Perception and Perspective in Robotics

Paul Fitzpatrick (paulfitz@ai.mit.edu)
MIT Artificial Intelligence Laboratory, Cambridge, MA 02139 USA

Abstract

To a robot, the world is a sea of ambiguity, in which it will sink or swim depending on the robustness of its perceptual abilities. But robust machine perception has proven difficult to achieve. This paper 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.

Introduction

Perception is key to intelligent behavior. While the field of Artificial Intelligence has made impressive strides in replicating some aspects of cognition, such as planning and plan execution, machine perception remains distressingly brittle and task-specific. This paper directly addresses this brittleness by supporting perception through active, developmental, and interpersonal means.

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 paper.

Active and interpersonal perception both act as sources of the "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

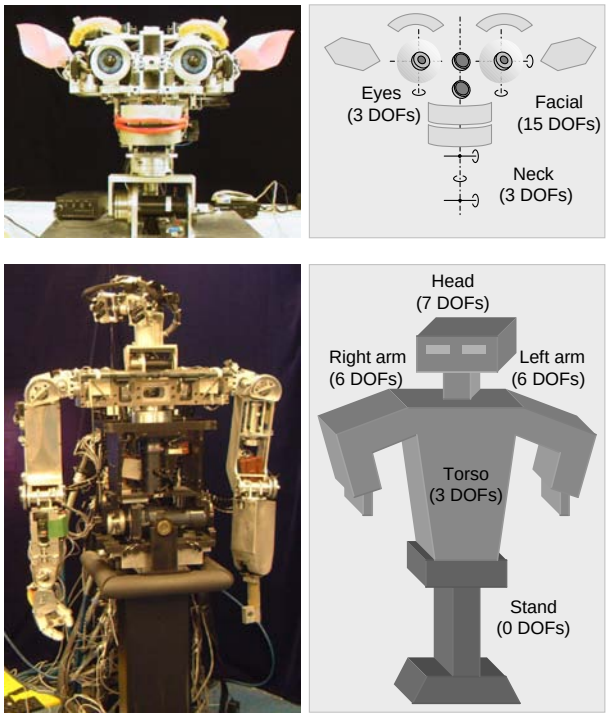


Figure 1: The robots Kismet (top) and Cog (bottom). Kismet is an expressive anthropomorphic head useful for human interaction work; Cog is an upper torso humanoid more adept at object interaction.

"caregiver". 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.

The robots

This work is implemented on two robots, Cog and Kismet (see Figure 1), Cog is an upper torso humanoid

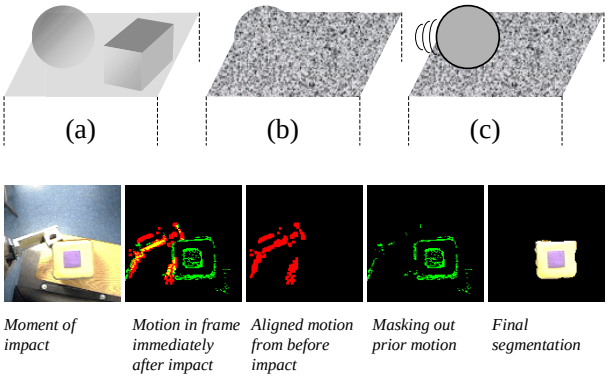


Figure 2: Cartoon motivation (top) for active segmentation (bottom). Human vision is excellent at figure/ground separation (top left), but machine vision is not (top center). Coherent motion is a powerful cue (top right) and the robot can invoke it by simply reaching out and poking around. The lower row of images show the processing steps involved. The moment of impact between the robot arm and an object, if it occurs, is easily detected – and then the total motion after contact, when compared to the motion before contact and grouped using a minimum cut approach, gives a very good indication of the object boundary (Fitzpatrick, 2003).

(Brooks et al., 1999) that has previously been applied to tasks such as visually-guided pointing (Marjanović et al., 1996), and rhythmic operations such as turning a crank or driving a slinky (Williamson, 1998). Kismet is an "infant-like" robot whose form and behavior is designed to elicit nurturing responses from humans (Breazeal et al., 2001). It is essentially an active vision head augmented with expressive facial features so that it can both send and receive human-like social cues.

