

International Journal of Humanoid Robotics
© World Scientific Publishing Company

Object Recognition from Multiple Percepts

Artur M. Arsenio

*Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
The Stata Center, 32 Vassar Street, Room G376
Cambridge, Massachusetts 02139, USA
arsenio@csail.mit.edu*

This paper presents a perceptual system that exploits human caregivers as catalysts for the humanoid robot Cog to perceive and learn about objects, scenes, people, and the robot itself. A broad spectrum of machine learning problems are addressed for object recognition across several categorization levels. The paper introduces a new complex approach to object recognition based on the integration of multiple percepts. Training data for all learning mechanisms is automatically generated from actions by an embodied agent, so that the robot develops categorization autonomously.

Cognitive capabilities of the humanoid robot are developmentally created, starting from infant-like abilities for detecting, segmenting, and recognizing percepts over multiple sensing modalities. Human caregivers provide a helping hand for communicating such information to the robot, by acting on the objects, inducing their compliant perception from these human-robot interactions.

Keywords: Object Recognition; Humanoid Robots; Robot Learning; System Integration

1. Introduction

This paper introduces a new complex approach to object recognition for the humanoid robot Cog. Objects may have different meanings in different contexts - a rod is labelled as a pendulum if oscillating with a fixed end-point. From a visual image, a large piece of fabric on the floor is most often a tapestry, while it is most likely a bed sheet if it is found on a bed. But if a person is able to feel the fabric's material or texture, or the sound that it makes (or not) when grasped with other materials, then he might determine easily the fabric's true function. Object recognition draws on many sensory modalities and object's behavior, which inspired this paper's approach. Hence, an object will be recognized based on:

- ▷ local features such as its color
- ▷ the sound it produces or often associated to it
- ▷ being estimated (or not) as a face or the robot own body
- ▷ cross-modal features (using visual/sound patterns)
- ▷ contextual features, used to identify the identity, location, size, angle and depth most probable for an object or person given the image of a scene
- ▷ typical scene in which they occur

2 *Artur M. Arsenio*

This paper exploits the fact that several recognition problems in different fields are closely tied. For instance, a common definition of the computer vision problem of object recognition¹ is:

Definition *We are given a database of object models and a view of the real world. For each object in the model database, the object recognition and localization problem consists in answering the following two questions:*

- ▷ *Is the object present in the observed scene?*
- ▷ *If present, what are the 3D pose parameters (translation and rotation parameters) with respect to the sensor coordinate system?*

If possible, the system should learn unknown objects from the observed data.

This definition relies on a very strong assumption: Objects are recognized by their appearance solely. This paper advocates a more general approach. All of the following are sometimes true:

- ▷ *Objects are recognized by their appearance - color, luminance, shape, texture*
- ▷ *Objects have other physical features, such as mass, of a dynamic nature*
- ▷ *The dynamic behavior of an object varies depending on its actuation*
- ▷ *Temporal information of the object motion structure - the kinematics - is necessary for identifying functional constraints²*
- ▷ *Objects are situated in the world, which may change an object's meaning depending on context*
- ▷ *Objects have an underlying hierarchic tree structure - which contains information concerning objects that are reducible to other objects, i.e., that originated by assembling several objects*
- ▷ *There is a set of actions that can be exerted on an object, and another set of actions that an object can be used for (e.g., a nail can be hammered, while a hammer is used for hammering). Therefore, a set of affordances³ also intrinsically define (albeit not uniquely) an object.*

1.1. Complex Integration of Cognitive Functions

Solutions to tackle all the previous issues led to the development of a complex cognitive system for the humanoid robot Cog⁴, as shown in Figure 1, to acquire categorical information about actions, scenes, objects, people and the robot itself. This paper concentrates on a specific key portion of such framework - object recognition - which is of paramount importance for the overall system performance.

2. Object Detection/Recognition

Object Segmentation: A vast set of human-robot interactive techniques⁵ were developed to introduce the humanoid robot Cog knowledge concerning the visual

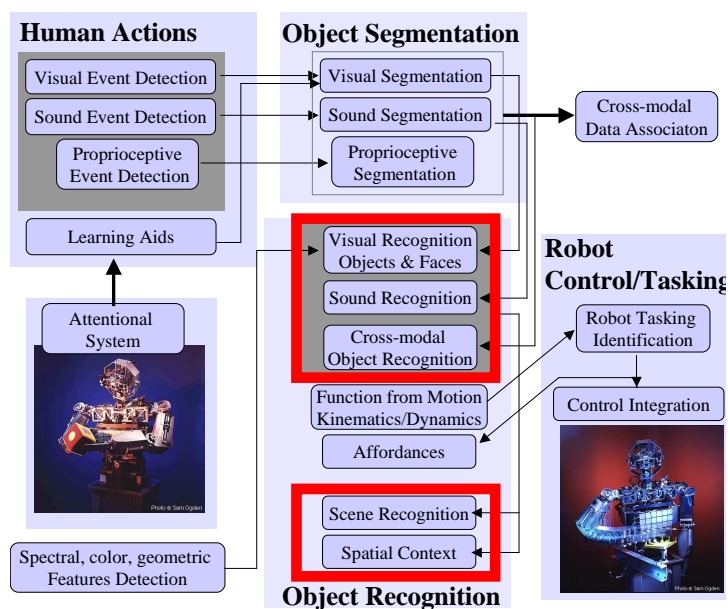


Fig. 1. Four main modules for Cog's cognitive system. It includes learning about objects from human demonstration. This is followed by object recognition that updates knowledge concerning an object. The robot then executes tasks according to its embedded knowledge of the object, to extract percepts *per se*. This paper concentrates on key issues for object recognition, presenting the modules enclosed by the large squares.

appearance of objects from human (or the robot itself) actions. Percepts segmented as object templates from the robot's surrounding world are then converted into an useful format through a template matching based object recognition scheme, which enables the robot to recognize object templates under different perspective views.

2.1. Template Matching – Color Histograms

The object recognition algorithm needs to cluster object templates by classes according to their identity. Such task was implemented through color histograms – objects are classified based on the relative distribution of their color pixels. Since object's masks are available, external global features do not affect recognition, and hence color histograms are appropriate. A multi-target tracking algorithm⁴ keeps track of object locations as the visual percepts change due to movement of the robot's active head. Ideally, a human actor should expose the robot to several views of the object being tracked (if the object appearance is view-dependent), in order to link them to the same object.

