The Helping Hand in Humanoid Robot Learning

Artur M. Arsenio
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA
Email: arsenio@csail.mit.edu

Abstract— **Human caregivers play an important role during child's development phases. A human tutor often modifies a task context so that information is easily perceived and learned by the child. We propose to use the same strategy to teach a humanoid robot. Contrary to standard supervised learning techniques relying on a-priori availability of training data obtained manually, actions by an embodied agent (the human) are used to automatically generate training data for the learning mechanisms, so that the robot develops categorization autonomously.**

The work presented in this paper follows a developmental approach to perception and learning. This framework based on human-robot interactive communication is demonstrated to apply naturally to a large spectrum of computer vision problems: object segmentation, visual and cross-modal object recognition, object depth extraction and localization from monocular contextual cues, and learning from visual aids – such as books.

I. Introduction

Embodied and situated perception [3] consists of boosting the vision capabilities of an artificial creature by fully exploiting the opportunities created by an embodied agent situated in the world [2]. Proponents for Active vision [1], [8], contrary to passive vision, argue for the active control of the visual perception mechanism so that perception is facilitated. Percepts can indeed be acquired in a purposive way by the active control of a camera [1]. This approach has been successfully applied to several computer vision problems, such as stereo vision - by dynamically changing the baseline distance between the cameras or by active focus selection [14].

We argue for solving a visual problem by not only actively controlling the perceptual mechanism, but also and foremost actively changing the environment through experimental manipulation [3], [12]. The human body plays an essential role in such a framework, being applied not only to facilitate perception, but also to change the world context so that it is easily understood by the robotic creature (the humanoid robot Cog used throughout this work is shown in Figure 1).

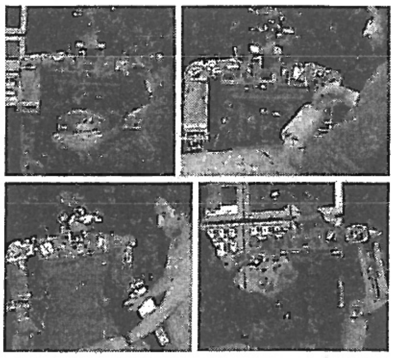


Fig. 1. The experimental platform. The humanoid robot Cog is equipped with cameras in an active vision head, a microphone array across the torso and two robotic arms. Some typical learning scenarios (from left to right, top to bottom) a human shows a book to Cog; a human describes the shape of an object to the robot; a repetitive action (hammering) is demonstrated to the robot; a human waves a yellow car to create a salient stimulis on Cog's attentional system.

Although a human can help the robot to extract meaningful percepts from the world, it should be emphasized that such help should not include constraining the world structure in anyway, such as the removal of environment clutering or careful luminosity setup, among others, since both children and robots exist in real, not virtual, worlds. Instead, the focus should be placed on communicating information to the robot which boosts its perceptual skills, helping *him* to filter out irrelevant information. Indeed, while teaching a toddler, parents do not remove the room's furniture or buy extra lights to just show the child a book. Help instead is given by facilitating the child's task of stimulus selection (for example, by pointing or tapping into a book's image).

This paper presents a human-centered approach to facilitate the robot's perception and learning, while showing the benefits that result from introducing humans in the robot's learning loop. This work aims at teaching humanoid robots as children, being the child's mother role attributed to a human tutor. With that in mind, next section will present software tools developed to enable human-robot interactions during important learning activities for children: playing with toys, tools, books and drawings. An approach for learning the structure of the robot's surrounding world is presented in Section III. Such structure is inferred from cues introduced by humans. Finally, conclusions are drawn in Section IV, together with a discussion on ongoing work.

