

What can I control? A framework for robot self-discovery

Aaron Edsinger

Computer Science and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 US
edsinger@csail.mit.edu

Charles C. Kemp

Computer Science and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 US
cckemp@csail.mit.edu

Abstract

In this paper we present a developmental progression for a humanoid robot that uses mutual information to discover controllable perceptual categories. Previously we have shown that the robot can discover a visual category that corresponds with its hand from less than 5 minutes of interaction with a human. Here, we show how this discovery can be used to adapt the robot's perceptual and motor systems such that the robot can subsequently discover its fingers. In this way, a robot can expand its control over the world so that newly mastered actions can open up new realms of influence, and new realms of influence can lead to the mastery of new actions.

1. Introduction

A robot can be more than a passive observer of the world as it learns and develops. It can exert influence over its immediate surroundings through, for example, manipulation, vocalizations, or social interaction. Ideally, a robot would incrementally discover what it can control, adapt its perceptual and motor systems to this discovery, and then refine its ability to influence these aspects of its world.

Touchette and Lloyd (Touchette and Lloyd, 2004) have noted in their work on information-theoretic control that the degree to which the final state of a system, Φ , is controlled by the state of a controller, A , can be represented by their mutual information, $I(\Phi; A)$. Mutual information provides a general way to quantify the extent to which a system's actions control observed characteristics of the world.

Mutual information can be used to rank visual categories by how well the robot can control them. The highest ranked categories should correspond to controllable parts of the robot's body. Previously, we demonstrated that a robot can use this approach to discover a visual category that corresponds with its

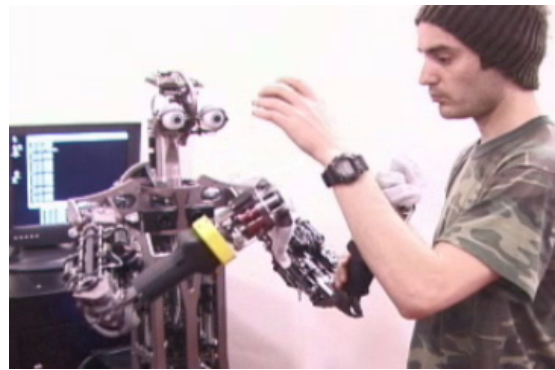


Figure 1: Domo, the humanoid robot used in this work, interacting with a person.

hand from less than 5 minutes of natural interaction with a human (Kemp and Edsinger, 2006b).

In this paper we extend and review this work. We present a developmental progression that is guided by the discovery of controllable perceptual categories. This discovery allows us to expand the robot's controllers and perceptual system, enabling further discovery. We present two stages of the developmental progression: hand-discovery and finger-discovery. The robot first detects the controllability of a visual category corresponding to its hand through random motion of its arm. Next, we extend the perceptual and control system to include this discovery. The robot then discovers the controllability of a visual category corresponding to its finger through finger motion while fixating its hand. Results are presented using the robot shown in Figure 1.

The next section of this paper discusses related work. In Section 3, we present the developmental progression and its application to hand and finger discovery. Section 4, describes implementation details for the robot's perceptual system. In Section 5, we describe quantitative results, and in Section 6, we conclude with a discussion of future work.

2. Background

Researchers have successfully used mutual information for developmental learning on robots. Roy, Schiele, and Pentland used a clustering algorithm based on mutual information to find visual and auditory clusters that link objects with words (Roy et al., 1999). Kaplan and Hafner, (Kaplan and Hafner, 2005), and Olsson, Nehaniv, and Polani, (Olsson et al., 2005), have used similarity measurements related to mutual information to autonomously develop sensorimotor maps for robots.

One of the main results in this paper is that the robot discovers its own hand and fingers through motion cues. Visual motion can serve as a powerful cue for robot development. For example, (Olsson et al., 2005, Gross et al., 1999) have implemented systems which can learn to predict the optic flow generated by robot ego-motion. In (Kemp and Edsinger, 2006b) we demonstrated that during everyday human-robot interaction, the fastest moving convex regions in the image often correspond to developmentally relevant regions such as hands, fingers, heads, and eyes. We review this motion feature in Section 4.1.