Recognition works as follows. Quantization of each of the three color channels originates 8^3 groups G_i of similar colors. The number of image pixels n_{G_i} indexed to a group is stored as a percentage of the total number of pixels. The first 20 color histograms of an object category are saved into memory and updated

Recog. objects	Errors %	Recog. objects	Errors %
Big sofa	4.35 (23)	Green chair	0.0 (4)
Small sofa	18.2 (11)	Door	3.57 (28)
Table	5.56 (18)	Blue door	14.29 (7)
Black chair	0.0 (18)	Total	5.5 (109)

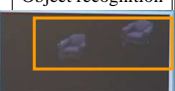



Object recognition	Object segmentation
	
	
Cog's wide field of view	Cog's foveal view • human showing an object

Table 1. (left) Recognition errors. It is shown the number of matches evaluated from a total of 11 scenes (objects are segmented and recognized more than once per scene). Incorrect matches occurred due to color similarity among big/small sofas or between different objects. Missed matches result from drastic variations in light sources (right) sofa is segmented and recognized.

thereafter. New object templates are classified according to their similarity with other object templates previously recognized for all object categories, by computing $p = \sum_{i=1}^{8^3} \text{minimum}(n_{G_i}, n_{G'_i})$. If $p < 0.7$ for all of the 20 histograms in an object category, then the object does not belong to that category. If this happens for all categories, then it is a new object. If $p \geq 0.7$, then a match occurs, and the object is assigned to the category with maximum p .

Whenever an object is recognized into a given category, the average color histogram which originated a better match is updated. Given an average histogram which is the result of averaging m color histograms, the updating consists of computing the weighted average between this histogram (weight m) and the new color histograms (unit weight). This has the advantage that color histograms evolve as more samples are obtained to represent different views of an object.

Experimental Results: Table 1 presents quantitative performance statistics for this algorithm (it was also applied for assisting the construction of 3D maps of scenes⁶). It shows the system running on the humanoid robot Cog, while recognizing previously learned objects. Incorrect matches occurred due to color similarity among different objects (such as a big and a small sofa). Errors arising from labelling an object in the database as a new object are chiefly due to drastic variations in light sources. Qualitative results from an on-line experiment of several minutes for object segmentation, tracking and recognition of new objects on the humanoid robot are shown in Figures 2 and 3.

Out of around 100 samples from on-line experiments, recognition accuracy average was of 95%. The recognition accuracy depends however on the object being recognized. For instance, several experiments shown the algorithm not capable to differentiate among different people's faces, although it differentiated correctly between faces and other objects. This demonstrates the need for an independent algorithm for face recognition, which is described hereafter.

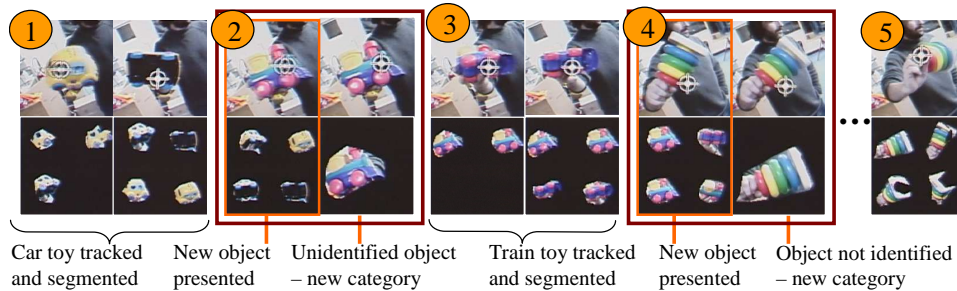


Fig. 2. Sequence from an on-line tracking and recognition experiment of several minutes on the humanoid robot Cog. (1) The robot is tracking a toy car (top row), and new template instances of it are being inserted into a database. A random set of templates from this database is shown on the bottom row. (2) A new object (a toy train) is presented. It was never seen before, so it is not recognized and a new category is created for it. (3) The toy train is tracked. (4) A new, unknown object presented, for which a new category is created on the object recognition database. (5) Templates from the new object are stored.

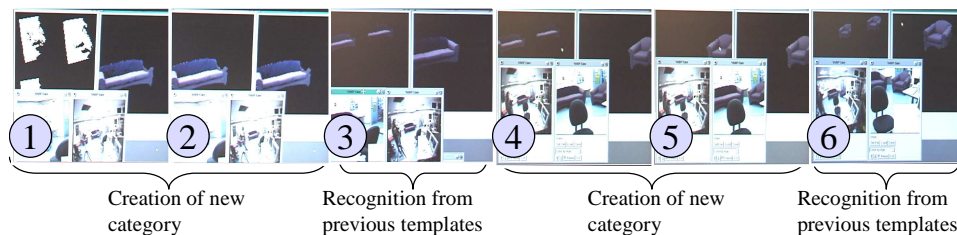


Fig. 3. Sequence from an on-line experiment of several minutes on the humanoid robot Cog. (1) The robot detects and segments a new object – a sofa; (2) New object is correctly assigned to a new category; (3) Object, not being tracked, is recognized from previous templates (as shown by the two sofa templates mapped to it); (4-5-6) Same sequence for a smaller sofa.

3. Face Detection/Recognition

Humans are especially good for recognizing faces, being such skill rather robust, considering the large variability of face features due to viewing conditions, poses, emotional expressions, and visual distractions such as haircut variability or glasses, among others. Faces play a major social role to convey identity and emotion, and also to extract information concerning others intentions and living habits.

The problem of face identification is organized as shown in Figure 4, which also shows the similar organization used to recognize objects. Training data is automatically generated by detecting and tracking multiple faces and objects over time⁴. Batches of faces from one person (or batches of object's templates) are then inserted into the database. If more than 60% of this batch matches an individual (or object, respectively), it is recognized as such. Otherwise, a new entry is created on the

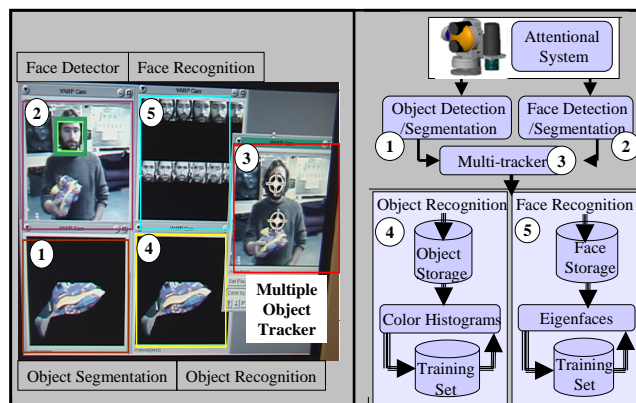


Fig. 4. Approach for segmenting and recognizing faces and objects. Training data for object/face recognition is extracted by keeping objects and others faces in memory for a while, generating this way a collection of training samples consisting of multiple segmentations of objects and faces. (left) on-line experiment on Cog (right) schematic organization. 1) Object segmentation 2) Face detection and segmentation 3) Multiple object tracking 4) Object Recognition 5) Face Recognition.

database which corresponds to a new person (or a new object), and the learning algorithm is updated. Indeed, people often have available a few seconds of visual information concerning the visual appearance of other people before having to take a recognition decision, which is the motivation behind this approach.

