Towards an Embodied and Situated AI

Artur M. Arsenio
MIT Computer Science and Artificial Intelligence Laboratory
200 Technology Square
Room NE43-936
arsenio@csail.mit.edu

Abstract

Real-time perception through experimental manipulation is developed using the robot arm to facilitate perception, or else exploiting human/robot social interactions (such as with a caregiver) so that the human changes the world in which it is situated, enhancing the robot's perceptions. Contrary to standard supervised learning techniques relying on a-priori availability of training data segmented manually, actions by an embodied agent are used to automatically generate training data for the learning mechanisms, so that the robot develops categorization autonomously. This framework 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. The theory is corroborated by experimental results.

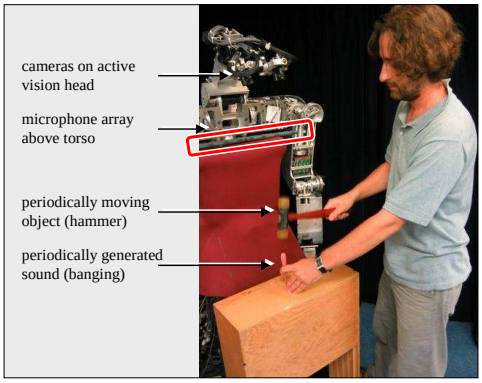

Figure 1: The experimental platform. The humanoid robot Cog is equipped with cameras in an active vision head, a microphone array across the torso and two robotic arms. A human demonstrates some repetitive action to the robot, such as using a hammer

Introduction

Embodied and situated perception (Arsenio 2002; 2003) consists of boosting the vision capabilities of an artificial creature by fully exploiting the concepts of an embodied agent situated in the world (Anderson 2003). Active vision (Aloimonos, Weiss, & Bandopadhay 1987; Bajcsy 1988), contrary to passive vision, argues 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 (Aloimonos, Weiss, & Bandopadhay 1987). 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 (Krotkov, Henriksen, & Kories 1990).

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. The human (and/or robot) body is used 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).

Real-time visual embodied strategies, which are not limited to active robotic heads, are described in this paper to boost perception capabilities. Embodied vision methods will be demonstrated with the goal of simplifying visual processing. This is achieved by selectively attending to the human actuator (*Hand*, *Arm* or *Finger*), or the robot actuator. Indeed, primates have specific brain areas to process the hand visual appearance (Perrett *et al.* 1990). Inspired on human development studies, I will first put emphasis on facilitating vision through the action of a human instructor. Through social interactions of a robot with the instructor, the latter facilitates robot's perception and learning, in the same way as human teachers facilitate children perception and learning during child development phases.

Although a human can interpret visual scenes perfectly well without acting on them, such competency is acquired developmentally by linking action and perception. Actions are not necessary for standard supervised learning, since off-line data is segmented manually. But whenever an actor has to autonomously acquire object categories using its own body to generate informative percepts, actions become indeed very useful. Therefore, the next section introduces the framework to detect visual events produced by such actions. The following sections apply this framework to tackle a variety of research issues.

Visual Perception Driven by Action

This Section presents the algorithms for identifying events at multiple spatial/frequency resolutions. Spatial event candidates are moving regions of the image that change velocity either periodically, or abruptly under contact.

Detection of Frequency Domain Events

Tools are often used in a manner that is composed of some repeated motion – consider hammers, saws, brushes, files, etc. This repetition can potentially aid a robot to perceive these objects robustly. Our approach is for the robot to detect simple repeated visual events at frequencies relevant for human interaction.

Periodic detection for events created by human teachers, such as tapping an object or waving their hand in front of the robot, is applied at multiple scales. Indeed, for objects 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 for each half.