Active perception

The most well-known instance of active perception is active vision. The term "active vision" is essentially synonymous with moving cameras. Active vision work on Cog is oriented towards opening up the potentially rich area of manipulation-aided vision, which is still largely unexplored. But 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.

As a concrete example of this idea, Cog was given a simple "poking" behavior, whereby it selects locations in its environment, and sweeps through them with its

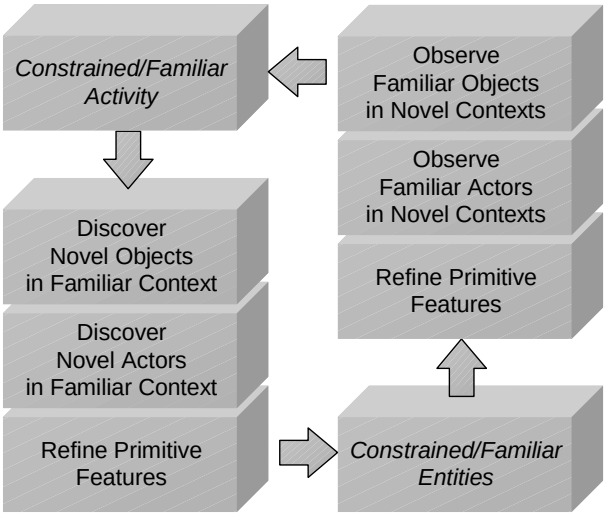


Figure 3: If the robot is engaged in a known activity (left), there may be sufficient constraint to identify novel elements within that activity. Similarly, if known elements take part in some unfamiliar activity, tracking those can help characterize that activity. Potentially, development is an open-ended loop of such discoveries.

arm (Fitzpatrick and Metta, 2002). If an object is within the area swept, then the motion signature generated by the impact of the arm with that object greatly simplifies segmenting that object from its background, and obtaining a reasonable estimate of its boundary (see Figure 2). The image processing involved relies only on the ability to fixate the robot's gaze in the direction of its arm. This coordination is easy to achieve either as a hard-wired primitive or through learning (Fitzpatrick and Metta, 2002). Within this context, it is possible to collect excellent views of the objects the robot pokes, and the robot's own arm.

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.

Developmental perception

The previous section showed how, with a particular behavior, the robot could reliably segment objects from the background (even if it is similar in appearance) by poking them. It can determine the shape of an object boundary in this special situation, even though it cannot do this normally. This is precisely the kind of situation that a developmental framework could exploit. Figure 3 shows how an open-ended developmental cycle might be possible. Particular, familiar situations allow the robot to perceive something about objects and actors (such as a human or the robot itself) that could not be per-

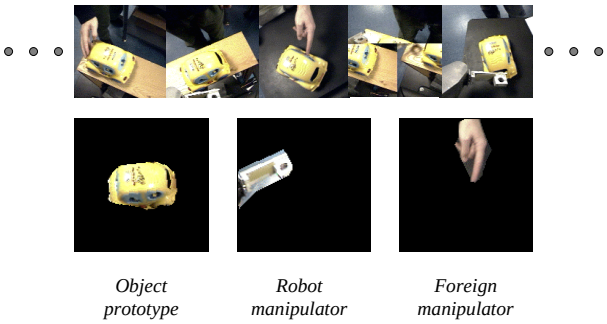


Figure 4: The top row shows sample views of a toy car that the robot sees during poking. Many such views are collected and segmented as described in (Fitzpatrick, 2003). The views are aligned to give an average prototype for the car (and the robot arm and human hand that acts upon it). To give a sense 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 arm and 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).

ceived outside those situations. These objects and actors can be tracked into other, less familiar situations, which can then be characterized and used for further discovery. Throughout, existing perceptual capabilities ("primitive features") can be refined as opportunities arise.