3.1. Face Detection:

Faces in cluttered scenes are located by a computationally efficient algorithm (developed by Paul Viola's group at MIT), which is applied to each video frame (acquired by a foveal camera). If a face is detected, the algorithm estimates a window containing that face, as shown in Figure 4.

3.2. Face Recognition – Eigenfaces

Appearance-based approaches project face images into a linear subspace of reduced dimensions⁷. In order to efficiently describe a collection of face images, this subspace is determined using Principal Component Analysis (PCA) on a set of training images. The corresponding eigenvectors are denominated eigenfaces⁷, because they are face-like in appearance (see Figure 5). Eigenfeatures, such as eigeneyes or eigenmouth for the detection of facial features⁸, is an alternative variant for the eigenfaces method. In feature-based approaches, geometric face features, such as eyebrow's thickness or invariant moments, are extracted to represent a face⁹. However, feature extraction poses serious problems for such techniques⁹.

Let the training set of M face images from a person n be $\{\phi_1, \phi_2, \dots, \phi_M\}$ (see Figure 5). The average face image of this set is defined by $\psi = 1/M \sum_{i=1}^M \phi_i$. The covariance matrix for the set of training faces is thus given by (1):

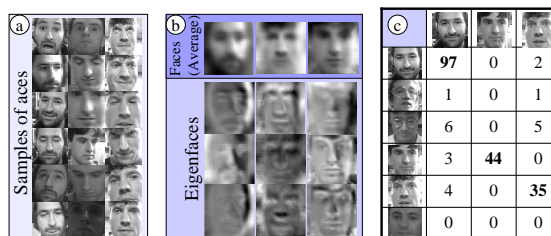


Fig. 5. a) Face image samples are shown for each of three people out of a database of six; b) Average face image for three people on the database, together with three eigenfaces for each one; c) Confusion table with face recognition results.

$$C_{\phi_n} = \frac{1}{M} \sum_{i=1}^M \Gamma_i \Gamma_i^T = AA^T \quad (1)$$

being $\Gamma_n = \phi_n - \psi$ the difference of each image from the mean, and $A = [\Gamma_1, \Gamma_2, \dots, \Gamma_M]$. Cropped faces are first re-scaled to 128×128 images (size $S = 128^2$). Determining the eigenvectors and eigenvalues of the S^2 size covariance matrix C is untractable. However, C rank does not exceed $M - 1$. For $M < S^2$ there are only $M - 1$ eigenvectors associated to non-zero eigenvalues, rather than S^2 . Let v_i be the eigenvectors of the $M \times M$ matrix $A^T A$. The eigenfaces μ_i are given by:

$$\mu_i = \sum_{k=1}^M v_{ik} \Gamma_k, i = 1, \dots, M \quad (2)$$

The number of basis functions is further reduced from M to M' by selecting only the most meaningful M' eigenvectors (with largest associated eigenvalues) and ignoring all the others. Classification of a face image ϕ consists of projecting it into eigenface components, by correlating the eigenvectors with it, for obtaining the coefficients $w_i = \mu_i(\phi - \psi), i = 1, \dots, M'$ of this projection. The weights w_i form a vector $\Omega = \{w_1, w_2, \dots, w_{M'}\}$. A face is then classified by selecting the minimum L_2 distance to each object's coefficients in the database $\varepsilon_\phi = \|\Omega - \Omega_k\|$ where Ω_k describes the k^{th} face class in the database. If ε_ϕ is below a threshold, then it corresponds to a new face.

Experimental Results: The training data set contains a lot of variation (see Figure fig-eigenfaces for a few demonstrative samples). Validation data corresponds to a random 20% of all the data. The confusion Table in Figure 5 presents results for recognizing three different people, being the average recognition accuracy of 88.9%.

4. Scene Recognition

Given the image of an object, its meaning is often a function of the surrounding context. Context cues are useful to remove such ambiguity. Ideally, contextual fea-

8 *Artur M. Arsenio*

tures should incorporate the functional constraints faced by people, objects or even scenes (eg. people cannot fly and offices have doors). Hence, functionality plays a more important role than more ambiguous and variable features (such as color, which selection might depend on human preferences). As such, texture properties seem appropriate as contextual features. Although environmental textures are also often human-dependent, global features such as door placement, desks and shelves location, wall division or furniture geometry usually follow a predetermined pattern which presents low variability. Therefore, in order to incorporate such global constraints, features will be averaged to a low resolution spatial configuration.

Wavelets¹⁰ were selected as contextual features. Processing is applied iteratively through the low frequency branch of the transform over $T = 5$ scales, while higher frequencies along the vertical, horizontal and diagonal orientations are stored (due to signal polarity, this corresponds to a compact representation of six orientations in three images). The input is thus represented by $v(x, y) = v(\vec{p}) = \{v_k(x, y), k = 1, \dots, N\}$, with $N=3T=15$. Each wavelet component at the i^{th} level has dimensions $256/2^i \times 256/2^i$, and is down-sampled to a 8×8 image:

$$\bar{v}(x, y) = \sum_{i,j} v(i, j)h(i - x, j - y) \quad (3)$$

where $h(x,y)$ is a Gaussian window. Thus, $\bar{v}(x, y)$ has dimension 960. Similarly to other approaches¹¹, the dimensionality problem is reduced to become tractable by applying Principal Component Analysis (PCA). The image features $\bar{v}(\vec{p})$ are decomposed into the basis functions given by the PCA, encoding the main spectral characteristics of a scene with a coarse description of its spatial structure:

$$v(\vec{p}) = \sum_{i=1}^D c_i \varphi_k^i(\vec{p}) \quad , \quad c_i = \sum_{\vec{p},k} v_k(\vec{p}) \varphi_k^i(\vec{p}) \quad (4)$$

where the functions $\varphi_k^i(\vec{p})$ are the eigenfunctions of the covariance operator given by $v_k(\vec{p})$. These functions incorporate both spatial and spectral information. The decomposition coefficients are obtained by projecting the image features $v_k(\vec{p})$ into the principal components c_i , used hereafter as input context features.