A grid of points homogeneously sampled from a moving region in the image is tracked over a time interval of approximately 2 seconds (65 frames). The motion trajectory for each point over this time interval is determined using the Lucas-Kanade pyramidal algorithm. A Short-Time Fourier Transform (STFT) is applied to each point's motion sequence,

$$I(t, f_t) = \sum_{t'=0}^{N-1} i(t')h(t'-t)e^{-j\frac{2\pi}{N}f_t t'} \qquad (1)$$

where h is a windowing function, and N the number of frames. Periodicity is estimated from a periodogram determined for all signals from the energy of the STFTs 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. Figure 2 shows STFTs for both periodic signals and signals filtered out.

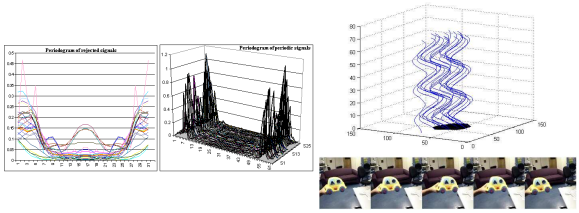

Figure 2: (left) STFTs of discarded points; (center) STFTs for a set of periodic points; (right) trajectories for periodic points from waving a toy car.

Detection of Spatial Domain Events

The algorithm to identify and track multiple objects in the image is described herein. A motion mask is first derived by subtracting gaussian filtered versions of successive images and placing non-convex polygons around any motion found. A region filling algorithm is applied to separate the mask into regions of disjoint polygons (using a 8-connectivity criterion). Each of these regions is used to mask a contour image computed by a Canny edge detector. The contour points are then tracked using the Lucas-Kanade algorithm. An affine model is built for each moving region from the position and velocity of the tracked points. Outliers are removed using the covariance estimate for such model.

Table 1: Categories of spatial events from the entities' (objects and actuator) motion.

	Type of Interaction	
Knowledge in memory	**Contact** eg. poking/grabbing an object, or assembling it to another object.	**Release** eg. throwing or dropping an object, or disassembling an object into two.
actuator/object (explicit)	▷ overlap of two entities ▷ large a priori velocities	▷ two moving entities loose contact ▷ large a priori velocities
actuator (implicit)	▷ abrupt grow of the actuator's motion area ▷ large actuator velocity ▷ abrupt velocity rise for previously stationary object	▷ large initial velocity of ensemble ▷ large a posteriori velocities for both entities ▷ motion flow of assembled region separates into two disjoint regions

Four categories of spatial events were defined, as shown in Table 1. Whenever an event occurs, the objects involved in such event are inserted into a short term memory, with a span of two seconds after the last instant the object moved. Figure 3 demonstrates the detection of two events.

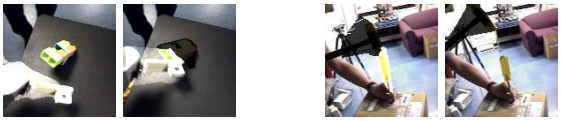

Figure 3: (left) sequence of two images for robotic arm approaching and impacting an object - shows an implicit contact. Two human legs are also moving in the background. (right) sequence of two images for hammering a nail - shows an explicit release. The arm that secures the nail also moves, but does not create an event.

Embodied object segmentation

A fundamental problem in computer vision - *Object Segmentation* - is dealt with by detecting and interpreting natural human/robot task behavior such as tapping, waving, shaking, poking, grabbing/dropping or throwing objects.

An active segmentation technique developed recently (Fitzpatrick 2003) relies on poking objects with a robot actuator. This strategy operates on first-person perspectives of the world: the robot watching its own motion. However, it is not suitable for segmenting objects based on external cues. The minimum cut algorithm (Shi & Malik 2000) is another good segmentation technique, though it suffers from problems inherent to non-embodied techniques. These problems will be dealt by exploiting shared world perspectives between a cooperative human and a robot. Object segmentation on unstructured, non-static, noisy and low resolution images is indeed hard because:

▷ objects may have similar color/texture as background

▷ multiple objects may be moving simultaneously in a scene

- ▷ significant luminosity variations
- ▷ need of real-time, fast segmentations, on low resolution images (128×128)

Segmentation Driven by Active Actuation

A scene perceived might contain several moving objects, which may have similar colors or textures as the background. Multiple moving objects create ambiguous segmentations from motion, while difficult figure/ground separation makes segmentation harder. However, a human teacher facilitates perception by waving an object (or acting on this object, such as grabbing it) in front of the robot, so that the motion of the object is used to segment it.

