International Journal of Humanoid Robotics

Teaching Humanoid Robots like Children: Explorations into the World of Toys and Learning Activities

Artur M. Arsenio

Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
The Stata Center, 32 Vassar Street, Room G32-376
Cambridge, Massachusetts 02139, USA
arsenio@csail.mit.edu

The goal of this work is to develop a humanoid robot's perceptual system through the use of learning aids. We describe methods to enable learning on a humanoid robot using learning aids such as books, drawing materials, boards, educational videos or other children toys. Visual properties of objects are learned and inserted into a recognition scheme, which is then applied to acquire new object representations – we propose learning through developmental stages.

Inspired in infant development, we will also boost the robot's perceptual capabilities by having a human caregiver performing educational and play activities with the robot (such as drawing, painting or playing with a toy train on a railway). We describe original algorithms to extract meaningful percepts from such learning experiments.

Keywords: Developmental Learning; Humanoid Robotics; Hand Gesture Recognition; Cross-modal perception; Learning Aids; Educational Activities

1. Introduction

This work argues for boosting a humanoid robot's object recognition capabilities – essential for developmental learning of a robot – through the use of learning aids. Teaching a humanoid robot information concerning its surrounding world is a difficult task, which takes several years for a child, equipped with evolutionary mechanisms stored in its genes, to accomplish. Learning aids are often used by human caregivers to introduce the child to a diverse set of (in)animate objects, exposing the latter to an outside world of colors, forms, shapes and contrasts, that otherwise could not be available to a child (such as the image of a Panda). A learning aid expands the child's knowledge of its surrounding world, and it is therefore a potentially useful tool to introduce new informative percepts to a robot.

Our strategy relies heavily in human-robot interactions. It is essential to have a human in the loop to introduce objects from a book to the robot (as a human caregiver does to a child). A more complete human-robot communication interface results from adding other aiding tools to the robot's portfolio (which facilitate as well the children' learning process). Embodied vision methods will be demonstrated with the goal of simplifying visual processing. This is achieved by selectively

attending to the human actuator (*Hand* or *Finger*). Indeed, primates have specific brain areas to process the hand visual appearance [1]. Inspired by human development studies, emphasis will be placed on facilitating vision through the action of a human instructor.

We argue for enculturated humanoid robots – introducing robots into our society and treating them as us – using child development as a metaphor for developmental learning of a humanoid robot. We exploit extensively childhood learning elements such as books (a child's learning aid) and other cognitive artifacts such as drawing boards. Multi-modal object properties are learned using these tools and inserted into several recognition schemes, which are then applied to developmentally acquire new object representations. The humanoid robot therefore sees the world through the caregivers eyes.

1.1. *Motivation*

For an autonomous robot to be capable of developing and adapting to its environment, it needs to be able to learn. The field of machine learning offers many powerful algorithms, but these require training data to operate. Infant development research suggests ways to acquire such training data from simple contexts, and use this experience to bootstrap to more complex contexts. We need to identify situations that enable the robot to temporarily reach beyond its current perceptual abilities, giving the opportunity for development to occur[2,3,4]. This led us to create children-like learning scenarios for teaching a humanoid robot. These learning experiments are used for the humanoid robot Cog (see Figure 1) to learn about object's multiple visual and auditory representations from books, other learning aids, musical instruments and education activities such as drawing and painting.

A human-robot interactive approach was thus implemented to introduce the humanoid robot to new percepts stored in books, as described in Section 2. Such percepts are converted into an useful format through an object recognition scheme,which enables the robot to recognize an object in several contexts or to acquire different object representations. Learning from other useful learning aid tools is also presented in this section. These learning tools are not only useful for acquiring and recognizing visual percepts, but also to associate words' auditory patterns to objects. Words might describe physical or functional properties of objects, or an object name (or other attributes) in a language, as discussed in Section 3.

Recognition of hand gestures is useful for a robot to extract meaning from human-robot educational activities such as painting or drawing in paper or jelly boards. These learning experiments are described in Section 4 for teaching the robot. This section also presents an approach for identifying repetitive hand gestures without an off-line database of manually annotated hand gestures. A method for associating the sounds produced by rough object surfaces to the corresponding visual textures is presented in Section 5. Finally, Section 6 draws the conclusions.

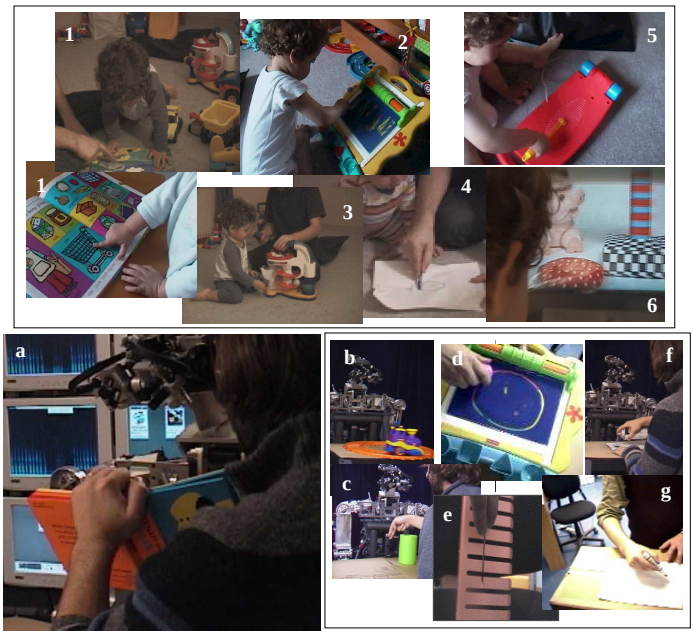

Fig. 1. Teaching humanoid robots as if they were children, exploring the large plethora of educational tool and activities widely available to human educators. Caregivers have available a rich set of tools to help a child to develop, such as (1) books; (2) drawing boards; (3) Toys or tools (e.g., hammer); (4) drawing material (e.g. pens or chalk pencils); (5) musical instruments; and (6) TV educational videos or programs, just to name a few. All these learning scenarios, and more, were implemented on the humanoid robot Cog, as shown: (a) an interactive reading scenario between the robot and a human *tutor*; (b) robot learning functional constraints between objects by observing a train moving on a railway track; (c) a human describes by gestures the boundaries of an object; (d) robot recognizing geometric shapes drawn by a human in a drawing board; (e) robot learns about rough textures from the sound they produce; (f) and (g) robot learns from educational activities played with a human, who paints with a ink can or draws with a pen, respectively.

