What Can I Control?: The Development of Visual Categories for a Robot's Body and the World that it Influences

Charles C. Kemp
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts
cckemp@csail.mit.edu

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

***Abstract*— We present a developmental perceptual system for a humanoid robot that autonomously discovers its hand from less than 2 minutes of natural interaction with a human. The perceptual system combines simple proprioceptive sensing with a visual attention system that uses motion to select salient regions. We show that during natural interactions with a person, the majority of the selected visual regions consist of significant body parts on the human and robot (hands, fingers, and the human's head). The system visually clusters the selected image regions, models their spatial distribution over a sensory sphere, and uses mutual information to determine how much the clusters are influenced by the robot's arm. In our tests, the visual cluster that most strongly relates to the robot's arm primarily contains images of the robot's hand, and has a spatial distribution that can predict the location of the robot's hand in the image as a function of the arm's configuration.** [1]

I. Introduction

What can I control? This is a critical question for any autonomous intelligent system. The incremental discovery of the factors that we have influence over allows us to direct our resources to more productive ends and expand our opportunities for action. For an embodied system, a first step on this path can be the discovery of the body. In this paper, we present a developmental perceptual system for a humanoid robot that autonomously discovers its hand in less than 2 minutes of natural interaction with a human.

We first present a method for visual attention that uses motion and shape to select salient regions that are important for a robot that cooperates with people. We show that during natural interactions with a person, these regions tend to correspond with the person's head and hand, and with the robot's hand. Second, we show that using mutual information and non-parametric density modeling of spatial distributions over a sensory sphere can be used to predict the locations of salient visual categories and determine which of these visual categories can be influenced by the robot's arm. In our tests, this method ranks visual categories associated with the robot's hand higher than those related to the person with whom the robot is interacting. It also predicts the location of the robot's hand in the image as a function of the configuration of the robot's arm.

[1]This work was sponsored by the NASA Systems Mission Directorate, Technical Development Program under contract 012461-001.

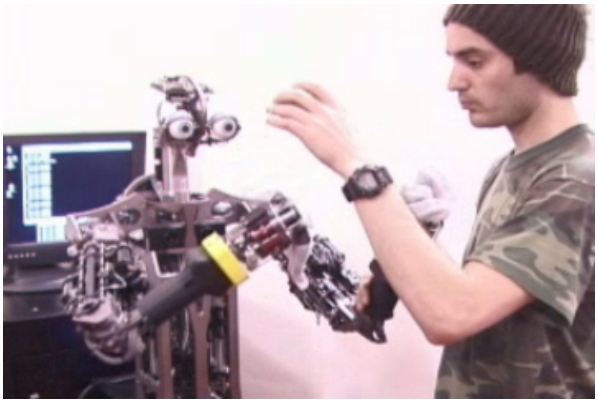


Fig. 1. The robot used in this work, Domo, interacting with a person.

We begin by discussing related work in Section II. In Section III we provide a review of the approach used in the visual attention system. Next, in Section IV we discuss the techniques for clustering, density estimation, and mutual information applied to discovery of the robot's hand. Section V describes our results, and Section VI outlines possible directions for future work. Finally, Section VII provides concluding remarks.

II. Related Work

The work for this paper includes a visual attention system that uses image motion and image edges to select salient regions of the image. Like the work of Itti, Koch, and Niebur [8], it attempts to rapidly find salient locations in the image. With respect to the computer vision literature, it is a form of spatio-temporal interest point operator that gives the position and scale of significant parts of the image [12]. The multi-scale histograms generated by the visual attention system can be related to scale-space methods and have similarities to classic image processing techniques such as the distance transform, medial axis transform, and hough transform for circles [5].

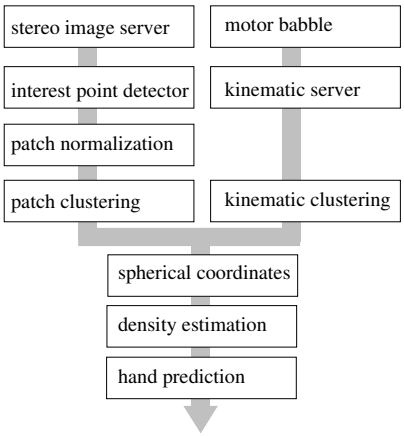


Fig. 2. The architecture of the described system. During human-robot interaction, the visual attention system produces scale and orientation normalized image patches of interest points. Image patches are clustered by HSV value and the arm configurations are clustered by joint angle. Each patch location is mapped into spherical camera coordinates and a density estimate of the spatial distribution is computed for each visual cluster. The mutual information between the patch locations and the arm configurations is used to predict the hand's appearance in the image.