The set of non-skin moving points tracked over time are sparse, and hence an algorithm is required to group then into a meaningful template of the object, as follows. First, an affine flow-model is applied to the flow data to recruit other points within uncertainty bounds. Clusters of points moving coherently are then covered by a non-convex polygon – the union of a collection of locally convex polygons (Arsenio 2003) – as shown in Figure 4. This algorithm is much faster than the minimum cut algorithm (Shi & Malik 2000), and provides segmentation of similar quality to the active minimum cut approach in (Fitzpatrick 2003).

Figure 4: Samples of object segmentations. (left) Top row shows original images, while bottom row shows segmentations (right) sample segmentations from a large corpora consisting of tens of thousands of computed segmentations.

Segmentation by Demonstration

This is a human aided object segmentation algorithm. A human teacher waves the arm/hand/finger on top of the object to be segmented. The motion of skin-tone pixels (Breazeal 2000) is tracked over a time interval and the energy per frequency content is determined. A template of the actuator is built from the trajectory defined by the set of periodic moving points. Points from these trajectories are collected together, and mapped onto a reference image. A standard color segmentation (Comaniciu & Meer 1997) algorithm is applied to this reference image. The differentiated clusters of colors hence obtained need to be grouped together into the colors that form an object. This grouping works by having 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 (see Figure 5).

Therefore, points taken from waving are used to both select and group a set of segmented regions into the full object. This strategy segments objects that cannot be moved independently, such as objects printed in a book, or heavy, stationary objects such as a table or a sofa.

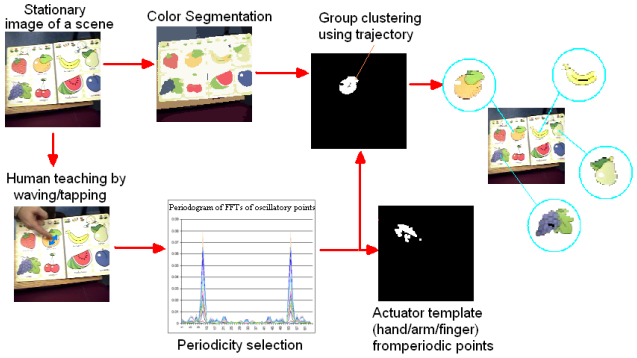

Figure 5: The actuator's trajectory is used to extract the object's color clusters.

Object Recognition

An object recognition scheme was developed, able to recognize objects from color, luminance and shape cues, or from combinations of them.

The object recognition algorithm consists of three independent algorithms. The input space for each of these algorithms consists of different features:

Color. Input features consist of groups of connected regions with similar color

Luminance. Input space consists of groups of connected regions with similar luminance

Shape. A Hough transform algorithm is applied to a contour image (which is the output of a Canny edge detector). Line orientation is determined using Sobel masks. Pairs of oriented lines are then used as input features

Geometric hashing (Wolfson & Rigoutsos 1997) is a rather useful technique for high-speed performance. In this method, invariants (or quasi-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. A Fuzzy Hash table (a hash table with variable-size buckets) was implemented to store affine color, luminance and shape invariants (which are view-independent for small perspective deformations). Figure 6 shows results for each input space, while results for real objects will be shown in the next sections.

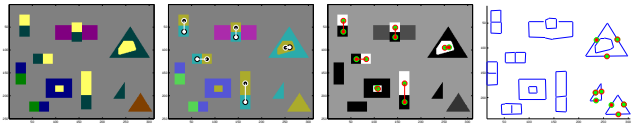

Figure 6: Within features conjunction searches for a yellow-green triangle. Lines mark features matched (left) The original image (middle-left) normalized color buckets for the original image, with results for a yellow-green query superimposed (middle-right) Luminance buckets of the original image. Shows also query results for a dark-light object (right) Search for triangles (conjunction of three oriented lines).