Researchers have also used visual motion to detect the robot’s hand. Fitzpatrick and Metta (Fitzpatrick et al., 2003) used image differencing to detect ballistic motion and optic-flow to detect periodic motion of the robot’s hand. Natale, (Natale, 2004), applied image differencing for detection of periodic hand motion with a known frequency, while (Arsenio and Fitzpatrick, 2003) used the periodic motion of tracked points. Gold and Scasselati explored the idea of temporal contingency for the detection of motion related to the robot’s body (Gold and Scasselati, 2006). They used image differencing and motor babbling to learn a time window that models the delay between executing a motor command and detecting visual motion. They then used this time window to detect the onset of motion that was likely to correspond with the body. In contrast to this work, our approach provides a prediction of where the hand is expected to appear in the image.

In the next section we present a framework for discovery-based development where transitions between stages are constructed by hand. This framework is compatible with autonomous approaches as well. Recent work by Oudeyer and Kaplan (Oudeyer and Kaplan, 2004, Oudeyer et al., 2005) has demonstrated autonomous development guided by an intrinsic curiosity drive. The drive balances a robot’s exploration, learning, and task difficulty, allowing it to pursue learning situations of increasing perceptual and motor complexity. Barto (Barto et al., 2004) has also developed intrinsically motivated behaviors to guide autonomous

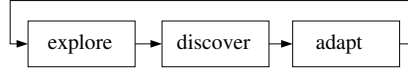


Figure 2: The process of self-discovery: The robot explores the space of possible motor actions and perceptual categories. A category is discovered to be controllable by a motor action. This discovery enables the robot to adapt its perceptual and motor systems. The process is then repeated.

development. In this context, discovering the controllability of perceptual categories could be viewed as an intrinsic motivation that guides autonomous development.

3. Discovery Framework

In this section we present a developmental framework for discovery and describe its application to a robot manipulator.

3.1 Framework

We can view our approach as comprising a succession of developmental stages, each one building from knowledge acquired during the previous stage. This progression is guided by the discovery of controllable perceptual categories. Our progression is illustrated in Figure 2. During a given stage $S^i = [A^i, O^i]$, the robot is capable of actions A^i and observations O^i . Given this general setup, we propose the following:

1. Development stage $S^i \rightarrow S^{i+1}$
 - (a) **Explore.** The robot performs actions, A^i , using available controllers. While doing so, it senses the world and collects observations O^i .
 - (b) **Discover.** The robot segregates O^i into perceptual categories. It then determines the extent to which its actions, A^i , control features, Φ^i , associated with these categories using mutual-information $I(\Phi^i; A^i)$. The system then selects the most controllable perceptual categories.
 - (c) **Adapt.** Given these controllable perceptual categories and the related actions, the robot adapts its perceptual system and controllers, which results in a new set of possible actions, A^{i+1} , and observations, O^{i+1} .

The key point is that given a controllable perceptual category, the robot can adapt its perceptual system and controllers so as to facilitate the next stage of development.

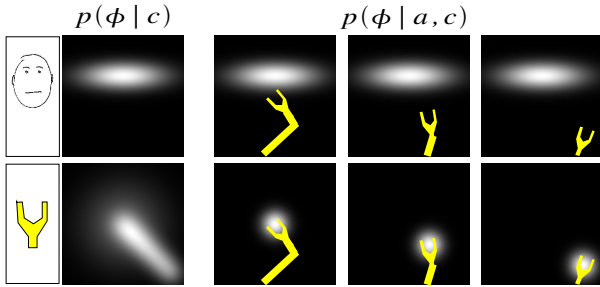


Figure 3: In this idealized example, the visual attention system selects image patches containing two visual categories: the robot’s hand (top row) and the human’s face (bottom row). These image patches are clustered into visual categories. The visual locations, ϕ , of the image patches for each visual category c are used to approximate the spatial distributions $p(\phi|c)$ (left column) and $p(\phi|a,c)$ (the three columns on the right), where a is the robot’s arm configuration. Conditioning the spatial distribution of the robot’s hand a substantially reduces the uncertainty, while the spatial distribution for the human’s face is unaffected by the robot’s arm. Mutual information is then used to rank the visual category for the robot’s hand as more controllable than the human’s face.