2. Robot Skill Augmentation through Cognitive Artifacts

Children's learning is often aided by the use of audiovisuals, and especially books, from social interactions with their mother or caregiver during the developmental sub-phases of re-approximation and individual consolidation, and afterwards. Indeed, humans often paint, draw or just read books to children during their childhood. Books are indeed a useful tool to teach robots different object representations and to communicate properties of unknown objects to them.

Learning aids are also often used by human caregivers to introduce the child to a diverse set of (in)animate objects, exposing the latter to an outside world of colors, forms, shapes and contrasts, that otherwise might not be available to a child (such as images of whales and cows). Since these learning aids help to expand the child's knowledge of the world, they are a potentially useful tool for introducing new informative percepts to a robot.

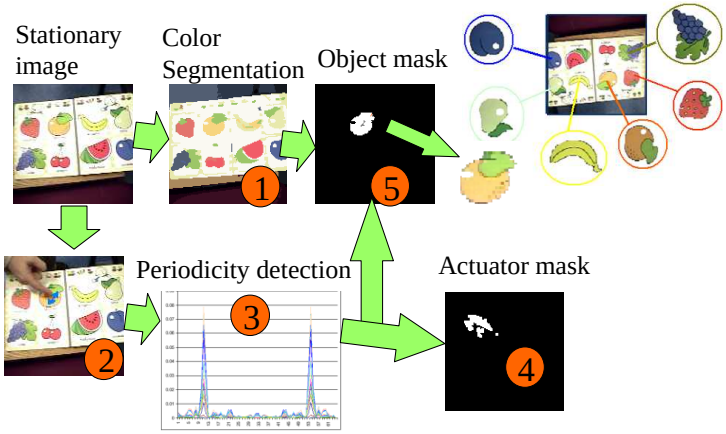

Fig. 2. A standard color segmentation algorithm computes a compact cover for the image. The actuator's periodic trajectory is used to extract the object's compact cover – a collection of color cluster sets.

2.1. *Teaching a Humanoid Robot from Books*

The author recently introduced a human aided perceptual grouping algorithm [5] for extracting informative percepts from books from a periodically moving actuator. This paper extends the original work with cardboard books, by presenting a statistical evaluation of the algorithm for a large set of samples. Experimental results are also introduced for human-robot interactions with more general books, including fabric and foam books.

The perceptual grouping problem can be framed as the estimation of $p(o_n|v_{B_{\vec{p},\epsilon}}, act^{per}_{\vec{p},S})$, the probability of extracting object o_n from a book 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 following algorithm implements the estimation process to solve this figure-ground segregation problem (see Figure 2):

(1) A standard color segmentation [5] algorithm is applied to a stationary image (stationary over a sequence of consecutive frames)
(2) A human actor waves on top of the object to be segmented
(3) The motion of skin-tone pixels is tracked over a time interval (using a pyramidal implementation of the Lucas-Kanade algorithm), and the energy per frequency content is determined for each point's trajectory
(4) Periodic, skin-tone points are grouped together into the finger mask [2].
(5) The trajectory of the arm's endpoint describes an algebraic variety [6] over N^2 (N represents the set of natural numbers). The target object's template is given by the union of all bounded subsets (the color regions of the stationary image) which intersect this variety.

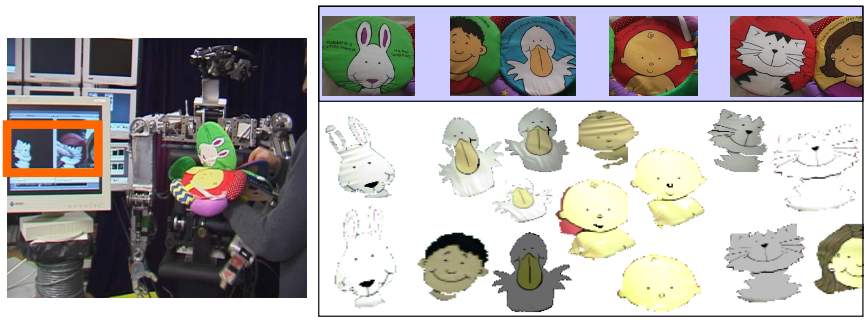

Fig. 3. (left) Teaching the visual appearance of objects to a robot, by having a human showing a fabric book to the robot, as if it was an infant. (right) Illustration shows segmentation results – on the bottom – of book pages (on top)-from a book made of fabric. Multiple segmentations were acquired from all the book pages. It is worth stressing that segmentation on books made of fabric textile poses additional difficulties, since pages deform easily, creating perspective deformations, shadows and object occlusions.

Periodic detection is applied at multiple scales. A movement might not appear periodic over a long time interval, but may 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. Periodicity is estimated from a periodogram determined for all signals from the energy of the Short-Time Fourier Transforms (STFT) over the spectrum of frequencies. These periodograms are processed by a collection of narrow bandwidth band-pass filters. Periodicity is found if, compared to the maximum filter output, all remaining outputs are negligible.

The algorithm consists of grouping together the colors that form an object 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. Points taken from waving are therefore used to both select and group a set of segmented regions into the full object.

Experimental Results

Figure 3 shows qualitative experimental results for a human-robot interaction in which the human introduces visual percepts to the robot from a fabric book, by tapping on relevant figures. These books correspond to the hardest cases to segment, since the fabric material is deformable, often making the visual image appear distorted, occluded or with large portions of shadowy regions.

It is essential to have a human in the 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. Since this learning aid helps to expand the child's knowledge of the world, it is a useful tool for introducing new informative percepts to a robot (whales do not appear often on research labs!). To corroborate this argument, this human-centered segmentation algorithm was successfully applied to extract tem-

Fig. 4. Figure/ground segregation from cardboard books (books targeted for infants and toddlers). Templates for several categories of objects (for which a representative sample is shown), were extracted from dozens of books. No noticeable performance degradation was found from subjects not acquainted with the algorithm interacting with the robot. (top) Clusters of colors were grouped together into an elephant, a piece of wood and a monkey. (middle) A bear and a cow are segmented from a book. These two animals are the union of a large set of local color clusters. (bottom) Segmentation of several elements from books.

plates for a large sample of animate and inanimate objects, as shown in Figure 4 for cardboard books. Segmentations were obtained for a large spectrum of categories: mammals (including a whale), reptiles, birds, dinosaurs, plants, clothes, fruits, geometric shapes, utilities, furniture, musical instruments, among others.