As a specific example of development, the segmented views provided by poking of objects and actors by poking can be collected and clustered as shown in Figure 4. Such views are precisely what is needed to train up an object detection and recognition system, and follow those objects and actors into other, non-poking contexts (Fitzpatrick, 2003).

As well as giving information about the appearance of objects, the segmented views of objects can be pooled to train up detectors for more basic visual features – for example, edge orientation. 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. Figure 5 shows an orientation filter trained up from such data that can work at much finer scales than normally possible when the filter is derived from an ideal edge model such as that of (Chen et al., 2000). The "catalog" of edge appearances found shows that the most frequent edge appearances is an "ideal" straight, noise-free edge, as might be expected (top of Figure 5) – but a remarkable diversity of other forms also occur which are far less obvious (bottom of Figure 5).

Interpersonal perception

Perception is not a completely objective process; there are choices to be made. For example, whether two ob-

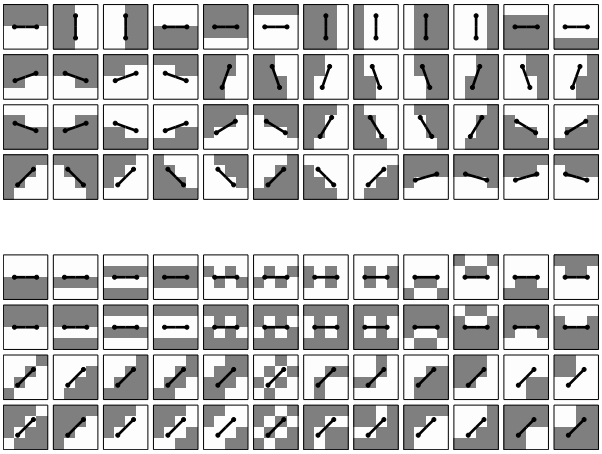


Figure 5: The empirical appearance of edges. Each 4×4 grid represents the possible appearance of an edge, quantized to just two luminance levels. The dark line centered in the grid is the average orientation that patch was observed to have in the training data. The upper set of patches are the most frequent ones that occur in training data consisting of about 500 object segmentations. The lower set of patches are a selection of patterns chosen to illustrated the diversity of possible patterns that can occur. The oriented features represented include edges, thin lines, thick lines, zig-zags, corners etc. It is difficult to imagine a set of conventional filters that could respond correctly to the full range of features seen here – all of which appeared multiple times in object boundaries in real images.

jects 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.

The most relevant literature at the intersection of communication, perception and development is that of language acquisition. It has been observed that language acquisition involves a search through a large search space of models guided by relatively sparse feedback and few examples. This so-called "poverty of the stimulus" relative to the complexity of the models being acquired is taken to imply that infants must have a good search strategy, with biases well matched to the nature of appropriate solution. This is a claim of innate constraints, and is historically controversial. Examples stressing under-determination in language learning include Quine's "Gavagai" example (Quine, 1960), where Quine invites us to imagine ourselves walking with a native guide in a foreign country, and seeing a rabbit

Human speech	Human action	Robot speech	Robot action
...	...	...	...
say yellow	[shows yellow toy]	yen	[looks at toy]
say yellow		yelo	
yes! say green		green	
...	...	...	...
yellow	[shows yellow toy]	yelo	[looks at toy]
left	[moves toy left]	left	[tracks toy]
green	[shows green toy]	green	[looks at toy]
right	[moves toy right]	right	[tracks toy]
yellow	[shows yellow toy]	yelo	[looks at toy]
left	[moves toy left]	left	[tracks toy]
...	...	...	...
yellow	[shows yellow toy]	yelo	[looks at toy]
next?		ahm left?	[nods left]
yes!	[moves toy left]	left!	
green	[shows green toy]	green	[looks at toy]
next?		ahm right?	[nods right]
yes!	[moves toy right]	right!	
...		...	

