In Berthouze, L., Kozima, H., Prince, C. G., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C. (Eds.)
Proceedings of the Fourth International Workshop on Epigenetic Robotics
Lund University Cognitive Studies, 117, ISBN 91-974741-3-4

The DayOne project: how far can a robot develop in 24 hours?

Paul Fitzpatrick
MIT CSAIL,
Cambridge, Massachusetts, USA

Abstract

What could a robot learn in one day? This paper describes the `DayOne` project, an endeavor to build an epigenetic robot that can bootstrap from a very rudimentary state to relatively sophisticated perception of objects and activities in a matter of hours. The project is inspired by the astonishingly rapidity with which many animals such as foals and lambs adapt to their surroundings on the first day of their life. While such plasticity may not be a sufficient basis for long-term cognitive development, it may be at least necessary, and share underlying infrastructure. This paper suggests that a sufficiently flexible perceptual system begins to look and act like it contains cognitive structures.

1. Introduction

Sometimes development is a rapid process. Consider the first day in the life of a foal, which can typically trot, gallop, groom itself, follow and feed from its mare, all within hours of birth (McCusker, 2003). Such precociousness is a common pattern for ungulates that evolved in habitats with sparse cover, where the newborn needs to (almost literally) hit the ground running or risk becoming a sitting target for predators.

In epigenetic robotics, we seek to create a "prolonged epigenetic developmental process through which increasingly more complex cognitive structures emerge in the system as a result of interactions with the physical and social environment" (Zlatev and Balkenius, 2001). Should the rapid development of the young of many species give us hope that this process could be much faster than we imagine? Perhaps not, since there is a difference between the development of perceptual and motor skills and the development of *cognitive* structures. Cognitive structures exhibit at least some flexibility of use and reuse, whereas perceptual and motor structures are closely tied to immediate sensing and actuation. But for those who see value in embodied, situated cognition, this distinction may seem unconvincing. Explicit in the work of Brooks was the suspicion that techniques for dealing with the uncertainty and ambiguity of perception and the subtleties of appropriate actuation are the work-horses of intelligence, and are presumably then key to cognitive structures: "This abstraction process is the essence of intelligence and the hard part of the problem being solved" (Brooks, 1991).

Work on the humanoid robot Cog has focused very much on rapid perceptual development. This paper describes the `DayOne` project, which was an attempt to integrate much of that work into a single, continuously running system. We hope to demonstrate that sufficiently advanced perceptual structures begin to look a lot like cognitive structures, since there is much flexibility in how they are constructed and used.

2. The stages of DayOne

The robot, upon startup, has the innate ability to turn towards and track movements, and to reach towards close objects, as described in earlier work (Metta and Fitzpatrick, 2003). Development of new perceptual skills begins in earnest right from the beginning, in the following stages :-

Low-level vision – The robot's low-level vision system is not complete upon startup. It has a filter which, by its construction, is fated to develop into an edge orientation detector, but to do so requires visual experience (Fitzpatrick, 2003b). This is an alternative to using carefully constructed model-based filters such as those developed in (Chen et al., 2000).

Mid-level vision – Once the low-level filters have stabilized, the robot learns to differentiate objects in its immediate surroundings. Again, the object recognition modules involved are fated to perform this task by their construction, but the actual set of objects that the robot learns to recognize is dependent on the contents of its environment (Fitzpatrick, 2003b,

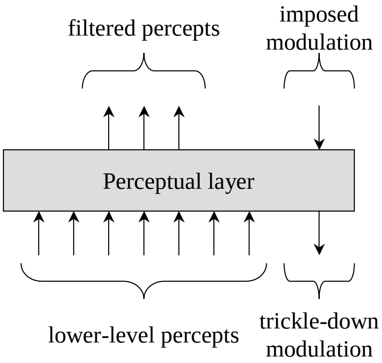


Figure 1: Each successively higher-level perceptual layer filters the one below it. For example, an object recognition layer finds clusters in the feature space presented to it and presents the clusters themselves as the features passed on to the next level. Modulation signals flow in the opposite direction to perception. They indicate when the output of the layer is overly detailed, or on the contrary insufficiently nuanced. If there is no way for the layer to make such a distinction, it passes the request on as 'trickle-down' modulation.

Metta and Fitzpatrick, 2003).

Mid-level audition – In parallel with visual development, the robot learns to differentiate utterances. This case is analogous to object differentiation. The actual set of utterances that the robot learns to recognize depends on what the humans in its environment choose to say (Fitzpatrick, 2003a).