Typical errors might arise from objects with similar color patterns as the book's background page, for which no perfect differentiation is possible, since the intersection of the object's compact cover of color regions with the book's page cover (complementary to the object) is not empty (see Figure 5). High color variability within an object create additional grouping difficulties (the object's compact cover of color clusters contains too many sets – harder to group). Another experiment is presented in Figure 6, which shows results for segmenting visual descriptions from a foam book, under varying lighting conditions. The algorithm depends to

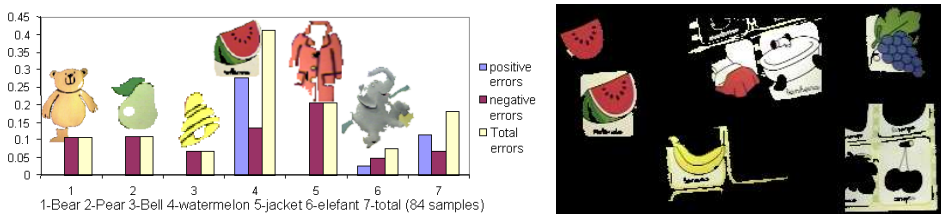

Fig. 5. (left) Statistical analysis for segmentation errors from books. Errors are given by (template area - object's real visual appearance area)/(real area). Positive errors stand solely for templates with larger area than the real area, while negative errors stand for the inverse. Total errors stand for both errors. The real area values were determined manually. (right) A representative set of templates illustrating sources of errors. The watermelon, banana and bed have color clusters with identical color – white – to its background, for which no differentiation is possible, since the intersection of the object's compact cover of color regions with the background is not empty. The cherries reflect another source of errors - very small images of objects are hard to segment.

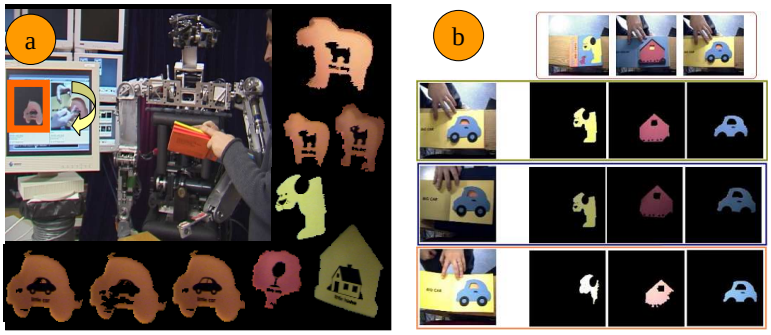

Fig. 6. a) Human shows elements from a foam book to the robot, and all objects within it were segmented (yellow dog on cover, dog, car, tree and house in inside pages). Foam parts of the book might be removed, resulting therefore a twofold of objects segmentations b) Variability with light sources. Top images show pages of a book. The other rows show segmentation results for three different luminosity conditions (from top to bottom – middle, low and high luminosity).

some extent on these variations, since different locations for light sources project shades differently. In addition, they may also add or remove color clusters depending on light sources as well as on the book material. This happens because lighting variability affects the performance of the standard color segmentation algorithm.

2.2. *Recognition of Geometric Patterns*

An object recognition technique previously developed [5,2] establishes the geometric link between an object's shape representation in a book and the *real* objects' shape recognized from the surrounding world. Geometric descriptors are obtained from local features, consisting of pairs of oriented lines, as follows:

(1) Contours are detected using a Canny detector

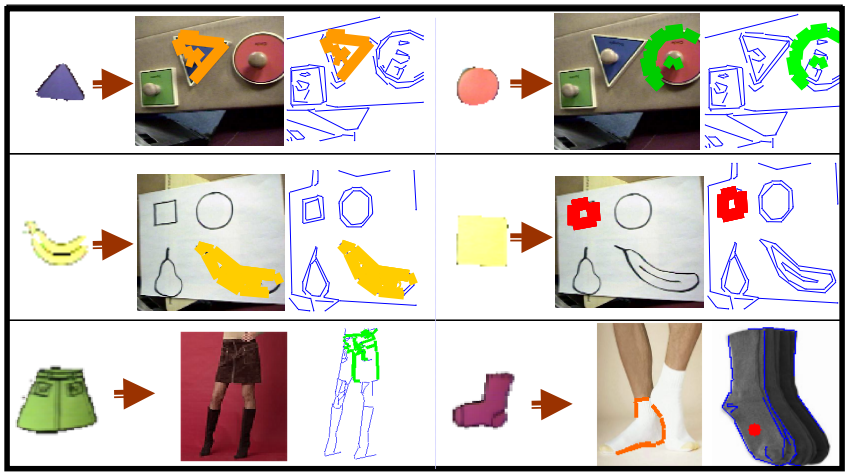

Fig. 7. (Top) Geometric shapes recognized using the descriptions from a book triangle–left– and circle–right. The recognition from chrominance features is trivial – objects have a single, identical color (Middle) Recognition of geometric, manual drawings from the description of objects learned using books (Bottom) Object descriptions extracted from books are used to recognize the geometric shapes of pictures of objects in catalogues.

(2) A Hough transform is applied to fit lines into the contour image
(3) A Sobel mask is applied at each contour point to extract the phase of the image gradient for that point
(4) All such phase measurements lying on a line are averaged to compute its phase.
(5) Two similarity invariants are used for recognition: (1) the angle between two lines and (2) the ratio of the two lines' length
(6) A coherence test requires all matches to lie within a small neighborhood.

Geometric hashing is a rather useful technique for high-speed recognition performance. Invariants are computed from training data in model images, and then stored in hash tables. Recognition consists of accessing and counting the contents of hash buckets. An Adaptive Hash table [5,2] (a hash table with variable-size buckets) algorithm was implemented to store affine color, luminance and shape invariants (which are view-independent for small perspective deformations).