This idea is related to efference copy mechanisms (S.J. Blakemore, 2001), where the robot learns a forward or inverse model of its body. For example, by learning a model of ego-motion, a robot can ignore the ego-motion from its optic flow measurement; by learning a model of gravity forces upon its arms, it can make the experienced joint torques pose invariant; and by discovering the location of its hand in the image, it can fixate the hand in order to focus perceptual resources at its fingers.

3.2 Discovery of the Hand and Fingers

We now describe the application of this framework to hand and finger discovery.

1. Stage 1: Discovery of the hand.

- (a) **Explore.** The robot randomly samples from the joint space of the manipulator, creating motion in front of its camera. The visual system collects a set of image patches that correspond with fast moving regions. The visual system is described in Section 4.1.
- (b) **Discover.** As depicted in Figure 3, the image patches are clustered into distinct categories. Image patches of the hand are assumed to cluster into at least one distinct visual category. The spatial distribution of each category is estimated. The mutual information between a category’s spatial dis-

tribution and its time-aligned joint posture is used to rank each visual category. The highest ranked category corresponds with the robot’s hand.

- (c) **Adapt.** The discovery of the hand allows for modification of the robot’s perceptual and motor system. The hand location can now be predicted within the image. A new, specialized controller can place the hand within the camera’s field-of-view at a known location and the head can fixate the hand. The salience of motion near the expected hand location can also be amplified and background motion discounted.

2. Stage 2: Discovery of the fingers

- (a) **Explore.** The robot brings its hand into the field-of-view and its head fixates the hand. The robot then explores the joint workspace of its finger. The hand is held still at a known visual location and the attention system selects image patches that correspond with fast moving regions near this location.
- (b) **Discover.** As in Stage 1, the collected image patches are clustered into distinct categories. Image patches of the finger-tip are assumed to cluster into at least one distinct visual category. The spatial distribution of each category is estimated. The mutual information between a category’s spatial distribution and its time-aligned joint posture is used to rank each visual category. The highest ranked category corresponds with the finger-tip.
- (c) **Adapt.** The location of the fingertip can now be predicted in the image given the predefined arm-pose. New perceptual and control systems can be added based on this discovery to prepare for subsequent development stages. For example, an appearance model of the fingertip could be constructed and used to visually detect and servo the finger.

Hand discovery is an important precursor to finger discovery. It allows the robot to focus its perceptual resources on the hand. Motion at the robot’s fingertip would otherwise be difficult to detect due to the large state-space of possible head, arm, and finger postures.

4. Implementation

Within this section we describe the details of the implementation. The visual system collects 10 salient image patches for each frame of video. The proprioceptive system associates each of the image patches



Figure 4: A sample output of the visual attention system selecting salient image regions during hand-discovery. The input is 120×160 image frames captured at 20 – 30 fps in an everyday, cluttered environment (right). The visual attention system prefers rapidly moving shapes that are approximately convex. The edge motion is measured relative to the global image motion (left). Image patches are selected from the top 10 regions (the circles around the robot’s hand).

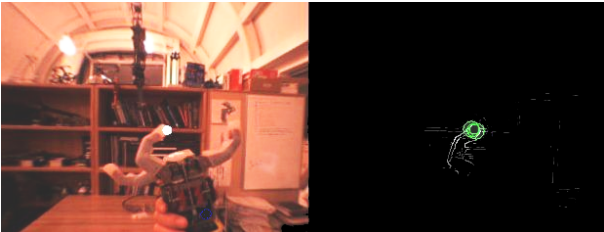


Figure 5: (Left) A typical view of the robot watching its hand during finger-discovery. Discovery of the hand enables the hand to be brought into the field-of-view and fixated. This facilitates finger discovery. The visual attention system selects the fastest moving convex shape (fingertip) in the image (Right).

with the time-aligned configuration of the robot’s joints. Next, the system clusters the image patches and the joint configurations, computes density estimates for the spatial distributions of the image clusters, and uses mutual information to rank the visual categories by their controllability.

4.1 Visual Attention

In the spirit of Itti, Koch, and Niebur (Itti et al., 1998), the first component of the visual system is a visual attention system that finds salient locations in the image. The visual attention system collects image patches that correspond with rapidly moving, approximately convex shapes, see Figure 4.