Lets now split image measurements \vec{v}_{IO} into two sets $v_{IO} = (v_I, v_O) = (v_{B_{\vec{p},\epsilon}}, v_{\bar{B}_{\vec{p},\epsilon}})$, where \bar{B} is the complementary of B . If it is assumed that in the presence of an object, intrinsic object features ($\vec{v}_I \in B$) and context features ($\vec{v}_O \in \bar{B}$) are independent, then $P(\vec{v}_{IO}|\vec{x}, o_n) = p(\vec{v}_I|\vec{x}, o_n)p(\vec{v}_O|\vec{x}, o_n)$. These two conditional PDFs obtained refer to different sets of features¹¹:

Local information – $p(\vec{v}_I|\vec{x}, o_n)$ Corresponds to the object/face recognition schemes based on local features described in Sections 2 and 3. Assumes $p(o_n|v_{IO}) \simeq p(o_n|\vec{v}_I)$, so that relevant features are inside the neighborhood B – local features which belong to the object and not to the background.

Contextual information – $p(\vec{v}_O|\vec{x}, o_n)$ Represents the object conditional PDF given a set of contextual features, which provides priors on the object presence, location, depth, size and orientation. Assumes $p(o_n|v_{IO}) \simeq p(o_n|\vec{v}_O)$

Neglecting \vec{v}_I , the vector $\vec{c} = \vec{v}_{IO} = \{c_i, i = 1, \dots, D\}$ denotes the resulting D-dimensional input vector, with $D = E_m, 2 \leq D \leq Th_o$, where m denotes a class, Th_o an upper threshold and E_m denotes the number of eigenvalues within 5% of the maximum eigenvalue. These features can be viewed as a scene's holistic¹² representation since all the regions of the image contribute to all the coefficients, as objects are not encoded individually. The effect of neglecting \vec{v}_I is reduced by mapping the foveal camera (which grabs data for the object recognition scheme based on local features) into the image from the peripheral view camera, where the weight of the local features \vec{v}_I is strongly attenuated. The vector \vec{p} is thus given in wide field of view retinal coordinates.

A collection of images is automatically annotated by the robot⁶ and used as training data. Mixture models are applied to find interesting places to put a bounded number of local kernels that can model large neighborhoods. In D-dimensions a mixture model is denoted by density factorization over multivariate Gaussians (spherical Gaussians were selected for faster processing times), for each object class n :

$$p(\vec{c}|o_n) = \sum_{m=1}^M b_m p(\vec{c}|o_n, g_m), \quad p(\vec{c}|o_n, g_m) = G(\vec{c}, \vec{\mu}_{m,n}, C_{m,n}) = \frac{e^{-1/2(\vec{c}-\vec{\mu}_m)C_m^{-1}(\vec{c}-\vec{\mu}_m)}}{(2\pi)^{D/2}|C_m|^{1/2}}$$

where $|\cdot|^{1/2}$ is the square root of the determinant, g_m refers to the m^{th} Gaussian with mean $\vec{\mu}_m$ and covariance matrix C_m , M is the number of Gaussian clusters, and $b_m = p(g_m)$ are the weights of the local models. The estimation of the parameters will follow the EM algorithm¹³:

E-step for k -iteration From the observed data \vec{c} , this step computes the a-posteriori probabilities $e_{m,n}^k(l)$ of the clusters:

$$e_{m,n}^k(l) = p(c_{m,n}|\vec{c}) = \frac{b_{m,n}^k G(\vec{c}, \vec{\mu}_{m,n}^k, C_{m,n}^k)}{\sum_{m=1}^M b_{m,n}^k G(\vec{c}, \vec{\mu}_{m,n}^k, C_{m,n}^k)} \quad (5)$$

M-step for k -iteration : cluster parameters are estimated according to the maximization of the join likelihood of the L training data samples:

$$b_{m,n}^{k+1} = 1/L \sum_{l=1}^L e_{m,n}^k(l) \quad (6)$$

$$\vec{\mu}_{m,n}^{k+1} = \langle \vec{c} \rangle_m = \frac{\sum_{l=1}^L e_{m,n}^k(l) \vec{c}_l}{\sum_{l=1}^L e_{m,n}^k(l)} \quad (7)$$

$$C_{m,n}^{k+1} = \frac{\sum_{l=1}^L e_{m,n}^k(l) (\vec{c}_l - \vec{\mu}_{m,n}^{k+1})(\vec{c}_l - \vec{\mu}_{m,n}^{k+1})^T}{\sum_{l=1}^L e_{m,n}^k(l)} \quad (8)$$

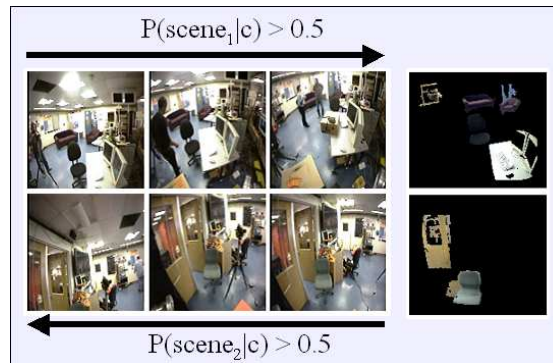


Fig. 6. Test images (wide field of view) organized with respect to $p(o_n|\vec{c})$. Top row: $o_n = scene_1$, $p(scene_1|\vec{c}) > 0.5$; Bottom row: $o_n = scene_2$, $p(scene_2|\vec{c}) > 0.5$. Scene descriptions shown in the right column are built on-line, automatically⁶.

All vectors are column vectors and $\langle \rangle_m$ in (7) represents the weighted average with respect to the posterior probabilities of cluster m . The EM algorithm converges as soon as the cost gradient is small enough or a maximum number of iterations is reached. The probability density function (PDF) for an object n is then given by Bayes' rule $p(o_n|\vec{c}) = p(\vec{c}|o_n)p(o_n)/p(\vec{c})$, where $p(\vec{c}) = p(\vec{c}|o_n)p(o_n) + p(\vec{c}|\neg o_n)p(\neg o_n)$. The same method applies for the out-of-class PDF $p(\vec{c}|\neg o_n)$ which represent the statistical feature distribution for the input data in which o_n is not present.

Finally, it is necessary to select the number M of gaussian clusters. This number can be selected as the one that maximizes the join likelihood of the data. An agglomerative clustering approach based on the Rissanen Minimum Description Length (MDL) order identification criterion¹⁴ was implemented to automatically estimate M . Figure 6 shows results for classifying two different scenes. Each time a human presents a scene object to the robot, both foveal and wide field of view images are saved and automatically annotated to the corresponding scene⁶.

Contextual cues are not only useful for scene or object classification, but also for the selection of an object's attentional focus, scale or orientation in an image.