Matching Representations: Drawings, Paintings, Pictures ...

Object descriptions may came in different formats - drawings, paintings, photos, etc. Hence, the link between an object representation in a book and *real* objects recognized from the surrounding world is established using the object recognition technique just briefly described, as shown by Figure 7. Except for a description contained in a book, which was previously segmented, 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,

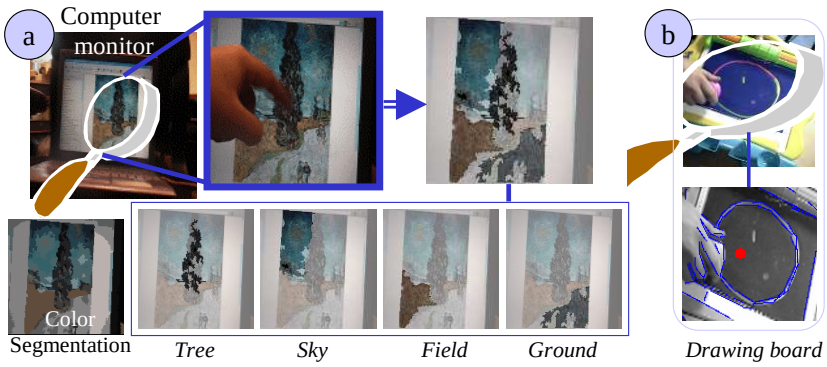

Fig. 8. a) Segmenting elements from a painting by Vincent van Gogh, *Road with Cypress and Star, 1890*. The image is displayed on a computer, from an internet website. A human then taps on several elements to segment them. The bottom row show the segmented elements individually segmented (darker on images): (from left to right) color segmentation of stationary image, and segmentations of the tree, sky, field and ground. b) Matching representations on a drawing board.

such as a drawing, as demonstrated also by results presented in Figure 7. A sketch of an object contains salient features concerning its shape, and therefore there are advantages to learning, and linking, these different representations. This framework is also a useful tool for linking other object descriptions in a book, such as a photo, a painting, or a printing (Figure 7).

2.3. *On the Use of Other Learning Aids*

A plethora of educational tools are used by educators to teach children helping them to develop. Examples of such tools are toys (such as drawing boards), educational TV programs or educational videos. The *Baby Einstein* collection includes videos to introduce infants and toddlers to colors, music, literature and art. Famous painters and their artistic creations are displayed to children on the *Baby Van Gogh* video, from the mentioned collection. This inspired the design of an experiment in which Cog is introduced to art using an artificial display (the computer monitor). The image of a painting by Vincent Van Gogh, *Road with Cypress and Star, 1890* is displayed on a computer screen. Painting are contextually different than pictures or photos, since the painter style changes the elements on the figure considerably. Van Gogh, a post-impressionist, painted with an aggressive use of brush strokes, as can be seen in his painting in Figure 8. But individual painting elements can still be grouped together by having a human actor tapping on their representation in the computer screen to group them together. Figure 8 shows results of this experiment for several individual elements: a cypress tree, the sky (with a star), a field and the road. Drawing boards are also very useful to design geometric shapes while interacting with a child, for which Figure 8 shows a circle drawn being matched to a circle shape previously learned.

3. Learning First Words

A human caregiver can introduce a robot to a rich world of visual information concerning objects' visual appearance and shape. But cognitive artifacts, which enhance perception, can also be applied to improve perception over other perceptual modalities, such as auditory processing.

We exploit repetition – rhythmic motion, repeated sounds – to achieve segmentation and recognition across multiple senses. We are interested in detecting conditions that repeat with some roughly constant rate, where that rate is consistent with what a human can easily produce and perceive. This is not a very well defined range, but we will consider anything above 10Hz to be too fast, and anything below 0.1Hz to be too slow. Repetitive signals in this range are considered to be *events* in our system. For example, waving a flag is an event, clapping is an event, but the vibration of a violin string is not an event (too fast), and neither is the daily rise and fall of the sun (too slow). Such a restriction is related to the idea of natural kinds [7], where perception is based on the physical dimensions and practical interests of the observer.

To find periodicity in signals, the most obvious approach is to use some version of the Fourier transform. And indeed our experience is that use of STFT demonstrates good performance when applied to the visual trajectory of periodically moving objects (as Section 2 shows). However, our experience also leads us to believe that this approach is not ideal for detecting periodicity of *acoustic* signals. Of course, acoustic signals have a rich structure around and above the kHz range, for which the Fourier transform and related transforms are very useful. But detecting gross repetition around the single Hz range is very different. 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. Acoustic signals may vary considerably in amplitude between repetitions, and there is significant variability or drift in the length of the periods. These two properties combine to reduce the efficacy of Fourier analysis. This led us to the development of a more robust method for periodicity detection, which is now described. In the following, the term *period* is used strictly to describe event-scale repetition (in the Hz range), as opposed to acoustic-scale oscillation (in the kHz range).

Period estimation – 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 – The previous procedure gives us an estimate $p_{estimate}$ of the event period. But the possibility of drift and variability in the period is explicitly

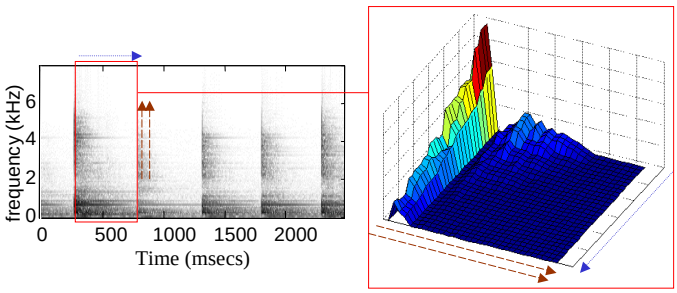

Fig. 9. Extraction of an acoustic pattern from a periodic sound (a hammer banging). The algorithm for signal segmentation is applied to each normalized frequency band. The box on the right shows one complete segmented period of the signal. Time and frequency axes are labelled with single and double arrows respectively.

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 signal 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 $\| c_i - c_j \|_2 < p_{estimate}/2$ where c_i and c_j are the cluster centers.

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 average interval is the final estimate of the signal period.