Previously, we have used the same algorithm to successfully detect the tip of a tool held within a robot’s rotating hand. We describe the visual attention system in (Kemp and Edsinger, 2006a), so we only present a high-level overview here. For each region in a frame of video the attention system assigns a saliency value that is higher for faster moving

shapes and higher for shapes that are more circular. In particular, rapidly moving edges that are approximately tangent to a circle of some radius will produce an especially strong response. In order to reduce the impact of camera motion, the motion of these edges is computed relative to the global image motion. The visual attention system also ensures that the selected regions are spatially distinct, and filters out any regions that have saliency values below threshold.

With respect to the computer vision literature, this attention system is a form of spatio-temporal interest point operator that gives the position and scale of significant parts of the image (Laptev, 2005). As we describe in (Kemp and Edsinger, 2006a), the algorithm has similarities to classic image processing techniques such as the distance transform, medial axis transform, and hough transform for circles (Forsyth and Ponce, 2002).

For each frame of video, we select the 10 most salient regions available, or fewer if necessary. For each of these selected salient regions, we collect a square image patch.

4.2 Clustering

We cluster the image patches and the joint configurations independently using K-means (Duda et al., 2001) to give us k_v visual categories and k_a joint configuration clusters. For each image patch, i , we create a feature vector for clustering that consists of a 16×16 hue and saturation histogram. For each joint configuration, a , we select either the four joint angles of an arm or the single joint angle of a finger, depending on the developmental stage. We convert each of the joint angles, θ_n , into a 2D Cartesian coordinate, x_n , resulting in the feature vector

$$x_n = [\cos \theta_n, \sin \theta_n]. \quad (1)$$

This feature vector allows us to use Euclidean distance when clustering the arm configurations without worrying about angular wrap-around.

4.3 Spatial Density Estimation

For each of the resulting visual categories, we model the visual locations of the member image patches with the probability distribution $p(\phi|c)$, which represents the chance of seeing an image patch of category, c , at location, ϕ . c is the index for one of the k_v visual categories, and ϕ is a 2D coordinate that describes the position of the image patch in a head-centered, spherical coordinate system. Like the sensory ego-sphere of (Peters et al., 2001), this spherical coordinate system is fixed to the body’s frame of reference in order to abstract away from the details of the head and camera configuration and allow image positions from distinct head poses to be related to one another. We estimate $p(\phi|c)$ for each visual

category, c , using 61x61 bin 2D histograms. If we define

$$g(x, y) = \delta(\text{round}(\frac{61}{2\pi}(x - y))), \quad (2)$$

where $\delta(d) = \begin{cases} 1 & \text{if } d = 0 \\ 0 & \text{otherwise} \end{cases}$, then

$$p(\phi|c) \approx \frac{1}{\sum_{i \in c} w(i, c)} \sum_{i \in c} w(i, c) g(T_{i_h}(i_x), \phi). \quad (3)$$

The kinematic transform T_{i_h} maps the pixel coordinate for image patch, i_x , to spherical coordinates given the configuration of the head when the patch was captured, i_h . We use a weighting function $w(i, c)$ to reduce the influence of image patches that are far from the mean of the cluster. For each visual category, c , $w(i, c)$ is a spherical Gaussian fit to the cluster's members. Each dimension of the 2D histogram maps to a $[-\pi, \pi]$ range of the corresponding dimension of ϕ , where $\phi = [0, 0]$ is directly in front of the robot and $\phi = [\pi, \pi]$ is directly behind the robot.

We compute these density estimates in order to estimate the controllability of each visual category using mutual information. $p(\phi|c)$ can also predict where visual category c is likely to appear, which can be advantageous. For example, Torralba (Torralba, 2003) describes a framework that uses similar information to improve visual search and modulate the saliency of image regions.