5. Object/People Recognition from Contextual Features

Objects in the world are situated, in the sense that they usually appear in specific places. Children are pretty good at learning the relative probability distribution of objects in a scene – for instance, books are often found on top of shelves. The scene context puts a very important constraint on the type of places in which a certain object might be found.

The same way we wish to be able to determine relations among objects (e.g., chairs are most probable in front of desks), it is also extremely useful to extract relations among people as well as in between people and objects. For instance, people

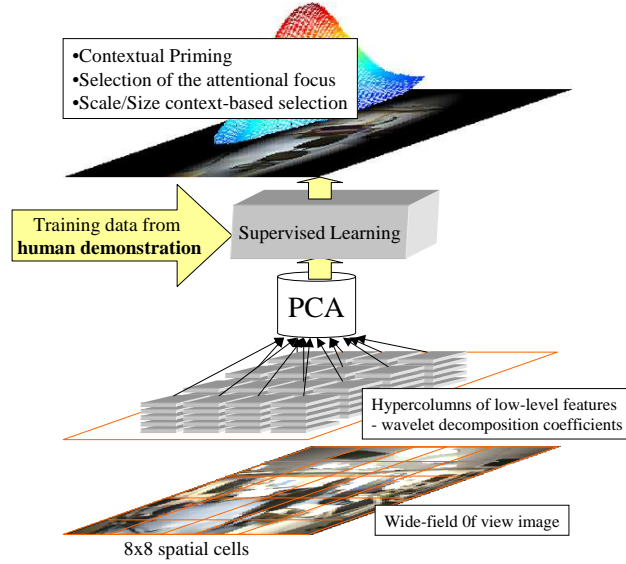


Fig. 7. Algorithmic structure for learning to locate objects and people from contextual and human cues. Training data is acquired from the object/face segmentation and recognition modules. Temporal constancy of such data is maintained by the multiple object tracking algorithm.

usually sit on desks, and therefore might be expected to appear on places on top of chairs or in front of tables, as well as in walking places such as corridors, but not in a ceiling! From a humanoid point of view, contextual selection of the attentional focus is very important both to constrain the search space for identifying or locating objects (optimizes computational resources) and also to determine common places on a scene to drop or store objects such as tools or toys.

A model for the contextual control of the attentional focus (location and orientation), scale selection and depth inference is hereafter presented which, contrary to other approaches¹¹, does not neglect dependency among the input state dimensions. The output space is defined by the 6-dimensional vector $\vec{x} = (\vec{p}, d, \vec{s}, \phi)$, where \vec{p} is a 2-dimensional position vector, d is the object's depth⁶, $\vec{s} = (w, h)$ is a vector containing the principal components of the ellipse that models the 2D size retinal size of the object, and ϕ is the orientation of such ellipse. Given the context \vec{c} , one needs to evaluate the PDF $p(\vec{x}|o_n, \vec{c})$ from a mixture of (spherical) Gaussians¹³,

$$p(\vec{x}, \vec{c}|o_n) = \sum_{m=1}^M b_{m,n} G(\vec{x}, \vec{\eta}_{m,n}, X_{m,n}) G(\vec{c}, \vec{\mu}_{m,n}, C_{m,n}) \quad (9)$$

The mean of the new Gaussian $G(\vec{x}, \vec{\eta}_{m,n}, X_{m,n})$ is now a function: $\vec{\eta} = f(\vec{c}, \beta_{m,n})$, that depends on \vec{c} and on a set of parameters $\beta_{m,n}$. A locally affine model was chosen for f , with $\{\beta_{m,n} = (\vec{a}_{m,n}, A_{i,n}): \vec{\eta}_{m,n} = \vec{a}_{m,n} + A^T \vec{c}\}$. The learning equations become now¹³:

E-step for k -iteration From the observed data \vec{c} and \vec{x} , this step computes the a-posteriori probabilities $e_{m,n}^k(l) = p(c_{m,n} | \vec{c}, \vec{x})$ of the clusters:

$$e_{m,n}^k(l) = \frac{b_{m,n}^k G(\vec{x}, \vec{\eta}_{m,n}^k, X_{m,n}^k) G(\vec{c}, \vec{\mu}_{m,n}^k, C_{m,n}^k)}{\sum_{m=1}^M b_{m,n}^k G(\vec{x}, \vec{\eta}_{m,n}^k, X_{m,n}^k) G(\vec{c}, \vec{\mu}_{m,n}^k, C_{m,n}^k)}$$

M-step for k -iteration : cluster parameters are estimated according to (where m indexes the M clusters, and l indexes the number L of samples):

$$C_{m,n}^{k+1} = \langle (\vec{c} - \vec{\mu}_{m,n}^{k+1})(\vec{c} - \vec{\mu}_{m,n}^{k+1})^T \rangle_m \quad (10)$$

$$A_{m,n}^{k+1} = (C_{m,n}^{k+1})^{-1} \langle (\vec{x} - \vec{\eta}_{m,n}^{k+1})(\vec{x} - \vec{\eta}_{m,n}^{k+1})^T \rangle_m \quad (11)$$

$$a_{m,n}^{k+1} = \langle (\vec{x} - A_{m,n}^{k+1} \vec{c})(\vec{x} - A_{m,n}^{k+1} \vec{c})^T \rangle_m \quad (12)$$

$$X_{m,n}^{k+1} = \langle (\vec{x} - a_{m,n}^{k+1} - (A_{m,n}^{k+1})^T \vec{c})(\vec{x} - a_{m,n}^{k+1} - (A_{m,n}^{k+1})^T \vec{c})^T \rangle_m$$

The parameters $b_{m,n}^{k+1}$ and means $\mu_{m,n}^{k+1}$ are estimated as before. The conditional probability follows then from the joint PDF of the presence of an object o_n , at the spatial location p , with pose ϕ , size \vec{s} and depth d , given a set of contextual image measurements \vec{c}

$$p(\vec{x}|o_n, \vec{c}) = \frac{\sum_{m=1}^M b_{m,n}^k G(\vec{x}, \vec{\eta}_{m,n}^k, X_{m,n}^k) G(\vec{c}, \vec{\mu}_{m,n}^k, C_{m,n}^k)}{\sum_{m=1}^M b_{m,n}^k G(\vec{c}, \vec{\mu}_{m,n}^k, C_{m,n}^k)}$$

Object detection and recognition requires the evaluation of this PDF at different locations in the parameter space. The mixture of gaussians is used to learn spatial distributions of objects from the spatial distribution of frequencies in an image.