We have previously used the visual attention system of this paper to detect the tip of a moving human tool within the robot's hand [11]. This work relied on a calibrated kinematic model of the robot and only modeled the tip of a rigidly grasped tool within the robot's hand. In this paper, no kinematic model of the arm is assumed and a diverse set of salient regions are considered to be important .

We use mutual information to find visual features that can be controlled by the robot. Many 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 [15]. Kaplan and Hafner, [3], and Olsson, Nehaniv, and Polani, [14], have used similarity measurements related to mutual information to autonomously develop sensorimotor maps for robots. Olsson, Nehaniv, and Polani's work included image sensors on the sensory map, and discovered the effect of actuators on global optical flow.

Our system makes use of a spherical camera model in the body's frame of reference to abstract away from the details of the head and camera configuration. This type of approximation has a strong relationship to the Sensory Ego-Sphere of [7].

One of the main results in this paper is that the robot autonomously discovers its own hand. Many researchers have created methods of visual hand detection through motion. Fitzpatrick and Metta [4] used image differencing to detect ballistic motion and optic-flow to detect periodic motion of the robot's hand. Natale, [13], applied image differencing for detection of periodic hand motion with a known frequency, while [1] used the periodic motion of tracked points. Kemp, [9], created a wearable system that discovers the hand of the wearer with methods that are similar to what we use in this paper. However, his method required a parametric kinematic model and absolute orientation measurements.

Gold and Scassellati explored the idea of temporal contingency for the detection of motion related to the robot's body [6]. 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 [6], our method focuses on spatial relationships rather than time. We do not explicitly model temporal contingency, nor do we make use of the measured velocity of the joint angles during motor babbling. Incorporating these temporal relationships could be a useful extension.

Our method for discovering what can be controlled by the robot is very strongly related to the formal information based model for controllability presented by Touchette and Lloyd in [16]. Their formal information-theoretic framework for analyzing control systems supports our approach.

III. The Visual System

The visual system provides robust detection of fast moving and roughly convex features for the robot's attention system. In this section we provide an overview of the system, which is described in more detail elsewhere [9]–[11].

As in [10], the optical flow computation first uses block matching to estimate the most likely motion for each edge and a 2D covariance matrix that models the matching error around this best match. Next, a global 2D affine motion model is fit to these measurements. Finally, the significance of the motion for each edge is computed as the Mahalanobis distance between the edge's measured motion model and the global motion model. This motion measurement incorporates both the magnitude of the edge's motion and the uncertainty of the measurement.

The visual attention system implicitly assumes that salient regions will consist of many strongly moving edges that are approximately tangent to a circle at some scale. Due to projection, the detector will tend to respond more strongly to objects moving close to the camera and fast moving objects. It will also respond strongly to shapes with approximately convex projections onto the image.

The input to the interest point detector consists of a set of weighted edges, e_i, where each edge i consists of a weight, w_i, an image location, x_i, and an angle, θ_i. We use a Canny edge detector to produce edge locations and orientations, to which we assign weights that are equal to the estimated motion. Each edge votes on locations in a scale-space that correspond with the centers of the coarse circular regions the edge borders. For each edge, we add two weighted votes to the appropriate bin locations at each integer scale s.

Fig. 3. An example of the set of 2D histograms, m_s, produced by the interest point detector when given a rectangle of edges weighted equally with unit motion. The scale, s, increases from left to right. Strong responses in the planes correspond with corners, parallel lines, and the ends of the rectangle.

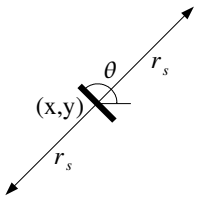


Fig. 4. This figure depicts the approximate locations in the image of the two votes at scale s cast by an edge with orientation θ and position (x, y).

As depicted in Figure 4, within the original image coordinates the two votes are approximately at a distance r_s from the edge's location and are located in positions orthogonal to the edge's length. We assume that the angle θ_i denotes the direction of the edge's length and is in the range $[-\frac{\pi}{2}, \frac{\pi}{2})$, so that no distinction is made between the two sides of the edge.