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. If a strong periodicity is not found at the default time scale, the time window is split in two and the procedure is repeated for each half. This constitutes a flexible compromise between both the time and frequency based views of a signal. Figure 9 shows segmentation of the sound of the hammer in the time-domain. A microphone array samples the sounds around the robot at 16kHz. The Fourier transform of this signal is taken with a window size of 512 samples and a repetition rate of 31.25Hz. The Fourier coefficients are grouped into a set of

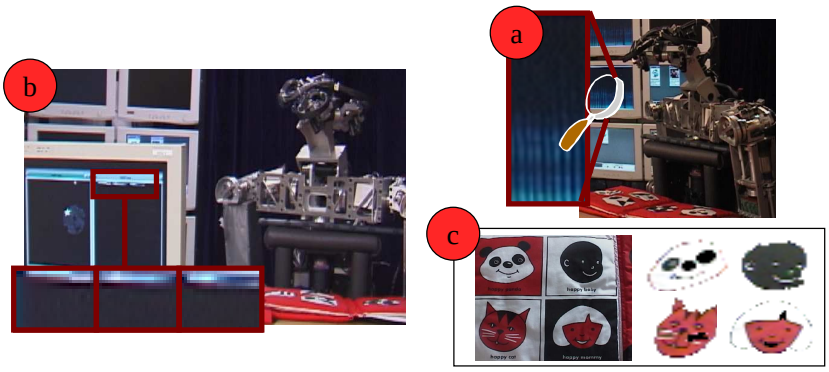

Fig. 10. a) Spectrogram of environment sounds heard by Cog. During this experiment, periodic sounds of words were being pronounced by a human actor, explaining the oscillatory pattern in the image. b) Sound segmented and associated with an object. It shows low resolution sound spectrograms resulting from an averaging process. It is worth noticing the pattern similarity among the three samples. c) Other visual segmentations extracted from the book's fabric pages.

frequency bands for the purpose of further analysis, along with the overall energy.

This sound segmentation algorithm generates training data for a sound recognition scheme (which is described in detail by Arsenio [2]). Auditory processing is also integrated with visual processing to extract the name and properties of objects. However, hand visual trajectory properties and sound properties are independent, since it is not the hand that generates sound while tapping on books, but the caregiver. Therefore, cross-modal events are associated together under a weak requirement: visual segmentations from periodic signals and sound segmentations are bound together if occurring temporally close, in contrast to Arsenio et al. [8] strong additional requirement of matching frequencies of oscillation for toys generating sounds, such as rattles.

Figure 10 presents results for sound segmentation and cross-modal binding from real-time, on-line experiments on the humanoid robot Cog. The three intra-category acoustic patterns are similar. More experimental results for naming three different objects are shown in Figure 11. The spectrograms obtained from sound segmentation differ considerably for the names of the three objects (inter-category).

4. Educational and Play Learning Activities

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 three different sources: from an attentional tracker [4], which tracks the attentional focus and is attracted to a new salient stimulus; from a multi-target tracking algorithm, implemented to track simultaneously multiple targets [2]; and from an algorithm that selectively attends to the human actuator

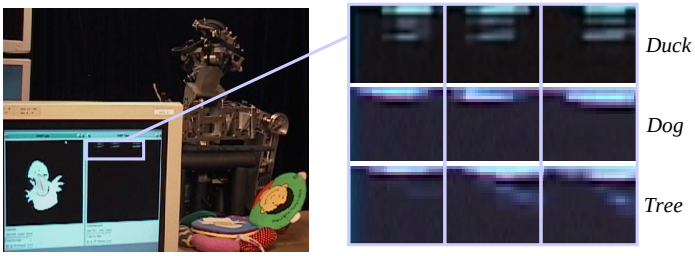

Fig. 11. Sounds of words: *Duck, dog barking and tree.*

for the extraction of periodic signals from its trajectory. This algorithm operates at temporal, pyramidal levels with a maximum time scale of 16 seconds, as follows:

(1) A skin detector [9] extracts skin-tone pixels over a sequence of images
(2) A blob detector [2] groups and labels the skin-tone pixels into connected regions
(3) Non-periodic blobs are tracked over the time sequence are filtered out. Visual periodicity detection (applying STFTs) is as described in Section 2.
(4) A trajectory if formed from the oscillating blob's center of mass over the temporal sequence.

Whenever a repetitive trajectory is detected from any of these parallel processes, it is partitioned into a collection of trajectories, each element being of a such collection described by the trajectory points between two zero velocity points with equal sign on a neighborhood, applying the partitioning process described in Section 3. The object recognition algorithm [2] is then applied to extract correlations between these sensory signals perceived from the world and geometric shapes presented in such world, or in the robot database of previously recognized objects [2], 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 the recognition scheme previously described, trajectory lines are matched to oriented edge lines (from a Canny detector) in
   (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.

4.1. *Learning Hand Gestures*

Standard hand gesture recognition algorithms require an annotated database of hand gestures, built off-line. Common approaches, such as Space-Time Gestures [10], rely on dynamic programming. Cutler et al. [11] developed a system for children to interact with lifelike characters and play virtual instruments by classifying optical

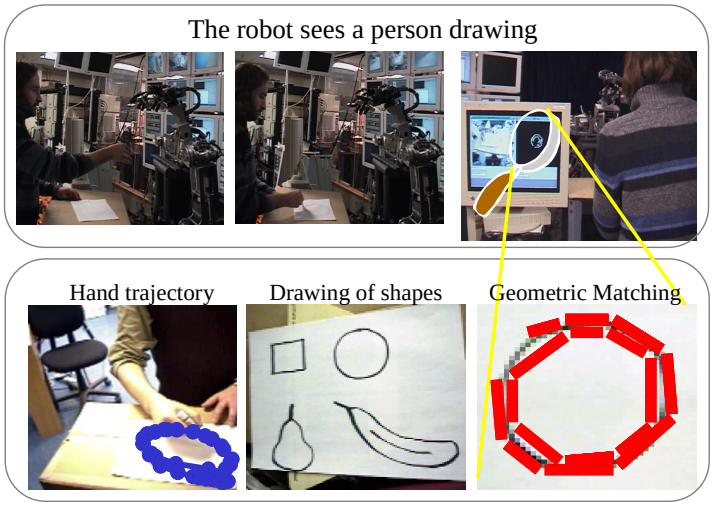

Fig. 12. (Top) A human draws a circle on a sheet of paper with a pen. (Bottom) The hand circular trajectory is matched to another circle previously recognized and stored (see Figure 7).