Results and Discussion: Figure 8 presents results for selection of the attentional focus for objects from the low-level cues given by the distribution of frequencies computed by wavelet decomposition. Some furniture objects were not moved (such as the sofas), while others were moved in different degrees: the chair appeared in several positions during the experiment, and so thus the chair's templates centroids, while the table and door suffered mildly displacements. Still, errors on the head gazing control added considerable location variability whenever a non-movable object was segmented and annotated. It demonstrates that, given an holistic characterization of a scene (by PCA on the image wavelet decomposition coefficients), one can estimate the appropriate places whether objects often appear, such as a chair in front of a table, even if no chair is visible at the time – which also informs that regions in front of tables are good candidates to place a chair. Object occlusions by people are not relevant, since local features are neglected, favoring contextual ones.



Fig. 8. Localizing and recognizing objects from contextual cues (top) Samples of scene images are shown on the first column. The next five columns show probable locations based on context for finding a door, the smaller sofa, the bigger sofa, the table and the chair, respectively. Even if the object is not visible or present, the system estimates the places at which there is a high probability of finding such object. Two such examples are shown for the chair. Occlusion by humans do not change significantly the context. (bottom) Results in another day, with different lightning.

6. Auditory Perception: Acoustic Segmentation and Recognition

A human caregiver introduces a robot to a rich world of visual information concerning objects' visual appearance and shape. But there are cases for which visual recognition is not appropriate (for instance, objects might not be visible to the robot). It is of paramount importance to refer to perception over other perceptual modalities in such situations.

We are interested in detecting conditions that repeat with a rate consistent with what a human can easily produce and perceive (e.g., waving a flag or clapping). We will consider anything above 10Hz to be too fast (e.g., the vibration of a violin string), and anything below 0.1Hz to be too slow (e.g., the daily rise and fall of the sun). Such a restriction is related to the idea of natural kinds¹⁵, where perception is based on the physical dimensions and practical interests of the observer.

6.1. Acoustic Segmentation

Acoustic signals have a rich structure around and above the kHz range, for which the Fourier transform and related transforms are very useful. But not so much for detecting gross repetition around the single Hz range. The sound generated by a moving object can be quite complicated, since any constraints due to inertia or continuity are much weaker than for the physical trajectory of a mass moving through space. This led us to the development of a robust periodicity detector:

Period estimation – The term *period* refers strictly to event-scale repetition (Hz range). For every sample of the signal, we determine how long it takes for the signal to return to the same value from the same direction (increasing or decreasing), if it ever does. For this comparison, signal values are quantizing adaptively into discrete ranges. Intervals are computed in one pass using a look-up table that, as we scan through the signal, stores the time of the last occurrence of a value/direction pair. The next step is to find the most common interval using a histogram (which requires quantization of interval values), giving us an initial estimate $p_{estimate}$ for the event period.

Clustering – Drift and variability in the period is explicitly taken into account as follows. We cluster samples in rising and falling intervals of the signal, using that estimate to limit the width of our clusters but not to constrain the distance between clusters. This is a good match with real signals we see that are generated from human action, where the periodicity is rarely very precise. Clustering is performed individually for each of the quantized ranges and directions (increasing or decreasing), and then combined afterwards. Starting from the first signal sample not assigned to a cluster, our algorithm runs iteratively until all samples are assigned, creating new clusters as necessary. A sample extracted at time t is assigned to a cluster with center c_i if $\|c_i - t\|_2 < p_{estimate}/2$. The cluster center is the average time coordinate of the samples assigned to it, weighted according to their values.

Merging – Clusters from different quantized ranges and directions are merged into a single cluster if $\|c_i - c_j\|_2 < p_{estimate}/2$.

Segmentation – We find the average interval between neighboring cluster centers for positive and negative derivatives, and break the signal into discrete periods based on these centers. Notice that we do not rely on an assumption of a *constant* period for segmenting the signal into repeating units.

The output of this entire process is an estimate of the average period of the signal, a segmentation of the signal into repeating units, and a confidence value that reflects how periodic the signal really is. The period estimation process is applied at multiple temporal scales. This sound segmentation algorithm generates incrementally training data for a sound recognition scheme: sound models of objects are first learned from experimental human/robot manipulation, enabling their a-posteriori identification with or without the agents actuation.

6.2. Sound Recognition

The repetitive nature of the sound generated by an object under periodic motion can be analyzed to extract an acoustic ‘signature’ for that object. We search for repetition in a set of frequency bands independently, then collect those frequency bands whose energies oscillate together with a similar period. Specifically, the acoustic signature for an object is obtained by applying the following steps:

- (1) Detection of the period of repetition for each frequency band (Section 6.1).
- (2) A *period histogram* is constructed to accumulate votes for frequency bands having the same estimated period (or half the period – it is common to have sounds that occur once per repetition, for example at one endpoint of the trajectory, or twice per repetition, for example at two instants of maximum velocity). The histogram is smoothed by adding votes for each bin of the histogram to their immediate neighbors as well.
- (3) The maximum entry in the period histogram is selected as the *reference* period. All frequency bands corresponding to this maximum are collected and their responses over the reference period are stored in a database of acoustic signatures. Since the same objects can be shaken or waved at different velocities, it is important to normalize temporal information relative to the reference period.

A collection of annotated acoustic signatures for each object are used as input data (see Figure 9) for a sound recognition algorithm by applying the eigenobjects method, which is also used for face recognition. A sound image is represented as a linear combination of base sound signatures (or *eigensounds*). Only eigensounds corresponding to the three highest eigenvalues – which represent a large portion of the sound’s energy – are retained. Classification consists of projecting novel sounds to this space, determining the coefficients of this projection, computing the L_2 distance to each object’s coefficients in the database, and selecting the class corresponding to the minimum distance.

Cross-modal information aids the acquisition and learning of unimodal percepts and consequent categorization in a child’s early infancy. Similarly, visual data is employed here to guide the annotation of auditory data to implement a sound recognition algorithm. Training samples for the sound recognition algorithm are classified into different categories by the visual object recognition system or from information from the visual object tracking system. This enables the system, after training, to classify the sounds of objects not visible.