For each scale s there is a 2D histogram that accumulates votes for interest points, as shown in Figure 3. The planar discretization of these histograms is determined by the integer bin length, l_s, which is set with respect to the discretization of the scale-space over scale, $l_s = \lceil \beta(r_{s+0.5} - r_{s-0.5}) \rceil$, where β is a scalar constant that is typically close to 1.

We define r_s such that r_{s+1} is a constant multiple of r_s, where s ranges from 1 to c inclusive. We also define r_s to be between r_{max} and r_{min} inclusive, so that

$$r_s = \exp(\frac{\log(r_{max}) - \log(r_{min})}{c-1}(s-1) + \log(r_{min})) \quad (1)$$

Setting r_{min} and r_{max} determines the volume of the scale-space that will be analyzed, while c determines the resolution at which the scale-space will be sampled.

We compute the bin indices, (b_x, b_y), for the 2D histogram at scale s with

$$b_s(x, \theta) = \text{round}(\frac{1}{l_s}(x + r_s \begin{bmatrix} \cos(\theta + \frac{\pi}{2}) \\ \sin(\theta + \frac{\pi}{2}) \end{bmatrix})), \quad (2)$$

which adds a vector of length r_s to the edge position x and then scales and quantizes the result to find the appropriate bin in the histogram.

Algorithmically, we iterate through the edges adding their weighted contributions to the appropriate bins. We can write the equation for the resulting interest point detection maps, m_s, using delta functions, δ, so that

$$m_s(u) = \sum_i w_i(\delta(u - b_s(x_i, \theta_i)) + \delta(u - b_s(x_i, \theta_i + \pi))), \quad (3)$$

where $\delta(x) = \begin{cases} 1 \text{ if } (x_x = 0) \wedge (x_y = 0) \\ 0 \text{ otherwise} \end{cases}$.

In order to soften the effects of our block discretization, we low-pass filter each 2D histogram, m_s, with a separable, truncated, FIR Gaussian, which is approximately equal to giving each edge a Gaussian vote distribution, since

$$G \star m_s = \sum_i w_i(G(u - b_s(x_i, \theta_i)) + G(u - b_s(x_i, \theta_i + \pi))), \quad (4)$$

where G is an ideal Gaussian. This is also approximately equal to blurring the weighted edge map by scale varying Gaussians, or blurring the scale-space volume across scale.

Ideally, the values of corresponding interest points resulting from a shape would be invariant to translation, scaling, and rotation of the shape. We introduce two scalar functions n_s and n_θ to reduce scale dependent variations and angle dependent variations respectively, so that

$$m_s(u) = n_s \sum_i n_{\theta_i} w_i(G(u - b_s(x_i, \theta_i)) + G(u - b_s(x_i, \theta_i + \pi))). \quad (5)$$

We determine the values for these two functions empirically using a calibration pattern.

We filter the scale-space for points that are locally maximal. We then use fourier-based shape descriptors from [9] to filter these local maxima for points in the scale-space that correspond with extended and enclosing curves in the image. Finally, we take the 10 remaining points with the highest responses, and use the corresponding positions and scales within the image to extract image patches.

IV. Discovery of the Hand

The visual system provides us with 10 salient image patches for each image. These patches are tagged with time-aligned joint angles from the robot's proprioceptive system. As described in this section, we then cluster these sensory inputs, compute density estimates of the patches' spatial distribution, and then use mutual information to determine which visual cluster relates most strongly to the robot's hand.

A. Clustering

We cluster the image patches and the arm configurations independently using K-means [2] to give us k_v visual categories and k_a arm configuration clusters. For each image patch, we first scale the patch to a standard size, and then multiply the result by a Gaussian mask in order to reduce the influence of the corners of the square image patch. We then create a feature vector consisting of a 16x16 hue and saturation histogram. For the arm configurations, we convert each of the 4 joint angles, θ_n, of an arm configuration into a 2D cartesian coordinate, x_n, resulting in an 8 dimensional feature vector, where

$$x_n = [cos(\theta_n), sin(\theta_n)]. \quad (6)$$

This allows us to use Euclidian distance when clustering without worrying about angular wrap-around.