flow measurements. Other classification techniques include state machines, dynamic time warping or HMMs. We follow a fundamentally different approach, being periodic hand trajectories mapped into geometric descriptions of these objects

(1) Objects are recognized on-line, following the method in Section 2.2. Similarity invariants are computed from such training data, and stored in hash tables
(2) The trajectory of the periodic hand gestures projected into the retinal image defines a contour image
(3) Oriented pairs of lines are fitted to such contour
(4) Similarity invariants computed from these pairs are then matched to the similarity invariants defined by pairs of lines stored in the hash tables

Figure 12 reports an experiment in which a human draws repetitively a geometric shape on a sheet of paper with a pen. The robot learns what was drawn by matching one period of the hand gesture to the previously learned shape (the hand gesture is recognized as circular). Hence, the geometry of periodic hand trajectories are on-line recognized to the geometry of objects in an object database, instead of being mapped to a database of annotated gestures.

4.2. *Object Recognition from Hand Gestures*

The problem of recognizing objects in a scene can be framed as the dual version of the hand gestures recognition problem. Instead of using previously learned object geometries to recognize hand gestures, hand gestures' trajectories are now applied to recover the geometric shape (defined by a set of lines) and appearance (given by an image template enclosing such lines) of a scene object (as seen by the robot):

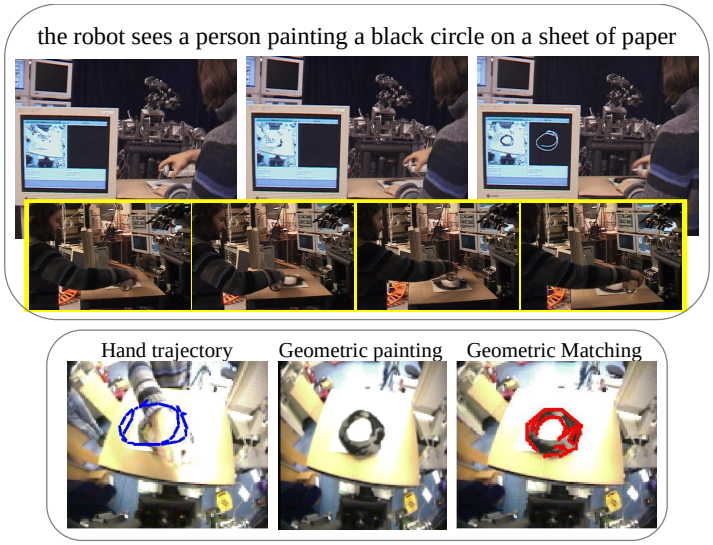

Fig. 13. (Top) Human paints a black circle with a ink can on a sheet of paper (two different views of an experiment running on Cog are shown). The circle is painted multiple times. (Bottom) The 1^{st} image from the left displays the hand trajectory, the 2^{nd} image shows the geometric circle drawn, and the last shows edges of the painted circle which were matched to the hand trajectory.

(1) The algorithm first detects oriented pairs of lines in a image of a world scene
(2) The geometry of periodic hand gestures is used to build a contour image
(3) The world image is masked by a dilated binary mask which encloses the arm trajectory on the contour image
(4) Oriented pairs of lines fitted to the contour image are then matched to the pairs of lines on the world image through the object recognition procedure.

Hence, visual geometries in a scene (such as circles) are recognized as such from hand gestures having the same geometry (as is the case of circular gestures). Figure 13 shows results for such task. The robot learns what was painted by matching the hand gesture to the shape defined by the ink on the paper. This algorithm is useful to identify shapes from drawing, painting or other educational activities.

4.2.1. *Shape from Human Cues*

This same framework is applied to extract object boundaries from human cues. Indeed, human manipulation provides the robot with extra perceptual information concerning objects, by actively describing (using human arm/hand/finger trajectories) object contours or the hollow parts of objects, such as a cup (Figure 14). Tactile perception of objects from the robot grasping activities has been actively pursued [12]. Although more precise, these techniques require hybrid position/force control of the robot's manipulator end-effector so as not to damage or break objects.

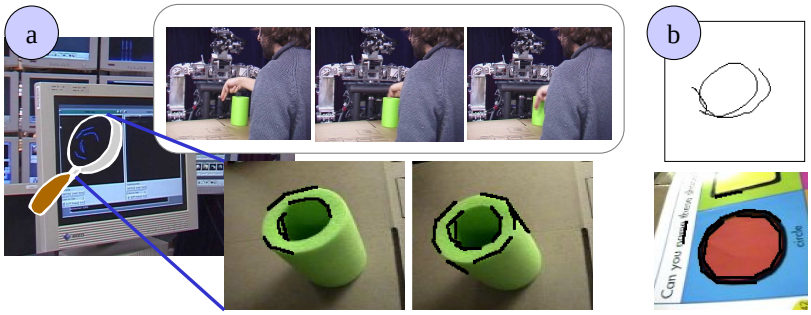

Fig. 14. a) A human moves his hand around an object boundary (top). The contour image extracted from the trajectory is then matched to the object contours. b) Some procedure applied to a geometric shape on a book.)

4.3. *Functional Constraints*

Not only hand gestures can be used to detect interesting geometric shapes in the world as seen by the robot. For instance, certain toys, such as trains, move periodically on rail tracks, with a functional constraint fixed both in time and space. Therefore, one might obtain information concerning the rail tracks by observing the train's visual trajectory. To accomplish such goal, objects are visually tracked by an attentional tracker [4] which is modulated by an attentional system [2]. The algorithm starts by masking the input world image to regions inside the moving object's visual trajectory (or outside but near the boundary). Lines modelling the object's trajectory are then mapped into lines fitting the scene edges. The output is the geometry of the stationary object which is imposing the functional constraint on the moving object.

Figure 15 shows experimental results for the specific case of extracting templates for train rail tracks from the train's motion (which is constrained by the railway circular geometry). Three such experiments were taken over a total time of 20 minutes. We opted by a conservative algorithm: only 8 recognitions were extracted from all the experiments (out of 200). Two of them originated poor quality segmentations. But the algorithm is robust to errors, since no false results were reported. Extracting a larger number of templates only depends on running the experiments for longer – or letting Cog play for some more time. The same applied for the other algorithms described in this section: tested under similar conditions, they originated recognition results of comparable performance.