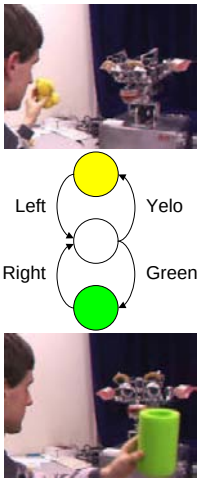


Figure 6: Extracts from a dialogue with Kismet. The first extract (say yellow...) illustrates how the robot's active vocabulary was extended. The second extract shows how a simple sorting activity was annotated for the robot. The final extract shows the robot being tested on its understanding of the form of the activity. The robot's utterances were transcribed phonetically, but are written in a simple form here for clarity. To the right is shown the very simple state machine model of the activity deduced by the robot.

pass just as the guide says "gavagai" – and then consider all the possible meanings this utterance might have. It is possible over time to learn from such situations (see (Steels and Kaplan, 1999) for an example of a robotic implementation). Pragmatic constraints can help speed the learner out of this sea of ambiguity. For example, (Markman, 1989) proposes a set of particular constraints infants might use to map words on to meanings. These constraints are along the style of the following (with many variations, elaborations and caveats) :-

- Whole-object assumption. If an adult labels something, assume they are referring to the whole object and not a part of it. categories" as opposed to thematic relationships. For example when child is asked to find "dog", may fetch the cat, but won't fetch dog-food.
- Mutual exclusivity. Assume objects have only one label. So look for an unnamed object to apply a new label to.

These constraints are intended to explain a spurt in vocabulary acquisition where infants begin to acquire words from one or a few examples – so-called fast-mapping. They are advanced not as absolute rules, but as biases on search.

Tomasello raises several objections to the constraint-based approach represented by Markman (Tomasello, 1997). Tomasello favors a "social-pragmatic" model of language acquisition that places language in the context of other joint referential activity, such as shared attention. He rejects the "word to meaning mapping" formulation of language acquisition. Rather, Tomasello proposes that language is used to invite others to experience the world in a particular way. From (Tomasello, 1997) :-

> The social-pragmatic approach to the problem of referential indeterminacy ... begins by rejecting truth conditional semantics in the form of the mapping metaphor (the child maps word onto world), adopting instead an experientialist and conceptualist view of language in which linguistic symbols are used by human beings to invite others to experience situations in particular ways. Thus, attempting to map word to world will not help in situations in which the very same piece of real estate may be called: "the shore" (by a sailor), "the coast" (by a hiker), "the ground" (by a skydiver), and "the beach" (by a sunbather).

Regardless of the utility of Tomasello's theory for its proper domain, language acquisition in infants, it seems a useful mindset for tackling interpersonal perception, which is in essence all about inviting the robot to view the world in a particular way.

Tomasello and his collaborators developed a series of experiments designed to systematically undermine the constraints approach to learning as typified by Markman and others. The experiments investigate word learning among children in the context of various games. The experiments are instructive in showing a range of situations in which simple rules based directly on gaze or affect would fail in at least one case or other. The experiments all avoid giving children (18-24 months old) ostentative naming contexts, and rather requiring them to pull out meanings from the "flow of interaction".

For example, in one experiment, an adult makes eye-contact with a child subject and says "Let's go find the toma." They then go to a row of buckets, each if which contains an object with which the child is not familiar.

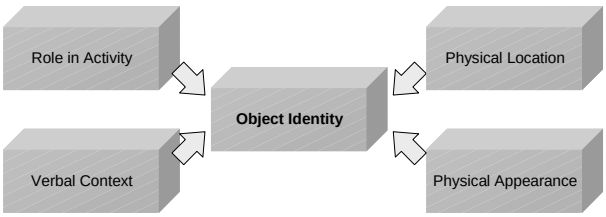
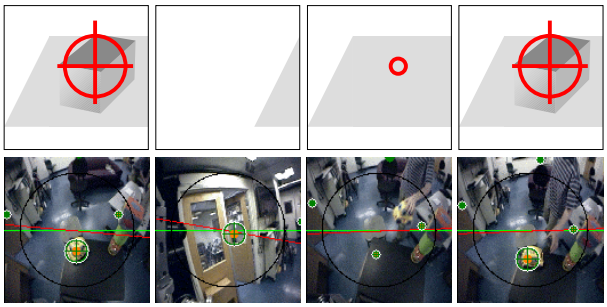


Figure 7: Perceptual judgements are fundamentally about identity: what is the same, what is different. Identity judgements should depend (at least) on activity, location, appearance, and verbal context. These in turn can be influenced by a caregiver.