Cross Modal Object Recognition

The advantage of combining rhythmic information across visual and acoustic modalities for object recognition is that they have complementary properties (Krotkov, Klatzky, & Zumel 1996). Since sound waves disperse more readily than light, vision retains more spatial structure – but for the same reason it is sensitive to occlusion and the relative angle of the robot's sensors. Due to physical constraints, the set of sounds that can be generated by manipulating an object is often quite small. For tools and toys which are suited to one specific kind of manipulation – as hammers encourage banging – there is even more structure to the sound they generate. 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. The spatial trajectory is used to extract visual and acoustic features – patches of pixels, and sound frequency bands – that are associated with the object (see Figure 7).[1]

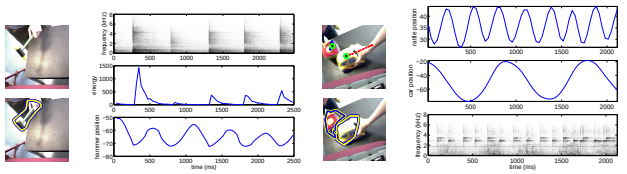

Figure 7: (left) a hammer bangs in a table. The visual trajectory of the hammer along the main axis of motion oscillates at the same frequency as the sound, with approximately zero phase-shift at the moment of impact (right) Two moving objects generating sound. Each temporal trajectory of a sound coefficient group is mapped into one of the visual trajectories if coherent with its periodicity. The sound energy has two peaks per period, since the sound of rolling is loudest during the two moments of high velocity motion between turning points in the car's trajectory (because of mechanical rubbing). The object's sound is segmented from the background by clustering the frequency bands with the same period (or half the period) as the visual target, and assigning those bands to the object (the estimated line on the spectogram separates the frequency bands associated to each sound).

Context Priming on Stationary Objects

The structure of the arm relative to a scene structure provides a natural way for constraining the object detection problem using global information. In addition, the environment surrounding the robot has also an embedded structure that can be learned using supervised learning techniques.

Given the image of an object (for instance, a car), its meaning depends often on the surrounding context. For example, in Figure 8-left, a *Ferrari* car may correspond to a real automobile or to a toy model. Context cues remove such ambiguity. In addition, for two images of an object at different scales (as illustrated by Figure 8-right) it is hard to determine, without contextual cues, if the real objects are of the same size. However, giving contextual cues, humans can assert with confidence such relationships.

Considering Figure 9, both sofa and table segmentations are hard cases to solve. The clustering of regions by table-

[1]This is collaborative work with *Paul Fitzpatrick.*

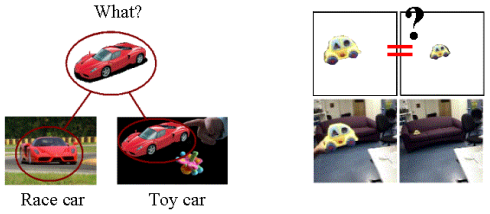

Figure 8: (left) Context cues are essential to remove ambiguity - an environment with trees, asphalt roads suggest a big, racing automobile, while a hand picking the car, side by side with another small toy, suggests a small toy car. (right) Contextual cues enable the estimation of the correct relative size of objects from the image.

like color content produces two disjoint regions. One of them corresponds to the table, but it is not possible to infer which just from the color content. But a human teacher can *show* the table to the robot by waving on the table's surface. The arm trajectory then links the table to the correct region. For the sofa case, segmentation is hard because the sofa appearance consists of a collection of color regions. It is necessary additional information to group such regions without including the background. Once more, a human teacher *describes* the object, so that the arm trajectory groups several color regions into the same object - the sofa.