The mutual information estimation also depends on $p(\phi|a, c)$ and $p(a|c)$, which we estimate using the approximations

$$p(a|c) \approx \frac{1}{\sum_{i \in c} w(i, c)} \sum_{i \in c} w(i, c) \delta(i_a - a) \quad (4)$$

$$p(\phi|a, c) \approx \frac{1}{\sum_{i \in c} w(i, c) \delta(i_a - a)} \sum_{i \in c} w(i, c) \delta(i_a - a) g(T_{i_h}(i_x), \phi), \quad (5)$$

where a is the index for one of the k_a joint clusters, and i_a is the index for the joint configuration at the time that image patch i was acquired. $p(a|c)$ gives a list of k_a numbers for each visual category c . $p(\phi|a, c)$ consists of a set of k_a 61x61 2D histograms for each visual category, c .

4.4 Mutual Information

We now wish to use mutual information to determine the extent to which the robot's joint configuration controls the visual location of each visual category. As illustrated in Figure 3, we can use the mutual information, $I_c(\Phi; A)$, between the random variables Φ and A (corresponding with arguments ϕ and a and to visual category c) to rank the k_v visual categories

according to how much they are controlled by the robot's arm or finger configuration. By definition

$$I_c(\Phi; A) = H_c(\Phi) - H_c(\Phi|A). \quad (6)$$

We can estimate the entropy, $H_c(\Phi)$, and conditional entropy, $H_c(\Phi|A)$, using our density estimates, since

$$H_c(\Phi) = - \sum_{\phi} p(\phi|c) \log p(\phi|c) \quad (7)$$

$$H_c(\Phi|A) = - \sum_a p(a|c) \sum_{\phi} p(\phi|a, c) \log p(\phi|a, c). \quad (8)$$

In Stage 1, the visual category, c_{best} , with the highest value for $I_c(\Phi; A)$, corresponds to the robot's hand. We can use our estimate for the distribution $p(\phi|a, c_{best})$ to predict the location of the hand within the image. In our tests, we used the maximum likelihood estimate of the hand's position,

$$\phi_{hand}(a) = \text{Argmax}_{\phi} p(\phi|a, c_{best}), \quad (9)$$

which returns the most likely spherical coordinate for the hand, ϕ_{hand} , given arm configuration a . In Stage 2, the most likely spherical coordinate for the fingertip is similarly estimated.

5. Results

We tested our approach on the 6 DOF arm, 4 DOF hand, and 8 DOF head of the humanoid robot, Domo, pictured in Figure 1. The visual attention system, clustering, and mutual information computations were performed off-line, though an online implementation is feasible.

5.1 The Protocol

Prior to the experiments we manually moved the arm around its reachable workspace and collected 4000 samples of arm postures that could be safely commanded to the motor controllers.

During the hand-discovery experiments a person interacted with the robot. These interactions involved full-body motion, hand waving, presentation of objects, as well as physical contact with the robot's arm. The robot periodically ($0.25Hz$) executed a reaching movement to a random arm posture generated from the collected data. In (Kemp and Edsinger, 2006b) we reported positive hand-discovery results with both periodic eye saccades and a stationary head. In this paper we review results with the stationary head.

During the finger-discovery phase, a single finger was swept periodically ($2Hz$) through its joint range. The robot's camera was fixated on the hand and the arm was held stationary, but background motion and small hand disturbances did occur. For both experiments, each interaction lasted about 2 minutes and

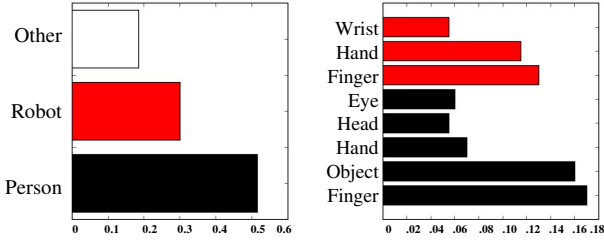


Figure 6: We hand-labelled categories for 200 image patches randomly selected from the image patches collected by the visual attention system. A patch was labelled as a person if it selected either a hand, finger, head, eye, or object in the hand. A patch was labelled as a robot if it selected either the robot’s hand, finger, or wrist. Patches that were neither person nor robot were labelled as other. The left plot shows the probability of each category and the right plot shows the probability of each sub-category.

generated approximately 2000 images and proprioceptive samples.