High-level perception – As soon as the robot is familiar with some objects and utterances, it can begin to learn the causal structure of simple activities. The modules involved are fated to perform this task, but the activities, the utterances, and the objects involved are all a function of the environment.

3. Layered, coupled development

At any moment in time, Cog's sensory input is distilled into a distributed set of percepts. In lower-level modules, these percepts are quantitative in nature, and closely tied to the details of the immediate sensor input – for example, the output of the edge orientation detector. In higher level modules, the percepts become more qualitative in nature, and less sensitive to accidental or irrelevant details – for example, the output of object recognition. Still higher-level percepts are even more qualitative, such as a percept that corresponds to seeing a familiar object, or hearing a familiar sound.

Figure 1 shows an abstract view of each perceptual layer. The primary direction of information flow is from lower levels to higher levels, with details being dropped along the way. A layer is useful if it drops irrelevant details; each layer has its own heuristics about what is relevant. For example, the object recognition module attempts to minimize the effects of pose. Of course, these heuristics will not always be appropriate, and only the overall task can determine that. Hence there is a *modulation* signal that operates in the reverse direction. It can request that more or less detail be supplied for recently activated percepts, or provide a training signal to drive differentiation. This is somewhat analogous to the behavior of neural networks.

The contract between each perceptual layer is as follows :-

- ▷ The "semantics" of what activation means for each line projecting to a higher layer will be preserved as much as possible over time. In other words, an output line will be activated for the same situations in the future as it was in the past.
- ▷ An important exception is that the semantics of an output line may change due to attempts to refine or purify it (so that it is less affected by noise, for example, or responds to the same basic property in an extended range of situations).
- ▷ Requests for refinement are handled locally if possible, otherwise passed back to a lower layer.
- ▷ Input lines with very similar activation should be detected and merged.

The contract is important because in actual implementation, layers change in both an incremental and batch manner, and this requires careful regulation to stay consistent over time. For example, the object recognition layer quickly creates a new output line when the robot appears to experience a novel object; periodically, all object clusters are examined and optimized using an off-the-shelf clustering algorithm in MATLAB, and the new output line may turn out to be redundant. The output of this clustering is mapped to the current output lines in such a way as to maximally preserve their semantics. Excess lines are never removed, but simply made identical, so that there are no abrupt changes in semantics.

4. Generalization of percepts

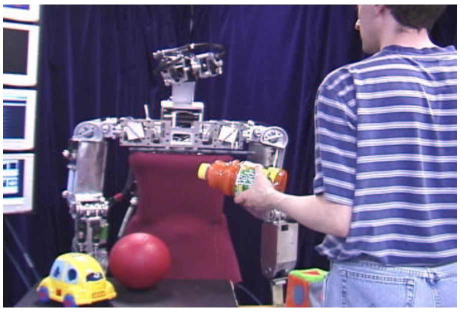
Cog continually searches for useful new ways to perceive the world, where being 'useful' means having predictive power. This search is performed by considering combinations of existing percepts, when heuristics suggest that such combinations may be fruitful. There are three categories of combinations :-

- **Conjunctions:** if two percepts are noted to occur frequently together, and rarely occur without each other, a composite percept called their conjunction is formed. From then on, this percept is activated whenever the two component percepts do in fact occur together in future.
- **Disjunctions:** if two percepts are noted to occur frequently together, but also occur independently in other situations, a composite percept called their disjunction is formed. This percept is activated whenever one or both of the two component percepts occur.
- **Implications:** Causal versions of the above composite percepts, which are sensitive to event order and timing, are also considered.

These composite percepts are intended to enable the robot to make meaningful generalizations, by allowing the same physical event to be viewed in ways that are sensitive to past history. Figure 2 demonstrates the use of such generalizations to link an object with its name through an extended search activity. This is a simplified version of an experiment carried out on human infants by Tomasello (Tomasello, 1997), which in combination with other experiments seeks to rule out many heuristics proposed for fast word learning in the infant development literature (Markman, 1989). A human and the robot engage in a simple search activity, where the human goes looking for an object, which they fail to find immediately. The robot is then tested to see if it can associate the object eventually found with its name, which is given at the start of the search and never mentioned in the presence of its referent.

Searches are presented to the robot as a game following a fairly strict script: first the word 'find' is uttered, then the name of the object to search for is mentioned. Then a series of objects are fixated. The word 'no' is uttered if the object is not the target of the search. The word 'yes' indicates that the search has succeeded, and the object currently fixated is the target of the search. The meaning of these words is initially entirely in the mind of the human. But the robot can discover them using event generalization, if it experiences a number of searches for objects whose name it already knows.