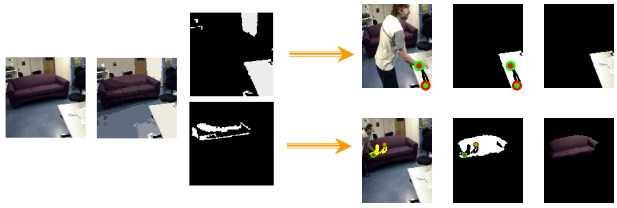

Figure 9: Segmentation of heavy, stationary objects. A human teacher *shows* the table and sofa to the robot, by waving on the objects' surface, so that the robot can then use the arm trajectory to link the objects to the correct color regions.

Inferring 2.5 Sketches

Besides binocular cues, the human visual system also processes monocular data for depth inference, such as focus, perspective distortion, gravitational light distortion, among others. Previous attempts have been made on exploring scene context for depth inference (Torralba & Sinha 2001). 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 its perception. I 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 reference for extracting relative depth information – without camera calibration. This measure is extracted from periodic signals of a human hand as follows. A skin detector is applied to extract skin-tone pixels over a sequence of images. A blob detector then groups the skin-

tone pixels into five regions. These regions are tracked over the sequence, and all non-periodic blobs are filtered out. A region filling algorithm is then applied to extract a mask for the arm. The smallest eigenvalue is used as 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. Figure 10 presents both coarse depth images and 3D plots for a typical scene.

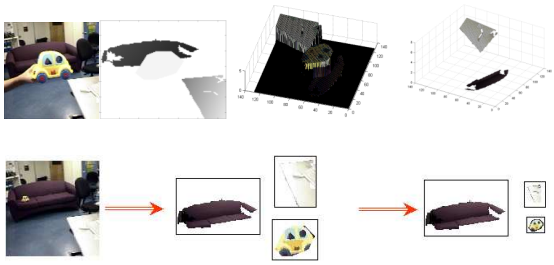

Figure 10: (top) coarse depth information for a scene on the robot's room (lighter corresponds to closer). Two right plots show 3D coarse Depth information on which the object is modelled by planes (bottom) re-scaling of objects' templates. Depth information is only available at a coarse resolution, and at discrete levels (and hence forming a 2.5 Sketch), in contrast to full 3D models.

The size of the objects on the segmented templates are not proportionally related to their true size – just compare the image sizes of the sofa, table and car in Figure 10. This is due to deformations introduced by the object's perspective projection into the retinal plane. But by using the arm diameter as reference, templates are proportionally re-scaled so that they reflect the true proportions between these objects.

Situated Vision

World structural information should be exploited in an active manner. For instance, the probability of an object being located on 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!

A method based on a weighted mixture of gaussians was developed to determine spatial context on images. Given an object, the probable location of other objects on the image, together with an estimate of their retinal size and orientation, are obtained by modelling the training data using a weighted mixture of gaussians (Luenberger 1991). This way, for each object on a scene, a link is established towards the other scene objects, as shown in Figure 11.

Learning from Books

This scheme boosts the robot's object recognition capabilities through the use of books - a learning aid. During children developmental phases, learning is often aided by the use of audiovisuals, and specially, books. Humans often paint, draw or just read books to children during the early months of childhood. Object descriptions in a book may

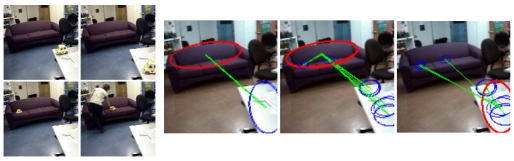

Figure 11: (left) Toy car placed along several locations on a scene. (right) Red ellipses represent a scene object, while blue ellipses represent the location, size, and orientation of other objects predicted by that object (with an associated uncertainty). From the left: 1^{st} and 2^{nd} images: predictions for table and toy car from the sofa, respectively; 3^{rd} image: prediction for toy car from the table.

came in different formats - drawings, paintings, photos, etc. Books are indeed a useful tool to teach robots different object representations or to communicate them properties of unknown objects.

The implemented strategies which enable the robot to learn from books rely 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), by tapping on their book's representations. The segmentation by demonstration method previously presented is then used to segment an object's image from book pages. This scheme was successfully applied to extract templates for fruits, geometric shapes and other elements from books, under varying light conditions (as shown in Figure 12).