II. Human-Robot Playing Activities

A. Books

During developmental phases, children's learning is often aided by the use of audiovisuals and especially, books. Humans often paint, draw or just read books to children during the early months of childhood. Books are indeed a very useful tool to teach robots different object representations or to communicate properties of unknown objects to them (such as a whale's visual appearance). We present a human aided object segmentation algorithm [7] to extract the visual appearance of objects from the background (mainly the book pages), which is illustrated in Figure 2:

1) A standard color segmentation algorithm is applied to a stationary image (stationary over a sequence of consecutive frames)
2) A human actor waves an arm on top of the object to be segmented
3) The motion of skin-tone pixels is tracked over a time interval (using the Lucas-Kanade Pyramidal algorithm), and the energy per frequency content is determined for each point's trajectory
4) Periodic, skin-tone points are grouped together into the arm mask [3].
5) The trajectory of the arm's endpoint describes an algebraic variety over N^2. The target object's template is then given by the union of all bounded subsets (the color regions of the stationary image) which intersect this variety

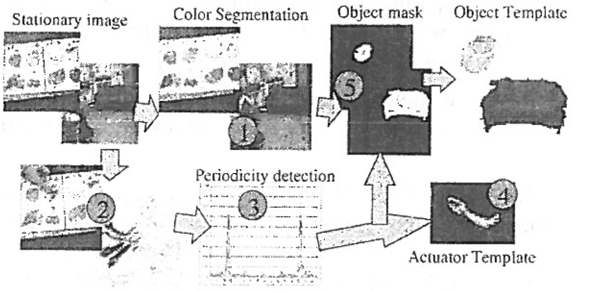
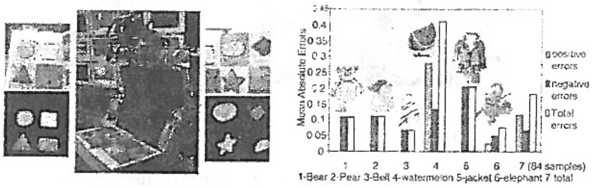


Fig. 2. A standard color segmentation algorithm computes a compact cover for the image. The actuator's periodic trajectory is used to extract the object's compact cover – a collection of color cluster sets.

The algorithm consists of grouping together the colors that form an object. This grouping works by having periodic trajectory points being used as seed pixels. The algorithm fills the regions of the color segmented image whose pixel values are closer to the seed pixel values, using a 8-connectivity strategy. Therefore, points taken from waving are used to both select and group a set of segmented regions into the full object. Clusters grouped by a single trajectory might either form or not form the smallest compact cover which includes the full object (depending on intersecting or not all the clusters that form the object). After two or more trajectories this problem vanishes.

Periodic detection [7] is applied at multiple scales. Indeed, for an arm oscillating during a short period of time, the movement might not appear periodic at a coarser scale, but appear as such at a finer scale. If a strong periodicity is not found at a larger scale, the window size is halved and the procedure is repeated again for each half.

This strategy, which enables the robot to learn from books, relies heavily in human-robot interactions. It is essential to have a human in the learning loop to introduce objects from a book to the robot (as a human caregiver does to a child), by tapping on their book's representations. This scheme was successfully applied to extract templates for fruits, geometric shapes and other elements from books, under varying light conditions – statistical results are presented in Figure 3, while Figure 4 shows a collection of segmentation samples.



a)

b)

Fig. 3. a) The humanoid robot looking at a book. Segmentation of geometric shapes from the book are also shown. b) Statistical analysis for object segmentation from books. Errors are given by (template area - object's real visual appearance area)/(real area). Positive/negative errors stand solely for templates with larger/smaller area than the real area, respectively. The real area values were determined manually.

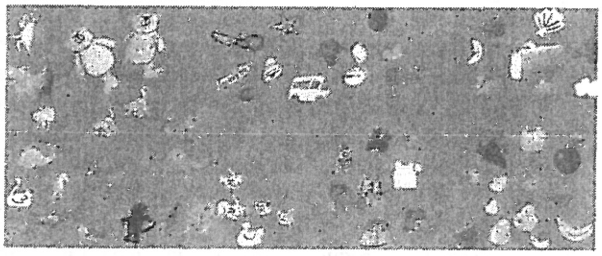


Fig. 4. Templates for several categories of objects (for which a representative sample is shown), were extracted from dozens of books. Two subjects not acquainted with the algorithm were also briefly instructed on the protocol for interacting with the robot. No noticeable performance degradation was found from such interactions.