B. Density Estimation

For the resulting visual clusters, we now model the positions of the image patches with a probability distribution $p(\phi, c)$, which represents the chance of seeing an image patch of category, c, at location, ϕ, where c is the index for one of the k_v visual categories, and ϕ is the 2D coordinate that describes the position of the patch in the head-centered spherical camera model. We estimate $p(\phi|c)$ for each visual category, c, using 2D histograms, so that

$$p(\phi|c) \approx \frac{1}{\sum_{i \in c} p(i|c)} \sum_{i \in c} p(i|c)\delta(\text{round}(T_h(i_x) - \phi)), \quad (7)$$

where $\delta(d) = \begin{cases} 1 \text{ if } d = (0,0) \\ 0 \text{ otherwise} \end{cases}$, $p(i|c)$ is the estimated probability of image patch i given visual category c, and T_h maps the pixel coordinate for a patch, i_x, to spherical coordinates given the configuration of the head, h. In this work, T_h uses a kinematic model of the robot's head/camera system. We model $p(i|c)$ using a spherical Gaussian. 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. Using non-parametric estimates of distributions, we can easily find estimates for $p(\phi|a,c)$ and $p(a|c)$ as well, where a is the index for one of the k_a arm clusters. By themselves, these density estimates have utility, since they can be used to predict where visual categories will appear, and hence serve as priors for visual search and object detection.

C. Mutual Information

We now wish to determine the extent to which the position distribution for each visual category, c, is influenced by the arm's configuration. Intuitively, we would expect that the person's face would remain in a similar visual position over the majority of the configurations of the robot's arm. Likewise, we would expect for the visual position of the robot's arm to be strongly related to the current proprioceptively sensed arm configuration.

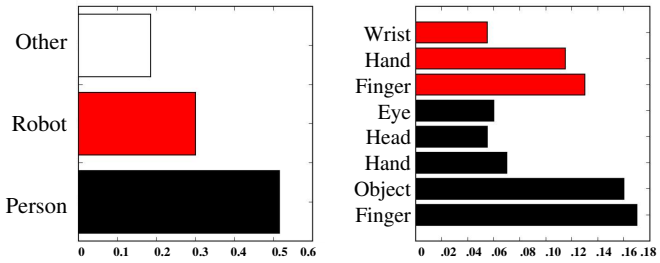


Fig. 5. We hand-labelled the categories for 200 image patches randomly collected from the attention system. A patch was labelled as a person if it selected either a hand, finger, head, eye, or object in the hand. A patch was labelled as a robot if it selected either the robot hand, finger, or wrist. Patches that were neither person or 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.

We can use the mutual information, $I_c(\Phi; A)$, between the random variables Φ and A (corresponding with arguments ϕ and a) to rank the k_v visual clusters according to how much they are influenced by the robot's arm configuration. By definition

$$I_c(\Phi; A) = H_c(\Phi) - H_c(\Phi|A). \quad (8)$$

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

$$H_c(\Phi) = -\sum_{\phi} p(\phi|c) \log(p(\phi|c)) \quad (9)$$

$$H_c(\Phi|A) = -\sum_{a} p(a|c) \sum_{\phi} p(\phi|a,c) \log(p(\phi|a,c)). \quad (10)$$

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

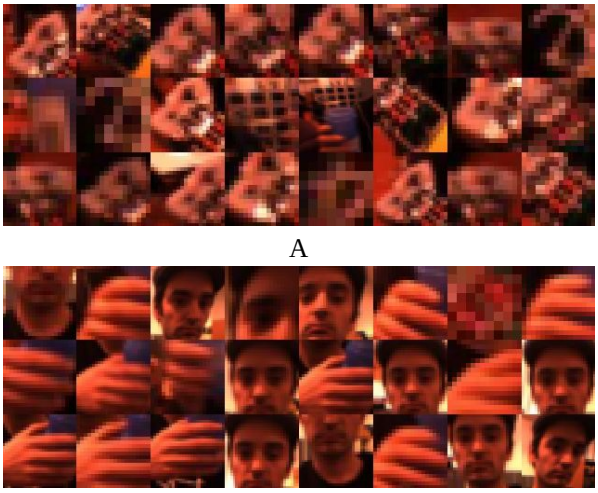
$$\phi_{hand}(a) = \text{Argmax}_{\phi}(p(\phi|a, c_{best})), \quad (11)$$

which returns the most likely spherical coordinate for the hand, ϕ_{hand}, given the arm configuration a.