One of these objects is randomly designated the "toma". If the session is a control, the adult goes directly to the bucket containing the toma, finds it excitedly and hands it to the child. Otherwise, the adult first goes to two other buckets in sequence, each time taking out the object, scowling at it, and replacing it, before "finding" the toma. Later, the child is tested for for the ability to comprehend and produce the new word appropriately. The results show equally good performance in the test and control scenarios. Tomasello argues that this situation counts against children using simple word learning rules such as "the object the adult is looking at while saying the novel word," "the first new object the adult looks at after saying the novel word," "the first new object the infant sees after hearing the novel word," or such variants.

Tomasello's theories and experiments are provocative, and suggest an approach quite different from the simple associative learning that is most often seen in robotics. Work on interpersonal perception on Cog draws heavily on (a grossly simplified caricature of) these ideas. The basic idea for interpersonal perception drawn from Tomasello's work is that information about the identity of an object needs to be easily coordinated between perception of activity, location, speech, and appearance (Figure 7). Without this flexibility, it is hard to imagine how scenarios such as the experiment described above or others proposed (Tomasello, 1997) could be dealt with.

It is currently unreasonable to expect the robot to understand the "flow of interaction" without help. Unaided segmentation of activity is a very challenging problem (see (Goldberg and Mataric, 1999) for one effort in the robotic domain). The human interacting with the robot can greatly simplify the task by making the structure of the activity unambiguous. Two mechanisms for this are particularly easy to deal with: vocalizations and location. If places and words are used consistently in an activity, then it is straightforward to model the basic "flow of interaction" they define. Figure 6 shows an example of this for a very simple sorting activity, implemented on the robot Kismet. Note that words are used here without the robot needing to know their meanings – it is sufficient that they be used consistently enough for the structure of the task to be made obvious.



Location marked, Target present | *Robot looks away* | *Robot looks back, Target is gone* | *Target reappears*

Figure 8: Keeping track of locations. Circles with cross-hairs represent locations that contain a particular object. If the object is removed, this is detected using color histograms (Swain and Ballard, 1991), and is indicated by a small circle without a cross-hair. The upper row is a cartoon sequence to illustrate what is happening in the views below, which are taken directly from Cog's egocentric map. Initially a yellow car is present on the table in front of Cog. The robot looks away to the door, and when it looks back, the car is no longer present. It then reappears and is immediately detected. This behavior, along with object tracking (which has also been implemented), give the basics of a representation of the robot's workspace.

The ability to interact verbally is currently being ported from Kismet to Cog, so that interpersonal perception can be integrated fully with the active and developmental work described earlier. Cog already has a well developed means to keep track of physical locations in an egocentric coordinate frame (see Figure 8). It is anticipated that this will be important in communicating the structure of activities to the robot, since even for adult humans cognition can often be traded off with physical space (Pelz, 1995; Kirsh, 1995). Recent work has focused on communicating the structure of search activity to the robot, and then using that to learn from a Tomasello-inspired 'find the toma' episode (Fitzpatrick, 2003).

Conclusions

This paper presented a snapshot of ongoing work to create an active, developing, malleable perceptual system for a robot. There is much remaining work to do. The immediate technical goal is to further develop mechanisms for communicating the structure of simple activities to a robot, translating this structure into a set of supervised learning problems for parts of the task which are difficult to communicate directly, and finally solving those problems with the guidance of a protocol for inducing feature selection. Figure 9 shows a schematic for how this may be achieved. The basic idea is for the robot to interact with the instructor vocally and through a shared workspace to acquire a "sequencing model" of an activity or task, and then to ground that model based on a