B. *Toys and Draw Sketches*

Object representations acquired from a book are inserted into a database, so that they become available for future recognition tasks. However, object descriptions may came in different formats - drawings, paintings, photos, etc. Hence, methods were developed to establish the link between an object representation in a book and *real* objects recognized from the surrounding world using the object recognition technique described in [7], as shown by Figure 5. Except for a description contained in a book, the robot had no other knowledge concerning the visual appearance or shape of such object.

Additional possibilities include linking different object descriptions in a book, such as a drawing (also shown in Figure 5). Other feasible descriptions to which this framework is being applied include paintings, prints, photos and computer generated objects.

C. *Playing with Tools and Toys*

Children extract meaningful percepts by playing with tools and toys with a human caregiver. The latter can show the child (or a robot) how to play a hammer and the *bang* sound that it makes upon impact. Due to physical constraints, the set of sounds that can be generated by

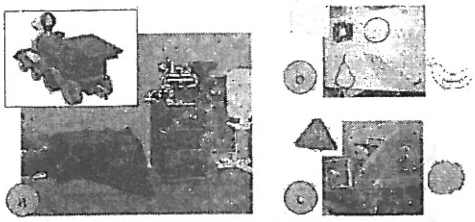


Fig. 5. a) Object recognition and location. The train appears under a perspective transformation in a bedroom scene - generated by the DataBecker Software. Estimated lines are also shown. Scene lines matched to the object are outlined. b) Recognition of geometric, manual drawings of a triangle and a circle from the description of objects learned using books. c) Geometric shapes of a square and a banana recognized using the descriptions from a book.

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 generate [11]. When sound is produced through motion for such objects the audio signal is highly correlated both with the motion of the object and the tools' identity. Therefore, the spatial trajectory can be applied to extract visual and audio features – patches of pixels, and sound frequency bands – that are associated with the object (see Figure 6), which enables the robot to map the visual appearance of objects manipulated by humans or itself to the sound they produce. Cross-modal integration from robot's multiple sensing modalities is described in detail in [5], [11].

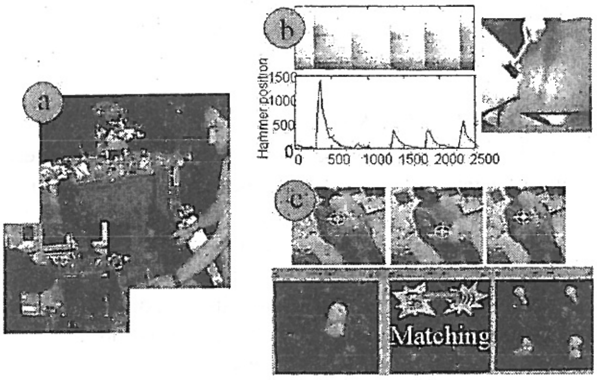


Fig. 6. a) A human playing with a hammer, which bangs in a table, producing a distinctive audio signal. b) A human moves a hammer repetitively producing sound upon impact, which is matched to the visual trajectory. c) top: tracking an oscillatory instrument; down: image of object segmentation and display of a detected visual/sound matching.

D. Educational Activities: Painting, Drawing

A common pattern of early human-child interactive communication is through activities that stimulate the child's brain, such as drawing or painting. Children are able to extract information from such activities while they are being performed on-line. This capability motivated the implementation of three parallel processes which receive input data from different sources: from an attentional tracker [12], which tracks the attentional focus and is attracted to a new salient stimulus; and from an algorithm that selectively attends to the human actuator for the extraction of periodic signals from the trajectory of oscillating skin blobs [6].

Whenever a repetitive trajectory is detected from any of these parallel processes, it is partitioned into a collection of trajectories, being each element of such collection described by the trajectory points between two zero velocity points with equal sign on a neighborhood (similarly to the partitioning process described in [11]). An object recognition algorithm previously described in [7] is then applied to extract correlations between these sensorial signals perceived from the world and geometric shapes present in such world, or on the robot object database (see Figure 7), as follows:

1) Each partition of the repetitive trajectory is mapped into a set of oriented lines by application of the Hough transform.
2) By applying a recognition scheme [7], trajectory lines are matched to oriented edge lines (from a Canny detector) on
   a) a stationary background,
   b) objects stored in the robot's object recognition database.

This way, the robot learns object properties not only through cross-modal data correlations, but also by correlating human gestures and information stored in the world structure (such as objects with a geometric shape) or on its own database.