V. Results

We tested our approach on the 6 DOF arm and 8 DOF head of the humanoid robot, Domo, pictured in Figure 1. The robot head was held fixed while the arm was allowed to explore its workspace through motor babble. During the exploration, a human subject was also allowed to freely interact with the robot through waving, presenting objects, and interacting with the arm. Each interaction lasted about 1 minute, generating approximately 1000 proprioceptive and

Fig. 6. A random sample of image patches collected by the motion based attention system.



A

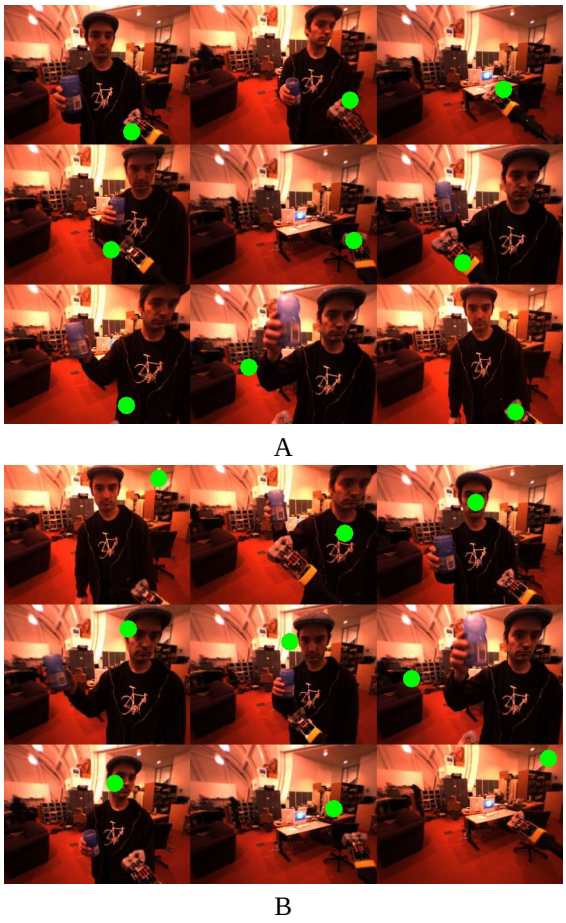
B

Fig. 7. Image clustering was done with K-means (k=2). Image patches were encoded as a $16x16$ hue and saturation histogram. The top 24 image patches for each cluster are shown. Cluster A predominantly contains the robot's hand while cluster B contains the human subject.

image points of data. The visual attention system, clustering, and mutual information computations were performed off-line, though an online implementation is feasible. The overall architecture of our system is illustrated in Figure 2.

Using motion and shape as cues, the visual attention system was robust at detecting the person's hands, fingers, head, eyes, and held object, as well as the robots hand, wrist, and fingers. As Figure 5 illustrates, over 50% of the image patches selected by the attention system were of salient human features, while over 30% were of the robot's hands. Figure 6 shows a random sample of the image patches selected by the interest point operators.

We successfully tested our approach over two different data sets using the parameters $k_v = [2, 10, 20]$ for the K-means clustering of the image patches. Figure 7 depicts the top image patches for each cluster when $k_v = 2$, showing



A

B

Fig. 8. Hand prediction results. In this example, two visual clusters, A and B, were learned. Cluster A has the higher mutual information and is the better predictor of the hand in the image. Cluster B is a poor predictor and has lower mutual information as expected. Each prediction, marked in green, corresponds to one of 9 arm configuration clusters. The maximum point of the 2D histogram, conditioned on an arm cluster, is taken as the predicted hand location within the spherical camera model. For visualization, the prediction is transformed into pixel coordinates and displayed with the camera image chosen when the arm was nearest the arm cluster mean. 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 prediction when the hand is out of view.

the ability to segregate the robot's hand from other image patches. The K-means clustering of the arm configurations used $k_a = 9$. After computing the mutual information between the spatial distribution of image patches and the arm pose, we consistently found that the cluster of image patches determined to be under the robot's control primarily contained the robot's hand. We also found that this cluster could reliably predict the appearance of the hand in the image, as shown in Figure 8.