Experimental Results: The system was evaluated quantitatively by random selection of 10% of the segmented data for validation, and the remaining data for training. This process was randomly repeated three times. It is worth noticing that even samples received within a short time of each other often do not look too similar, due to background acoustic noise, noise on the segmentation process, other objects’ sounds during experiments, and variability on how objects are moved and

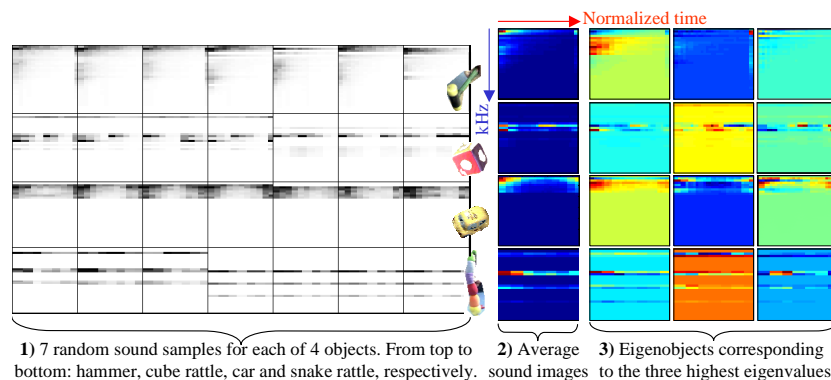


Fig. 9. Sound segmentation and recognition. Acoustic signatures for four objects are shown along the rows. (1) Seven sound segmentation samples are shown for each object, from a total of 28 (car), 49 (cube rattle), 23 (snake rattle) and 34 (hammer) samples. (2) Average acoustic signature for each object. The vertical axis corresponds to the frequency bands and the horizontal axis to time normalized by the period. (3) Eigensounds corresponding to the three highest eigenvalues.

presented to the robot. For example, the car object is heard both alone and with a rattle (either visible or hidden).

The recognition rate for the three runs averaged to 82% (86.7%, 80% and 80%). Recognition rates by object category were: 67% for the car, 91.7% for the cube rattle, 77.8% for the snake rattle and 83.3% for the hammer. Most errors arise from mismatches between car and hammer sounds. Such errors could be avoided by extending our sound recognition method to use derived features such as the onset/decay rate of a sound, which is clearly distinct for the car and the hammer (the latter generates sounds with abrupt rises of energy and exponential decays, while sound energy from the toy car is much smoother). Instead, we will show in the following section that these differences can be captured by cross-modal features to correctly classify these objects.

7. Cross-Modal Object Segmentation/Recognition

Different objects have distinct acoustic-visual patterns which are a rich source of information for object recognition, if we can recover them. The relationship between object motion and the sound generated varies in an object-specific way. A hammer causes sound after striking an object. A toy truck causes sound while moving rapidly with wheels spinning; it is quiet when changing direction. These statements are truly cross-modal in nature. Features extracted from the visual and acoustic segmentations are what is needed to build an object recognition system. Each type of features are important for recognition when the other is absent. But when both are present, then we can do even better by looking at the relationship between the visual motion of an object and the sound it generates. Is there a loud bang at an extreme of the physical trajectory? If so we might be looking at a hammer. Are the bangs soft relative to the visual trajectory? Perhaps it is a bell. Such relational

features can only be defined and factored into recognition if we can relate or *bind* visual and acoustic signals.

7.1. *Binding*

Due to physical constraints, the set of sounds that can be generated by manipulating an object is often quite small. For toys which are suited to one specific kind of manipulation - as rattles encourage shaking - there is even more structure to the sound they generate¹⁶. When sound is produced through motion for such objects the audio signal is highly correlated both with the motion of the object and its identity. The spatial trajectory can be applied to associate the visual appearance of objects manipulated by humans or the robot itself to the sound they produce.

A binding algorithm was developed to associate cross-modal, locally periodic signals, by which we mean signals that have locally consistent periodicity, but may experience global drift and variation in that rhythm over time. The detection of periodic cross-modal signals over an interval of seconds as described in the previous section is a necessary (but not sufficient) condition for a binding. The extra constraints that must be met for binding to occur are now described. Signals are compared by matching the cluster centers determined as in the previous section. Each peak within a cluster from the visual signal is associated to a temporally close (within a maximum distance of half a visual period) peak from the acoustic signal. Binding occurs if the visual period matches the acoustic one, or if it matches half the acoustic period, within a tolerance of 60ms.

7.2. *Recognition*

The feature space for cross-modal recognition consists therefore of:

- ▷ Sound/Visual period ratios – the sound energy of a hammer peaks once per visual period, while the sound energy of a car peaks twice (for forward and backward movement).
- ▷ Visual/Sound peak energy ratios – the hammer upon impact creates high peaks of sound energy relative to the amplitude of the visual trajectory. Although such measure depends on the distance of the object to the robot, the energy of both acoustic and visual signals will generally decrease with depth (the sound energy disperses through the air and the visual trajectory reduces in apparent scale).

Human or robot actions are therefore used to create associations along different sensor modalities, and objects can be recognized from the characteristics of such associations. Our original approach can differentiate objects from both their visual and acoustic backgrounds by finding pixels and frequency bands (respectively) that are oscillating together. This is accomplished through dynamic programming, applied to match the sound energy to the visual trajectory signal. Formally, let $S = (S_1, \dots, S_n)$ and $V = (V_1, \dots, V_m)$ be sequences of sound and visual trajectory energies segmented from n and m periods of the sound and visual trajectory

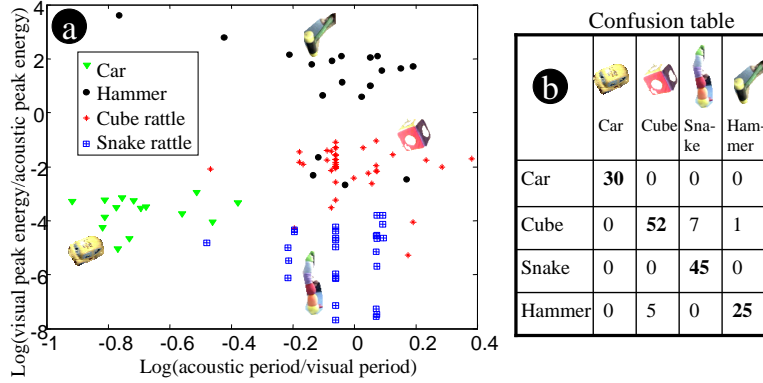


Fig. 10. Object recognition from cross-modal clues. The feature space consists of period and peak energy ratios. The confusion matrix for a four-class recognition experiment is shown. The period ratio is enough to separate the cluster of the car object from all the others. Similarly, the snake rattle is very distinct, since it requires large visual trajectories for producing soft sounds. Errors for categorizing a hammer originated exclusively from erroneous matches with the cube rattle, because hammering is characterized by high energy ratios, and very soft bangs are hard to identify correctly. The cube rattle generates higher energy ratios than the snake rattle. False cube recognitions resulted mostly from samples with low energy ratios being mistaken for the snake.

signals, respectively. Due to noise, n may be different to m . If the estimated sound period is half the visual one, then V corresponds to energies segmented with $2m$ half periods (given by the distance between maximum and minimum peaks). A matching path $P = (P_1, \dots, P_l)$ defines an alignment between S and M , where $\max(m, n) \leq l \leq m + n - 1$, and $P_k = (i, j)$, a match k between sound cluster j and visual cluster i . The matching constraints are imposed by:

The boundary conditions are $P_1 = (1, 1)$ and $P_l = (m, n)$.