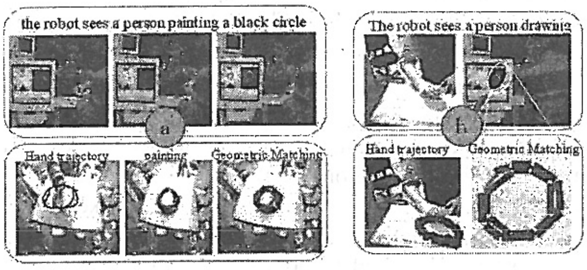

the robot sees a person painting a black circle
The robot sees a person drawing
a
b
Hand trajectory
painting
Geometric Matching
Hand trajectory
Geometric Matching

Fig. 7. a) A human is painting a black circle on a sheet of paper with a ink can. The circle is painted multiple times. The top row images show the activity being performed. The first image from the left on the bottom row shows the hand trajectory. Edge lines of the background image – middle image – are matched to such trajectory – last image. b) A human draws a circle on a sheet of paper with a pen, which is matched into a circle drawn and recognized previously (see Figure 5-b).

III. Learning about the World

Autonomous agents, such as robots and humans, are situated in a dynamic world [9], full of information stored on its own structure. For instance, the probability of a chair being located in front of a table is much bigger than that of being located on the ceiling. A robot should place an object where it can easily find it - if one places a book on the fridge, she will hardly find it later!

Therefore, a statistical framework was developed to capture such knowledge stored in the world. This framework consists of: learning 3D scenes from cues provided by a human actor; and learning the spatial configuration of objects within a scene.

A. Building Scene Descriptions

The human arm structure relative to a scene structure provides a natural way for constraining the object detection problem using *global* information (see Figure 8). The environment surrounding the robot provides also additional structure that can be learned through supervised learning techniques.

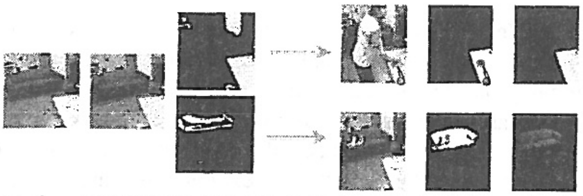


Fig. 8. Segmentation of heavy, stationary objects. A human teacher *shows* the table and sofa to the robot, by waving on the objects' surface, so that the robot can then use the arm trajectory to link the objects to the correct color regions.

A significant amount of contextual information may be extracted from a periodically moving actuator – most often such motions are from interactions with objects of interest – which can be framed as the problem of estimating $p(o_n|v_{B_{\vec{p},\epsilon}}, act^{per}_{\vec{p},S})$, the probability of finding object n given a set of local, stationary features v on a neighborhood ball B of radius ϵ centered on location p, and a periodic actuator on such neighborhood with trajectory points in the set $S \subseteq B$. The algorithm previously described to learn information from books also offers a solution for this problem. Segmentations for several furniture items on a scene, together with statistical results for such objects, are shown in Figure 9.

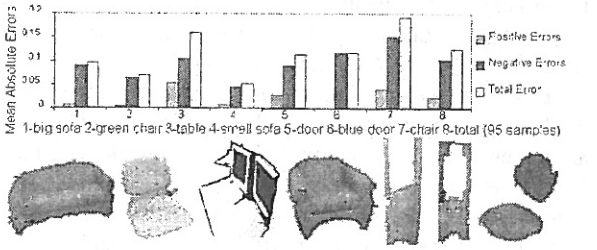


Fig. 9. Statistics for the furniture items (a set of segmentation samples is also shown). A chair is grouped from two disconnect regions by merging temporally and spatially close segmentations.

