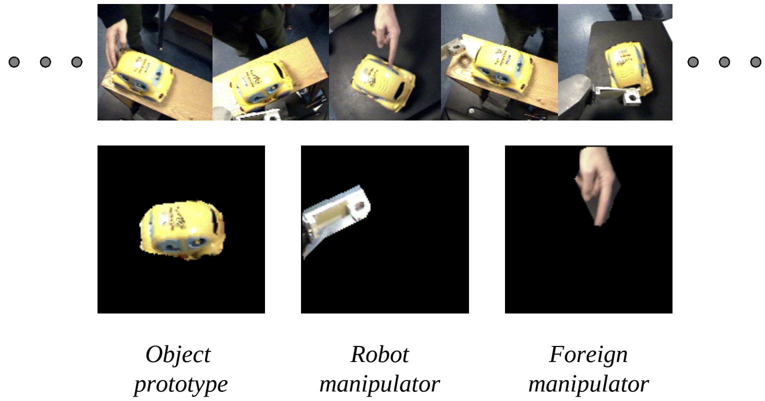

Figure 1-4: The top row shows sample views of a toy car that the robot sees during poking. Many such views are collected and segmented. The views are aligned to give an average prototype for the car (and the robot arm and human hand that acts upon it). To give a sense of the quality of the data, the bottom row shows the segmented views that are the best match with these prototypes. The car, the robot arm, and the hand belong to fundamentally different categories. The robot arm and human hand cause movement (are actors), the car suffers movement (is an object), and the arm is under the robot's control (is part of the self).

1.5 Developmental perception

Active segmentation provides a special situation in which the robot can observe the boundary of an object. Outside of this situation, locating the object boundary is basically guesswork. This is precisely the kind of situation that a developmental framework could exploit. The simplest use of this information is to empirically characterize the appearance of boundaries and oriented visual features in general. Once an object boundary is known, the appearance of the edge between the object and the background can be sampled along it, and labelled with the orientation of the boundary in their neighborhood. This is the subject of Chapter 4. At a higher-level, the segmented views provided by poking objects can be collected and clustered as shown in Figure 1-4. Such views are just what is needed to train an object detection and recognition system, which will allow the robot to locate objects in other, non-poking contexts. Developing object localization and recognition is the topic of Chapter 5.

Poking moves us one step outwards on a causal chain away from the robot and into the world, and gives a simple experimental procedure for segmenting objects. One way to extend this chain out further is to try to extract useful information from seeing a familiar object manipulated by someone else. This offers another opportunity for development – in this case, learning about other manipulators. Locating manipulators is covered in Chapter 6.

Another opportunity that poking provides is to learn how objects move when struck – both in general, for all objects, and for specific objects such as cars or bottles that tend to roll in particular directions. Given this information, the robot can strike an object in the direction it tends to move most, hence getting the strongest response and essentially evoking the "rolling affordance" offered by these objects. This is the subject of Chapter 7.

1.6 Interpersonal perception

Perception is not a completely objective process; there are choices to be made. For example, whether two objects are judged to be the same depends on which of their many features are considered essential and which are considered incidental. For a robot to be useful, it should draw the same distinctions a human would for a given task. To achieve this, there must be mechanisms that allow the robot's perceptual judgements to be channeled and moulded by a caregiver. This is also useful in situations where the robot's own abilities are simply not up to the challenge, and need a helping hand. This thesis identifies three channels that are particularly accessible sources of shared state: space, speech, and task structure. Robot and human both inhabit the same space. Both can observe the state of their workspace, and both can manipulate it, although not to equal extents. Chapter 8 covers a set of techniques for observing and maintaining spatial state. Another useful channel for communicating state is speech, covered in Chapter 9. Finally, the temporal structure of states and state transitions is the topic of Chapter 10.

1.7 Roadmap

Chapter 2	Overview of robot platforms and computational architecture
Chapter 3	Active segmentation of objects using poking
Chapter 4	Learning the appearance of oriented features
Chapter 5	Learning the appearance of objects
Chapter 6	Learning the appearance of manipulators
Chapter 7	Exploring an object affordance
Chapter 8	Spatially organized knowledge
Chapter 9	Recognizing and responding to words
Chapter 10	Interpersonal perception and task structure
Chapter 11	Discussion and conclusions