Temporal continuity satisfies $P_{k+1} \in \{(i + 1, j + 1), (i + 1, j), (i, j + 1)\}$. This restricts steps to adjacent elements of P .

The function cost $c_{i,j}$ is given by the square difference between V_i and S_j periods. The best matching path W can be found efficiently using dynamic programming, by incrementally building an $m \times n$ table caching the optimum cost at each table cell, together with the link corresponding to that optimum. The binding W will then result by tracing back through these links, as in the Viterbi algorithm.

Experimental Results: Figure 10 shows cross-modal features for a set of four objects. It would be hard to cluster automatically such data into groups for classification. But as in the sound recognition algorithm, training data is automatically annotated by visual recognition and tracking. After training, objects can be categorized from cross-modal cues alone. The system was evaluated by selecting randomly 10% of the data for validation, and the remaining data for training. This process was randomly repeated 15 times. The recognition rate averaged over all these runs were, by object category: 86.7% for the cube rattle, 100% for both the car and the

snake rattle, and 83% for the hammer. The overall recognition rate was 92.1%. Such results demonstrate the potential for recognition using cross-modal cues.

7.3. Self-Recognition

Lets turn now to the robot's perception of its own body¹⁶. Cog treats proprioceptive feedback from its joints as just another sensory modality in which periodic events may occur. These events can be bound to the visual appearance of its moving body part (if it is visible) and the sound that the part makes, if any (in fact Cog's arms are quite noisy, making an audible "whirr – whirr" when they move back and forth). Proprioceptive feedback provides very useful reference signals to identify appearances of the robot's body in different modalities. That is why the binding algorithm is extended to include proprioceptive data. As shown in Figure 11, the binding algorithm was used not only to identify the robot's own acoustic rhythms, but also to identify visually the robot's mirror image (an important milestone towards development of a child's theory of mind¹⁷).

8. Conclusions

Although the potential for expanding this work is vast, from a practical perspective complex levels of functionality have already been accomplished. Lets consider again Figures 4 and 11. Figure 11 shows a partial snapshot of the robot's state during one experiment. The robot's experience of an event is rich, with many visual and acoustic segmentations generated as the event continues, relevant prior segmentations recalled using object recognition, the relationship between data from different senses detected and stored, and objects tracked to be further used by statistical learning processes for object location from contextual features.

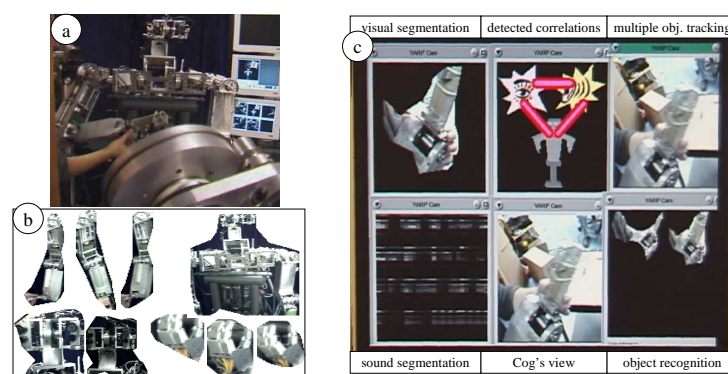


Fig. 11. Self-recognition. a) Like a child interested in his own reflection, the robot needs to integrate visual cues from a mirror with proprioceptive cues from its body for self-recognition; b) Visual templates associated to the robot's body; c) On-line recognition experiment on the robot Cog. It shows both object visual and sound segmentation, visual object tracking and recognition, and linking of acoustic and visual percepts to the robot's body.

Objects have complex uses, come in various colors, sound differently and appear in the world in ways constrained by the surrounding scene structure. Since these object properties are multi-modal, multiple sensorial input have to be processed. Each individual object property is important for recognition, but if more than one communication channel is available, such as the visual and auditory channels, both sources of information can be correlated for extracting richer pieces of information, as exemplified by the cross-modal object recognition algorithm. This new approach enabled us to disambiguate object identity in cases where one perceptual modality alone could fail. We demonstrated in this paper how to take advantage of multiple perceptual information – the whole is truly greater than the sum of the parts.

Acknowledgements

Project funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement. Author supported by Portuguese grant PRAXIS XXI BD/15851/98.

References

1. Z. Zhang, O. Faugeras, 3D Dynamic Scene Analysis (Springer-Verlag, 1992).
2. Z. Duric, J. Fayman, E. Rivlin, Recognizing functionality, in Proc. International Symposium on Computer Vision (1995) .
3. K. E. Adolph, M. A. Eppler, E. J. Gibson, Development of perception of affordances, *Advances in Infancy Research* **8** (1993) 51–98.
4. A. M. Arsenio, Cognitive-Developmental Learning for a Humanoid Robot: A Caregivers' gift, Ph.D. thesis, MIT, May/June 2004.
5. A. M. Arsenio, An embodied approach to perceptual grouping, in IEEE CVPR Workshop on Perceptual Organization in Computer Vision (2004) .
6. A. M. Arsenio, Map building from human-computer interactions, in IEEE CVPR Workshop on Real-time Vision for Human Computer Interaction (2004) .
7. M. Turk, A. Pentland, Eigenfaces for recognition, *J. Cognitive Neuroscience* (1991).
8. P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE PAMI* **19** (1997) (7) 711–720.
9. R. Chellappa, C. Wilson, S. Sirohey, Human and machine recognition of faces: A survey, in Proceedings of the IEEE (1995), vol. 83 705–740.
10. G. Strang, T. Nguyen, Wavelets and Filter Banks (Wellesley-Cambridge Press, 1996).
11. A. Torralba, Contextual priming for object detection, *Int. J. Computer Vision* (2003).
12. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, *International Journal of Computer Vision* (2001) 145–175.
13. N. Gershenfeld, The nature of mathematical modeling (Cambridge univ. press, 1999).
14. J. Rissanen, A universal prior for integers and estimation by minimum description length, *Annals of Statistics* **11** (1983) 417–431.
15. H. Hendriks-Jansen, *Catching Ourselves in the Act* (MIT Press, Cambridge, 1996).
16. P. Fitzpatrick, A. Arsenio, Feel the beat: using cross-modal rhythm to integrate robot perception (International Workshop on Epigenetic Robotics, 2004).
17. S. Baron-Cohen, *Mindblindness* (MIT Press, 1995).