The word spoken after 'find' gets a special composite implication percept associated with it,



Human speech	**Human action**	**Robot speech**	**Robot action**
...	...	...	...
say	[shows ball]	say	[looks at ball]
beh		ball	
say	[shows car]	say	[looks at car]
keh		car	
say	[shows cube]	say	[looks at cube]
keh		cube	
say		say	
	[waits]	cube	
	[shows ball]		[looks at ball]
say		say	
	[waits]	ball	
...	...	...	...
	[attracts attention]		[looks at person]
find		find	
ball		ball	
no	[shows cube]	no	[looks at cube]
no	[shows car]	no	[looks at car]
yes	[shows ball]	yes	[looks at ball]
...	...	...	...
	[attracts attention]		[looks at person]
find		find	
toma		toma	
no	[shows ball]	no	[looks at ball]
no	[shows cube]	no	[looks at cube]
yes	[shows bottle]	yes	[looks at bottle]
say	[shows cube]	say	[looks at cube]
		cube	
say	[shows bottle]	say	[looks at bottle]
		toma	
...	...	...	...

Figure 2: Extracts from a dialogue with Cog. First, the robot is taught to name the object it is looking at when the word 'say' is spoken. This is done by speaking the word, then prompting the robot with a short utterance (beh and keh in this example). Short utterances prompt the robot to take responsibility for saying what it sees. A link is formed between 'say' and prompting so that 'say' becomes an alternate way to prompt the robot. Then the robot is shown instances of searching for an object whose name it knows (in the one example given here, the ball is the target). Finally, the robot is shown an instance of searching where an unfamiliar object name is mentioned ('toma'). This allows it to demonstrate that it has learned the structure of the search task, by correctly linking the unfamiliar name ('toma') with the target of search (a bottle). This experiment is similar, but not identical, to one considered by Tomasello for human infants (Tomasello, 1997).

let us call it `word-after-find` (of course, no such symbols are used internally, and the word 'find' initially has no special significance – it could be replaced with any other word, such as 'seek,' 'cherchez,' or 'fizzle-tizzle'). When the search is for an object whose name the robot knows (through a pre-established disjunction) that is also noted as a simultaneous event with `word-after-find`. The object seen when 'yes' (`object-with-yes`) is said matches this and an implication is formed between the two. This implication is sufficient to link an *unknown* word following 'find' with the object seen when 'yes' is said, via the `word-after-find` and `object-with-yes` generalizations (again, the choice of the word 'yes' has no special significance, and could be replaced with 'frob').

5. Discussion and conclusions

This paper gave a snapshot of project implemented on the humanoid robot Cog, and showed one task the robot could learn in a single day – which involved training low-level orientation filters, object and utterance recognition modules, and activity understanding, all in one integrated system. There are many exciting research projects that continue to press the boundaries of what can be achieved through robot learning and development – (Weng et al., 2000, Metta, 2000, Roy and Pentland, 2002) etc. It seems that by its nature, the field of epigenetic robotics will advance by a combination of innovation, aggregation, and consolidation. In our own system, a promising direction of research seems to be to take the most 'cognitive-like' ability of the robot (understanding patterns of activity through composite percepts) and find ways to push something analogous back into the very lowest levels of perception. It is interesting to imagine what a robot of tomorrow could learn in a single hour, if everyone's most advanced methods of today become just a part of the smallest building blocks of tomorrow's systems!

Acknowledgements

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

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence Journal*, 47:139–160. originally appeared as MIT AI Memo 899 in May 1986.

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.

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

Fitzpatrick, P. (2003b). Object Lesson: Discovering and learning to recognize objects. In *Proceedings of the 3rd International IEEE/RAS Conference on Humanoid Robots*, Karlsruhe, Germany.

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

McCusker, M. (2003). Investigation of the effects of social experience on snapping intensity in Equus caballus foals. Master's thesis, Virginia Polytechnic Institute and State University.

Metta, G. (2000). *Babybot: a study into sensorimotor development.* PhD thesis, LIRA-Lab, DIST.

Metta, G. and Fitzpatrick, P. (2003). Early integration of vision and manipulation. *Adaptive Behavior*, 11(2):109–128.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

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

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2000). Autonomous mental development by robots and animals. *Science*, 291(5504):599–600.

Zlatev, J. and Balkenius, C. (2001). Why "epigenetic robotics"? In *Proceedings of the First International Workshop on Epigenetic Robotics*, volume 85, pages 1–4. Lund University Cognitive Studies.