5.2 Visual Attention System

We first evaluated the ability of the visual attention system to autonomously select potentially controllable perceptual categories. Using only motion and shape as cues, the majority of the image patches selected by the visual system correspond with categories relevant to human-robot interaction, including the person’s head, eyes, hands, and fingers, the objects being presented, and the robot’s hand and fingers. Figure 4 shows a sample output. Even during eye saccades the system rarely selects the background. Figure 6 shows results during the hand-discovery phase, where over 50% of 200 test patches were of human features, and over 30% were of the robot’s hand.

5.3 Discovery of the Hand

We first tested the ability of the system to discover the robot’s hand using the stationary head data set. Image clustering was tested using the parameters $k_v \in \{2, 10, 20\}$, and clustering over the arm configurations used $k_a = 9$. Figure 7 depicts the image patches closest to the mean for each cluster when $k_v = 2$, demonstrating the ability of the system to segregate the robot’s hand from other image patches. The clusters were ranked according to mutual-information, and the clusters containing the robot’s hand ranked the highest for all values of k_v . As shown in Figure 8, this top cluster can predict the location of the hand in the image,.



A



B

Figure 7: The results of image clustering during the hand-detection experiment with $k_v = 2$. Image patches were encoded as a 16×16 hue and saturation histogram. For each cluster, the 24 image patches closest to the cluster mean are shown. Cluster A predominantly contains the robot’s hand while cluster B contains the human subject.

5.4 Discovery of the Fingers

For finger-discovery, the arm was servoed to an arm configuration that brings the hand into view. For this experiment a kinematically derived posture was used, though, as Figure 8 shows, a discovered posture could be used instead. The robot watched as a single finger was periodically moved through its joint range. The expected hand location was used to discount motion feature detections far from the hand.

We tested our approach over two different data sets using $k_v \in \{2, 10\}$ and $k_a = 3$ for the K-means clustering. The relatively small amount of fingertip motion visible within the scene required a coarse discretization of the joint space. The histogram sizes were the same as for the hand. Figure 9 depicts the image patches closest to the mean for each cluster when $k_v = 2$. The robot was able to correctly detect the controllability of its finger and predict its location in the image given $k_v \in \{2, 10\}$. The results for $k_v = 2$ are shown in Figure 10.

6. Discussion

In this paper we have described a developmental progression that uses mutual information to guide robot discovery of controllable perceptual categories. We have shown how a humanoid robot can detect the controllability of a visual category corresponding to its hand. We have also demonstrated how this discovery allows the robot to then discover the controllability of its fingers. There are several avenues for future work. First, the transition be-

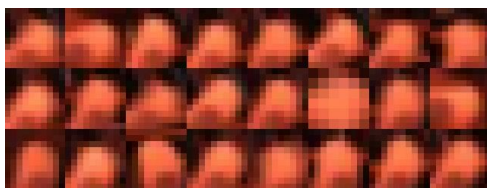


A



B

Figure 8: Hand prediction results for $k_v = 2$. Cluster A has the higher mutual information and is the better predictor of the hand in the image. Each prediction, marked by a circle, corresponds to an arm configuration cluster ($k_a = 9$). The images are ranked (left-to-right, top-to-bottom) according to the maximum value of $p(\phi|c)$. As expected, the system is unable to make a correct prediction when the hand is out of view.



A



B

Figure 9: The results of image clustering during the finger-detection experiment with $k_v = 2$. For each cluster, the 24 image patches closest to the cluster mean are shown. Cluster A predominantly contains the robot's finger while cluster B contains the human subject.



A



B

Figure 10: Finger prediction results for $k_v = 2$ and $k_a = 3$. Cluster A is ranked with the highest mutual information and correctly predicts the location of the finger in the image. Cluster B is ranked lower.

tween the hand-discovery and finger-discovery phases was programmed by hand. Also, the motor exploration consisted of simple random movements. Random arm movements could instead evolve towards refined finger motions as the controllability of these objects is discovered. This relates to the work of (Oudeyer et al., 2005), where motor exploration increases in complexity during development and is closer to the motor exploration seen in infants (von Hofsten, 2004). Finally, we have assumed that the visual system can meaningfully categorize the salient image patches using K-means with color histograms. Although this was sufficient for our data and provided a clean example, it would be inadequate for categories that require more subtle visual distinctions.