5. The Robot's First Musical Tones

There is experimental evidence that an amodal relation (in this case texture, which is common to visual and tactile sensing) provides a basis for learning arbitrary relations between modality-specific properties [13] (in this case the particular colored surface of a textured object). This motivated a strategy to extract image textures

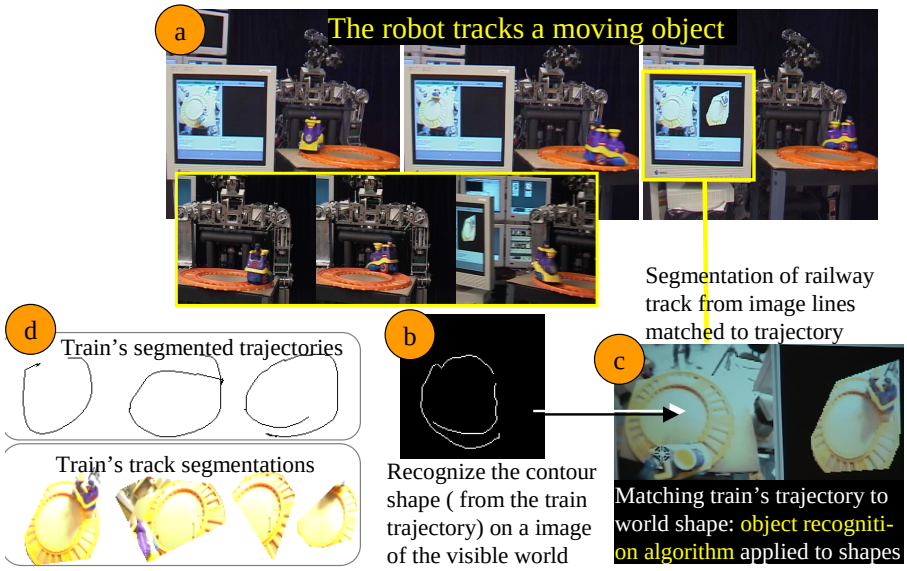

Fig. 15. a) Two different views of a toy train moving periodically around a railway track b) Trajectory extracted from one period of the train's motion c) Cog's view of the train on the rail (left), and railway tracks segmentation recovered from the lines matched d) Some result samples for train trajectories and rail tracks' segmentations.

from visual-sound patterns, i.e., by processing acoustic *textures* (the sound signatures) between visual trajectory peaks. The algorithm works by having a human probe the world by creating rhythmic sounds on a textured, roughed surface. Visual and acoustic textures are then linked as follows:

(1) Hand tracking of periodic gestures using the procedure applied in the previous section to learn from educational activities, that selectively attends to the human actuator for the extraction of periodic signals from its trajectory
(2) Tracking and mapping of the x and y human hand visual trajectories (horizontal and vertical directions in images, respectively) into coordinates along eigenvectors given by the Singular Value Decomposition, resulting in new axes x_1, y_1 (x_1 corresponding to the eigenvector along highest data variance). Three measures are then estimated:
    - The angle β of axis x_1 relative to x
    - The visual trajectory period (after trajectory smoothing to reduce noise) by detecting periodicity along the x_1 direction – using the STFT based approach
    - The amplitude difference A_v between the maximum trajectory value along x_1 and the minimum value, over one visual period
(3) Partition of the acoustic signal according to the visual periodicity, and sound periodic detection applied over multiple scales on such a window (Section 3). The goal is to estimate the spatial frequency F_s of the object's texture in the

image (with the highest energy over the spectrum). This is done by computing the number n of acoustic periods during one (or half) visual period. The spatial frequency estimate is then given by $F_s = A_v/n$, which means that the sound peaks n times from a visual minimum location to a maximum (the human is producing sounds with n peaks of energy along the hand trajectory)

(4) Spectral processing of a stationary image by applying to each image point a 1-dimensional STFT along the direction of maximum variation, with length given by the trajectory amplitude and window centered on that point, and storing for such point the energy of the F_s component of this transform. This energy image is converted to binary by a threshold given as a percentage of the maximum energy (or a lower bound, whichever is higher). The object is then segmented by applying this mask to the stationary image.

All these steps are demonstrated by the experiment in Figure 16. It shows a human hand playing rhythmic sounds with a textured metal piece (Figure 16-a), producing sound chiefly along the vertical direction of motion. The x and y visual trajectories of the human hand are tracked during a period of approximately 4 seconds (128 frames). The axis x_1 is at an angle of $\beta = 100.92^o$ with the x axis for the experiment shown. Periodic detection along x_1 (after smoothing to reduce noise) estimates a visual period of 1.28 seconds. The visual trajectory's amplitude difference A_v is 78 pixels over one visual period (Figure 16-b). Sound periodic detection is applied on the visually segmented acoustic signal over 2 scales of temporal resolution. For this experiment, the ratio between half the visual period and the sound period is $n \simeq 5$ (Figure 16-c).

The sound presents therefore 5 peaks of energy along the hand trajectory, which corresponds to a frequency of $F_s = 16Hz$. The stationary image in Figure 16-d is processed by selecting 16Hz components of the STFTs, resulting an energy image – middle – which masks the texture which produced such sound. Children toys like xylophones have a similar structure to the metal piece, which motivated this experiment. The algorithm extracted two such templates out of a 2 minute run. This corresponds to about 20% of the total number of templates which were theoretically possible to extract over this time interval. But once again, a conservative approach led to algorithm robustness to errors. It is also worth stressing however that this approach could also be applied by replacing sound with proprioceptive or tactile sensing, and the human action by robotic manipulation.