1) 3D Environment Maps: Besides binocular cues, the human visual system also processes monocular data for depth inference, such as focus, perspective effects, among others. Previous attempts have been made on exploring scene context for depth inference [17]. However, these passive techniques make use of contextual clues already present on the scene. They do not actively change the context of the scene through manipulation to improve the robot's perception. We propose an active, embodied approach that actively changes the context of a scene, extracting monocular depth measures. The human arm diameter (which is assumed to remain approximately constant for the same depth, except for degenerate cases) is used as a reference for extracting relative depth information. This measure is extracted from periodic signals of a human hand as follows:

1) Detection of skin-tone pixels over a image sequence
2) A blob detector labels these pixels into regions
3) These regions are tracked over the image sequence, and all non-periodic blobs are filtered out
4) A region filling algorithm (8-connectivity) extracts a mask for the arm
5) A color histogram is built for the background image. Points in the arm's mask having a large frequency on such histogram are labelled as background.
6) The smallest eigenvalue of the arm's mask gives an approximate measure of a fraction of the arm radius.

Once a reference measure is available, coarse depth information can be extracted relative to the arm diameter, for each arm trajectory's point. A plane is then fitted (in the least square sense) to this 3D data. A scene is defined by the uncertain geometric configuration of all the objects. Figure 10 presents both coarse depth images and 3D reconstruction data for a typical scene in the robot's lab.

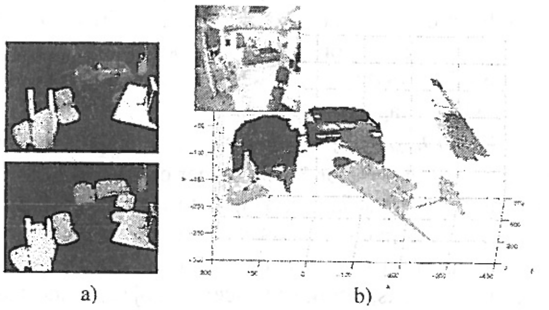


Fig. 10. An image of all the scene is built by converting retinal coordinates into egocentric coordinates (the pan and tilt viewing angle of the robot's head). a) Furniture image segmentations– on top – and depth map – bottom – for a scene in Cog's room (lighter corresponds to closer); b) Coarse 3D map built for the scene shown.

B. Context Driven Data Selection

World structural information will be exploited in an active manner. Contextual control of the attentional focus (location and orientation), scale selection and depth inference will be presented at two different categorical levels. From a humanoid point of view, contextual selection of the attentional focus is very important both to constrain the search space for locating objects (optimizes computational resources) and also to determine common places on a scene to drop or store objects such as tools or toys.

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 o_n depth. $\vec{s} = (w, h)$ 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. Mixture models are applied to find interesting places to put a bounded number of local kernels that can model large neighborhoods. Therefore, given the context $\vec{c}$, one needs to evaluate the PDF p from a mixture of (spherical) Gaussians G and G_x [13],

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

where $\vec{\mu}_{m,n}$ is the $\vec{c}$ mean and $C_{m,m}$ the covariance for cluster m and object n. The mean $\vec{\eta}_{m,n}$ of Gaussian G_x is a function that depends on $\vec{c}$ and on a set of parameters $\beta_{m,n}$. A locally affine model was chosen for the mean: $\beta_{m,n} = (\vec{a}_{m,n}, A_{i,n})$: $\vec{\eta}_{m,n} = \vec{a}_{m,n} + A^T \vec{c}$.

The EM algorithm is then used to learn the cluster parameters (see [13] for a detailed description of the EM algorithm). The EM algorithm converges as soon as the cost gradient is small enough or a maximum number of iterations is reached. The number M of gaussian clusters is selected automatically in order to maximize the join likelihood of the data, using the Minimum Description Length agglomerative clustering approach based on the Rissanen order identification criterion [15].

Contextual information will be exploited at two different categorization levels, through two complementary approaches:

- ▷ Object-centered context - which requires no visual input, operating on egocentric coordinates
- ▷ Holistic-based context - which operates on wide-field-of-view image coordinates

1) Object-based representation: This approach determines an object's vector $\vec{x}$ using other objects in a scene as contextual information. Training data consists of the data stored and automatically annotated while building scene descriptions from human cues. A scene is modelled as a collection of links, being each scene's object connected to any other object in the same scene. Each time an object is recognized or detected from human demonstration, the algorithm creates or updates connecting links. Each link from object a to object b, given x_a, is defined by the probability of finding object a at state x_a and object b with state $\vec{x}_b = (p_b, d_b, \vec{s}_b, \phi_b)$. On such approach, the contextual feature vector is $\vec{c} = x_a$, $o_n = o_a$ and $\vec{x} = \vec{x}_b$. The vector p is the object's location in the robot's head egocentric gazing coordinates (this mapping is estimated using a supervised learning technique [4]). Figure 11 shows results for selection of the attentional focus for objects given the state data $\vec{x}$ of another object. It is worth stressing that context priming prunes the set of candidate objects to match the primed object, and therefore reduces the computational resources required for object detection, which is only applied into the more likely spatial locations.