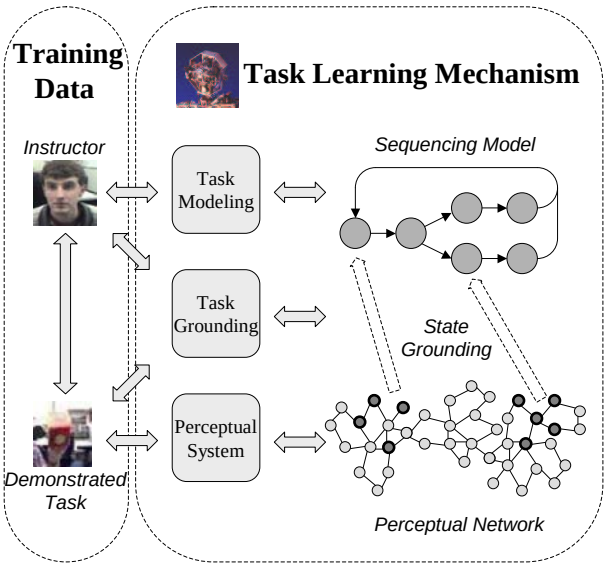


Figure 9: A summary of how task learning will be implemented. The instructor demonstrates the task while providing verbal and spatial cues. The cues are used to construct a model of the task. Generic machine learning methods are then used to ground this model in the robot's perceptual network, guided by previously grounded feature selection cues. The idea is to avoid ever presenting the robot with a hard learning problem; the learning algorithms are intended to be "decoders" allowing the human to communicate changes in representation, rather than to learn in the conventional sense.

demonstration of the task. This goal of this work is not to deal with general-purpose problem solving ability – for which better models are available (Clancey, 2002) – but to capture something of the quite general statistical learning abilities of young infants (Kirkham et al., 2002).

Acknowledgements

The author would like to thank the anonymous reviewers for their constructive feedback. Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement.

References

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48(1):57–86.

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics, A*, 31(5):443–453.

Brooks, R. A., Breazeal, C., Marjanovic, M., and Scassellati, B. (1999). The Cog project: Building a humanoid robot. *Lecture Notes in Computer Science*, 1562:52–87.

Chen, J., Sato, Y., and Tamura, S. (2000). Orientation space filtering for multiple orientation line segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(5):417–429.

Clancey, W. J. (2002). Simulating activities: Relating motives, deliberation, and attentive coordination. *Cognitive Systems Research*, 3(3):471–499.

Fitzpatrick, P. (2003). *From First Contact to Close Encounters: A developmentally deep perceptual system for a humanoid robot*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Fitzpatrick, P. and Metta, G. (2002). Towards manipulation-driven vision. In *IEEE/RSJ Conference on Intelligent Robots and Systems*, volume 1, pages 43–48.

Goldberg, D. and Mataric, M. J. (1999). Augmented markov models. Technical report, USC Institute for Robotics and Intelligent Systems.

Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. *Cognition*, 83(2):B35–B42.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, 73:31–68.

Marjanović, M. J., Scassellati, B., and Williamson, M. M. (1996). Self-taught visually-guided pointing for a humanoid robot. In *From Animals to Animats: Proceedings of 1996 Society of Adaptive Behavior*, pages 35–44, Cape Cod, Massachusetts.

Markman, E. M. (1989). *Categorization and naming in children: problems of induction*. MIT Press, Cambridge, Massachusetts.

Pelz, J. B. (1995). *Visual Representations in a Natural Visuo-motor Task*. PhD thesis, Brain and Cognitive Science, University of Rochester.

Quine, W. V. O. (1960). *Word and object*. Harvard University Press, Cambridge, Massachusetts.

Steels, L. and Kaplan, F. (1999). Collective learning and semiotic dynamics. In *European Conference on Artificial Life*, pages 679–688.

Swain, M. J. and Ballard, D. H. (1991). Colour indexing. *International Journal of Computer Vision*, 7(1):11–32.

Tomasello, M. (1997). The pragmatics of word learning. *Japanese Journal of Cognitive Science*, 4:59–74.

Williamson, M. (1998). Neural control of rhythmic arm movements. *Neural Networks*, 11(7-8):1379–1394.