6. Conclusions

This paper presented a framework to boost the robot's perception capabilities through the use of books and other learning aids. A frequency domain technique was presented for the extraction of appearance templates for a variety of objects, such as animals, clothes, plants, utilities, fruits, furniture, among others. An object recognition scheme incorporates such appearance templates to identify common features along several objects' representations, such as paintings, drawings, photos,

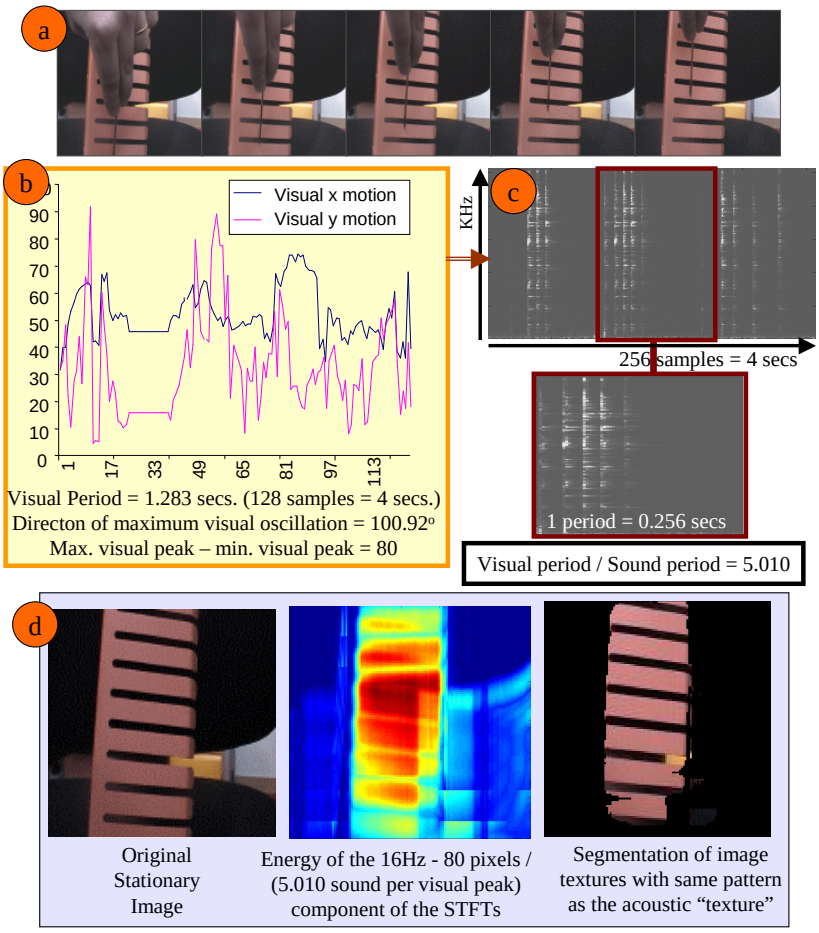

Fig. 16. Matching visual/acoustic textures to visual textures. a) Sequence of images showing a human hand playing rhythmic sounds with a textured metal piece. Sound is only produced along one direction of motion. b) Horizontal and vertical visual trajectories of the human hand, during a time window of approximately 4 seconds (128 frames). The visual period is estimated as 1.28 seconds. The amplitude difference between the maximum and minimum trajectory values is 78 pixels. Maximum variation makes a 100.92^o angle with the horizontal. c) Ratio between half the visual period and the sound period is $\simeq 5$, which means that sound peaks five times from a visual minimum location to a maximum. d) Stationary image – left – is segmented using a mask – middle – computed from the $16Hz$ energy component of the STFTs applied at each point, selecting the relevant object's texture – right.

computer generated models or real objects.

It is generally accepted [14] that playing activities are healthy for children developmental learning. We argue also to let humanoid robots play with human caregivers. To achieve such goal, algorithms were presented that enable the acquisition and recognition of percepts from educational activities such as drawing, painting, or playing with toys. Learning aids/activities were shown very useful to teach simple

language skills, by exploiting correlations between spoken words and visual events. This way, a human tutor introduces the robot both to an object's appearance and the set of phonemes used on a specific language to describe it.

If in the future humanoid robots are to behave like humans, a promising venue to achieve this goal is by treating then as such, and initially as children. Learning aids such as books or educational, play activities that stimulate a child's brain are important tools that caregivers extensively apply to communicate with children. And they also are important for human-robot interactions, towards the goal of creating a 2-year-old-infant-like artificial creature.

Acknowledgements
Project funded by DARPA's "Natural Tasking of Robots Based on Human Interaction Cues", contract 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. D. I. Perrett, A. J. Mistlin, M. H. Harries, A. J. Chitty, Understanding the visual appearance and consequence of hand action, in Vision and action: the control of grasping (Ablex, Norwood, NJ, 1990) 163–180.
2. A. Arsenio, Cognitive-Developmental Learning for a Humanoid Robot: A Caregivers' gift, Ph.D. thesis, MIT, May/June 2004.
3. A. Arsenio, Developmental Learning on a Humanoid Robot (International Joint Conference on Neural Networks, 2004).
4. P. Fitzpatrick, From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot, Ph.D. thesis, MIT, Cambridge, MA, 2003.
5. A. Arsenio, Teaching a humanoid robot from books, in International Symposium on Robotics (2004) .
6. J. Harris, Algebraic Geometry: A First Course (Graduate Texts in Mathematics, 133) (Springer-Verlag, 1994).
7. H. Hendriks-Jansen, Catching Ourselves in the Act (MIT Press, Cambridge, Massachusetts, 1996).
8. P. Fitzpatrick, A. Arsenio, Feel the beat: using cross-modal rhythm to integrate robot perception (International Workshop on Epigenetic Robotics, 2004).
9. B. Scassellati, Foundations for a Theory of Mind for a Humanoid Robot, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, 2001.
10. T. Darrel, A. Pentland, Space-time gestures, in IEEE Conference on Computer Vision and Pattern Recognition (New York, NY, 1993) 335–340.
11. R. Cutler, M. Turk, View-based interpretation of real-time optical flow for gesture recognition, in Int. Conference on Automatic Face and Gesture Recognition (1998) .
12. K. Rao, G. Medioni, H. Liu, B. G.A., Shape description and grasping for robot hand-eye coordination, *IEEE Control Systems Magazine* **9** (1989) (2) 22–29.
13. M. Hernandez-Reif, L. E. Bahrick, The development of visual-tactual perception of objects: Amodal relations provide the basis for learning arbitrary relations, *Infancy* **2** (2001) (1) 51–72.
14. S. Shelov, Your Baby's First Year (The American Academy of Pediatrics. Bantam Books, 1998).