2) Holistic-based approach: Given the image of an object, its meaning is often a function of the surrounding context. Ideally, contextual features should incorporate the functional constraints faced by people, objects or even scenes (eg. people cannot fly, a hammer needs an external force to be moved and offices have doors). Therefore, functionality plays a more important role than more ambiguous and variable features (such as color, for which selection, for instance, depends on the taste of a decorator). Functionality constraints have been previously exploited for multi-modal object recognition [5] and for determining function from motion [10], just to name a few applications.

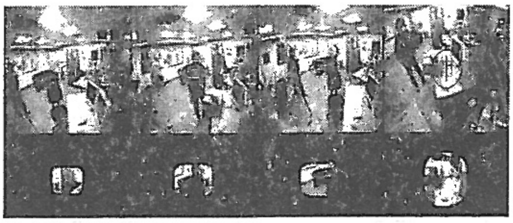


Fig. 11. Top row shows the results of saccade movements by the robot's head to find previously learned objects. It is shown the object predicted position, size and orientation. Bottom row shows image regions were the object is predicted to lie within.

As such, texture properties seem appropriate, which led to the selection of Wavelets [16] as contextual features, since they are much faster to compute than Gabor filters and provide a more compact representation. Input monochrome images are transformed using a Daubechies-4 wavelet tree, along 5 depth scales. The input is represented by $v(\vec{p}) = \{v_k(x,y), k = 1,\ldots,N\}$, with N=15. Each wavelet component is down-sampled to a 8×8 image, so that $\bar{v}(x,y)$ has dimension 960.

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 provided by the PCA, encoding the main spectral characteristics of a scene with a coarse description of its spatial arrangement. The decomposition coefficients are obtained by projecting the image features $v_k(\vec{p})$ into the principal components $\vec{c} = \{c_i, i = 1,\ldots,D\}$ ($\vec{c}$ denotes the resulting D-dimensional input vector, used thereafter as input context features). These coefficients can be viewed as a scene's holistic representation since all the regions of the image contribute to all the coefficients, as objects are not encoded individually. The effect of neglecting local features is reduced by mapping the foveal camera (which grabs data for the object recognition scheme based on local features) into the image from the wide field of view camera, where the weight of the local features is strongly attenuated.The vector $\vec{p}$ is now given in the wide field of view retinal coordinates. Figure 12 presents results for selection of the attentional focus for objects from low-level cues (wavelet decomposition) for a scene in the robot's laboratory.

There is still a lot of information that cannot be extracted from scenes familiar to a robot (real elephants are not common in humanoid research labs). But such information from the robot's outside world can be transmitted to the robot by a human tutor using books, as previously described.

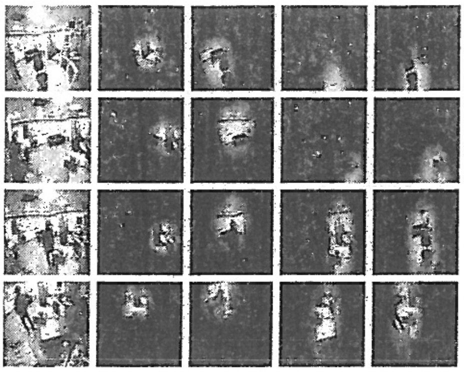


Fig. 12. Samples of scene images are shown on the first column. The next four columns show probable locations based on context for the smaller sofa, the bigger sofa, the table and the chair, respectively. Notice that, even if the object is not visible or present, the system estimates the places at which there is a high probability of finding such object. Two such examples are shown for the chair – no matter the viewing angle, chairs are predicted to appear in front of the table. It is also shown that occlusion by humans do not change significantly the global context.