VI. Simplifying Assumptions and Future Work

We have made several simplifying assumptions throughout this paper that could be worth revisiting in future work. First,

we have assumed that the visual system can meaningfully categorize the salient image patches without feedback from the proprioceptive system and the position distribution for Φ. A full developmental system would benefit from gradual differentiation of the visual categories based on feedback from other modalities, or simultaneous clustering akin to Roy's work [15]. Likewise, we chose to use a simple clustering algorithm, K-means, and feature vector, color histogram, to produce the visual clusters in order to provide a clean example. The visual system could more effectively categorize the image patches with a more sophisticated clustering algorithm and patch descriptor.

When using the 2D histograms as non-parametric density estimators for distributions involving Φ, one should ideally compensate for the distortions caused by mapping the surface of a sphere to a plane. For this work we ignored this complexity and computed a simple linear conversion. This is a reasonable approximation, since in our data set the image patches appear over a small portion of the sphere. Similarly, the motion of the head will bias the distributions to be stronger over the regions of the sphere that are looked at more frequently. We could attempt to ensure that the head's motion samples some region of the sphere with near uniformity, or normalize the areas of the histogram based on how often they have been observed. For this paper, we chose the simplest model, which ignores the impact of the head motion on the distribution, and implicitly incorporates it as another source of randomness.

Finally, the algorithm should work when the head is moving, due to the global motion model used in the optic flow and the spherical camera model. Currently we have only tested it with data taken while the head was held at a fixed position.

VII. Conclusions

In this paper, we have demonstrated a perceptual system for a humanoid robot that can quickly discover its own hand. The learning occurs autonomously during natural human interaction in an everyday, unstructured setting. The visual attention system, using motion and shape cues, is capable of autonomously selecting regions corresponding to significant body parts of the robot and the human. We model the spatial distribution of these regions over a sensory sphere, and use mutual information to help determine what the robot can influence within its environment.

References

[1] A. Arsenio and P. Fitzpatrick. Exploiting cross-modal rhythm for robot perception of objects. In *Proceedings of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems*, December 2003.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, New York, 2001.

[3] Kaplan F. and Hafner V. V. 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, 2005.

[4] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. 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, May 2003.

[5] D. A. Forsyth and Jean Ponce. *Computer Vision: a modern approach*. Prentice Hall, 2002.

[6] Kevin Gold and Brian Scassellati. Learning acceptable windows of contingency. *Connection Science: special issue on developmental learning*, In Press.

[7] R. A. Peters II, K. E. Hambuchen, K. Kawamura, and D. M. Wilkes. The sensory ego-sphere as a shortterm memory for humanoids. In *IEEE-RAS International Conference on Humanoid Robots*, pages 451–459, Tokyo, Japan, November 2001.

[8] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[9] Charles C. Kemp. *A Wearable System that Learns a Kinematic Model and Finds Structure in Everyday Manipulation by using Absolute Orientation Sensors and a Camera*. PhD thesis, Massachusetts Institute of Technology, May 2005.

[10] Charles C. Kemp and Aaron Edsinger. Visual Tool Tip Detection and Position Estimation for Robotic Manipulation of Unknown Human Tools. Technical Report AIM-2005-037, MIT Computer Science and Artificial Intelligence Laboratory, 2005.

[11] Charles C. Kemp and Aaron Edsinger. Robot manipulation of human tools: Autonomous detection and control of task relevant features. In *In Submission to: 5th IEEE International Conference on Development and Learning (ICDL-06)*, Bloomington, Indiana, 2006.

[12] I. Laptev. On space-time interest points. *Int. J. Computer Vision*, 64(2):107–123, 2005.

[13] Lorenzo Natale. *Linking Action to Perception in a Humanoid Robot: A Developmental Approach to Grasping*. PhD thesis, LIRA-Lab, DIST, University of Genoa, 2004.

[14] Lars Olsson, Chrystopher L. Nehaniv, and Daniel Polani. From unknown sensors and actuators to visually guided movement. In *4th IEEE International Conference on Development and Learning (ICDL-05)*.

[15] Deb Roy, Bernt Schiele, and Alex Pentland. Learning audio-visual associations using mutual information. In *Workshop on Integrating Speech and Image Understanding*, Greece, 1999.

[16] Hugo Touchette and Seth Lloyd. Information-theoretic approach to the study of control systems. *PHYSICA A*, 331:140, 2004.