The approach described in this paper is an initial exploration of a broader theme. We have demonstrated how a robot can expand its control over the world so that newly mastered actions open up new realms of influence, and new realms of influence lead to the mastery of new actions.

Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful and thorough comments. This work was sponsored by Toyota Motor Corporation: Autonomous Manipulation Capabilities for Partner Robots in the Home.

References

- Arsenio, A. and Fitzpatrick, P. (2003). Exploiting cross-modal rhythm for robot perception of objects. In *Proceedings of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems*.
- Barto, A. G., Singh, S., and Chentanez, N. (2004).

- Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of International Conference on Developmental Learning (ICDL)*.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons, New York.
- Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning About Objects Through Action: Initial Steps Towards Artificial Cognition. In *Proceedings of the 2003 IEEE International Conference on Robotics and Automation (ICRA)*, Taipei, Taiwan.
- Forsyth, D. A. and Ponce, J. (2002). *Computer Vision: a modern approach*. Prentice Hall.
- Gold, K. and Scassellati, B. (2006). Learning acceptable windows of contingency. *Connection Science: Special issue on developmental learning*, 18(2):217–228.
- Gross, H., Heinze, A., Seiler, T., and Stephan, V. (1999). A neural architecture for sensorimotor anticipation. *Neural Networks*.
- Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259.
- Kaplan, F. and Hafner, V. (2005). Mapping the space of skills: An approach for comparing embodied sensorimotor organizations. In *4th IEEE International Conference on Development and Learning (ICDL-05)*, pages 129–134, Osaka, Japan.
- Kemp, C. C. and Edsinger, A. (2006a). Robot manipulation of human tools: Autonomous detection and control of task relevant features. In *5th IEEE International Conference on Development and Learning (ICDL-06), Special Session on Classifying Activities in Manual Tasks*, Bloomington, Indiana.
- Kemp, C. C. and Edsinger, A. (2006b). What can i control?: The development of visual categories for a robot’s body and the world that it influences. In *5th IEEE International Conference on Development and Learning (ICDL-06), Special Session on Autonomous Mental Development*.
- Laptev, I. (2005). On space-time interest points. *Int. J. Computer Vision*, 64(2):107–123.
- Natale, L. (2004). *Linking Action to Perception in a Humanoid Robot: A Developmental Approach to Grasping*. PhD thesis, LIRA-Lab, DIST, University of Genoa.
- Olsson, L., Nehaniv, C. L., and Polani, D. (2005). From unknown sensors and actuators to visually guided movement. In *4th IEEE International Conference on Development and Learning (ICDL-05)*.
- Oudeyer, P.-Y. and Kaplan, F. (2004). Intelligent adaptive curiosity: a source of self-development. In Berthouze, L., Kozima, H., Prince, C. G., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C., (Eds.), *Proceedings of the 4th International Workshop on Epigenetic Robotics*, volume 117, pages 127–130.
- Oudeyer, P.-Y., Kaplan, F., Hafner, V. V., and Whyte, A. (2005). The playground experiment: Task-independent development of a curious robot. In Bank, D. and Meeden, L., (Eds.), *Proceedings of the AAAI Spring Symposium on Developmental Robotics*, pages 42–47.
- Peters, R. A., Hambuchen, K. E., Kawamura, K., and Wilkes, D. M. (2001). The sensory egosphere as a shortterm memory for humanoids. In *IEEE-RAS International Conference on Humanoid Robots*, pages 451–459, Tokyo, Japan.
- Roy, D., Schiele, B., and Pentland, A. (1999). Learning audio-visual associations using mutual information. In *Workshop on Integrating Speech and Image Understanding*, Greece.
- S.J. Blakemore, C.D. Frith, D. W. (2001). The cerebellum is involved in predicting the sensory consequences of action. *NeuroReport*, 12(11):1879–1884.
- Torralba, A. (2003). Modeling global scene factors in attention. *Journal of Optical Society of America A: Special Issue on Bayesian and Statistical Approaches to Vision*, 20(7):1407–1418.
- Touchette, H. and Lloyd, S. (2004). Information-theoretic approach to the study of control systems. *PHYSICA A*, 331:140.
- von Hofsten, C. (2004). An action perspective on motor development. *Trends in Cognitive Science*, 8:266–272.