IV. Conclusions

Humans were introduced in the robot's learning loop to facilitate robot perception. We presented a comprehensive set of experiments that support such claim. We demonstrated how on-line input from a human instructor during playing activities can facilitate robot's perception, using books and other learning aids, and learning activities such as drawing or painting. The implemented strategies are suitable to extract cross-modal auditory and visual information from tools (e.g., hammer), toys (such as rattles) or musical instruments.

A human in the learning loop can also introduce the robot knowledge concerning the robot's surrounding world. By actively describing objects on a scene, such as furniture items, the robot was able to segment objects and further built scene descriptions from such data. A probabilistic framework was then developed to determine probable locations of objects.

A. Ongoing and Future work

This human-centered framework is currently being applied to other research problems,

a) Shape from Human Cues: Human cues were shown useful to extract coarse depth measures, but they can be also applied to extract information concerning the object's shape (at a coarse level), such as hollow parts or object boundaries. This is achieved by having a human to describe actively, with its fingers, the object contours (possible benefits include removal of shadows).

b) Robot Localization and Map Building: This framework for building scenes from human cues is also been evaluated for simultaneously map building and mobile robot localization, using objects as natural landmarks (under large uncertainty, however).

c) Task Detection: A task can be defined as a collection of events on objects. A hybrid Markov Chain is being used to model complex tasks such as sawing, hammering, painting, drawing, among others [4].

d) Functional Object Recognition: A tool may have different uses. For instance, a knife can be use to cut (motion orthogonal to the knife's edge) or to stab (motion parallel to the knife's edge), which involves two different functions for the same object. We are processing the motion of a tool while executing a task to classify its function.

e) Recognition of Acoustic Patterns: A large corpora of sounds annotated to objects is extracted from the algorithm for cross-modal association. Such data is currently being used to train a classifier.

f) Control Integration Grounded On Perception: The integration of control strategies for both oscillatory and reaching movements should be grounded on the perception, which determines the mapping between the perceived motion of objects and how they should be manipulated.

And there are still other potential research directions to explore for which humans can really *give a hand* to help learning on an embodied and situated robotic agent.

Acknowledgements

Project funded by DARPA's "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] J.Y. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *Int. Journal on Computer Vision*, 2:333–356, 1987.

[2] M. Anderson. Embodied cognition: A field guide. *Artificial Intelligence*, pages 91–130, 2003.

[3] A. Arsenio. Embodied vision - perceiving objects from actions. *IEEE Workshop on Human-Robot Interactive Communication*, 2003.

[4] A. Arsenio. *Developmental Learning on a Humanoid Robot*. Accepted for the Int. Joint Conference on Neural Networks, 2004.

[5] 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*, Singapore, December 2003.

[6] A. M. Arsenio. Map building from human-computer interactions. In *Submitted to the CVPR Workshop on Real-time Vision from Human Computer Interactions*, 2004.

[7] A.M. Arsenio. Teaching a humanoid robot from books. In *International Symposium on Robotics*, March 2004.

[8] R.K. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, August 1988.

[9] R. A. Brooks, C. Breazeal, R. Irie, C. C. Kemp, M. Marjanović, B. Scassellati, and M. M. Williamson. Alternative essences of intelligence. In *Proceedings of the American Association of Artificial Intelligence*, pages 961–968, 1998.

[10] Z. Duric, J. Fayman, and E. Rivlin. Recognizing functionality. In *Proc. International Symposium on Computer Vision*, 1995.

[11] P. Fitzpatrick and A. Arsenio. *Feel the beat: using cross-modal rhythm to integrate robot perception*. Accepted to fourth International Workshop on Epigenetic Robotics, 2004.

[12] Paul Fitzpatrick. *From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot*. PhD thesis, MIT, Cambridge, MA, 2003.

[13] N. Gershenfeld. *The nature of mathematical modeling*. Cambridge university press, 1999.

[14] E. Krotkov, K. Henriksen, and R. Kories. Stereo ranging from verging cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):1200–1205, December 1990.

[15] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:417–431, 1983.

[16] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1996.

[17] A. Torralba and A. Oliva. *Global depth perception from familiar scene structure*. MIT AI-Memo 2001-036, CBCL Memo 213, 2001.