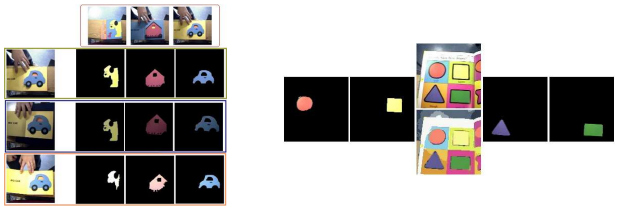

Figure 12: (left) top images show pages of a book. The other rows show segmentations for different luminosity conditions. (right) segmentations of geometric shapes from another book.

Developmental Learning from Demonstration

Object representations acquired from a book are inserted into a database, so that they become available for future recognition tasks. This way, the robot will be able to recognize (using the algorithms described in Section) real objects that, except for a description contained in a book, it has never seen before, as shown in Figure 13 for several objects.

The human ability to segment objects is not general-purpose, improving with experience. As soon as the robot acquires a complex repertoire of object representations on its database, books become useful as a means to test and validate knowledge. In addition, the robot is also able to recognize stationary objects on its field or view. A human teacher corrects eventual errors on-line by describing actively that object representation to the robot. Whenever recognition ambiguity occurs, objects can be actively segmented through actions of the robot's manipulator, such as poking the object (Arsenio 2003; Fitzpatrick 2003).

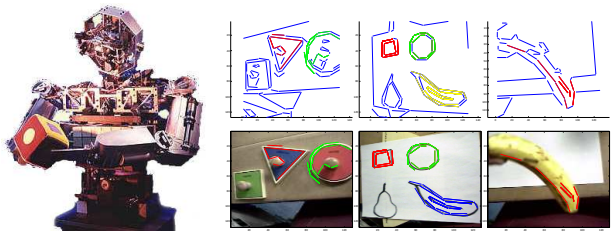

Figure 13: (left) the humanoid robot looks at an object to recognize it (right) The database of objects' templates, learned from books, is used to recognize *real objects*, or drawings of them.

Conclusions and Future Work

Embodiment and situatedness of an agent were exploited to boost its perception capabilities. We also introduced the robot/human in the learning loop to facilitate robot perception. Events originated from human/robot actions were detected at different time scales for a better compromise on frequency and spatial resolution. Objects were segmented and recognized from tasks being executed in real time, such as sawing. We also proposed strategies to segment and recognize objects that are not allowed to move. Such techniques proofed especially powerful to segment heavy objects in a scene or to teach a robot through the use of visual aids.

Undergoing and Future work

This framework is currently being applied to other research problems,

Scene Recognition from High-Level Features Given a configuration of objects, we expect not only to predict the location of other categorized objects, but also to infer the scene which these objects describe (eg. a sofa, a table with chairs and a TV are often not far apart, being common elements of a living room). Scene recognition will thus operate on higher level features (objects already categorized) compared to other research approaches (Oliva & Torralba 2001), based on low level features such as spatial distribution of frequencies

Task Detection A task can be defined as a collection of events on objects, being described by continuous states (such as a hammer oscillating or a hammer moving connected to a nail), and discrete probabilistic transitions (from the repetitive execution of tasks) among them (eg. grabbing, dropping an object). Therefore, an hybrid Markov Chain is being used to model complex tasks such as sawing, hammering, painting, among others

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 describe two different functions for the same object. We are using the shape and the motion of a tool while executing a task to classify its function

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 human-robot interactions can boost the capabilities of an embodied and situated robotic agent.

Acknowledgements

Work funded by DARPA project "Natural Tasking of Robots Based on Human Interaction Clues", contract DABT 63-00-C-10102. Author supported by Portuguese grant PRAXIS XXI BD/15851/98. The author would like to thank the anonymous reviewers for their constructive feedback.

References

Aloimonos, J.; Weiss, I.; and Bandopadhay, A. 1987. Active vision. *Int. Journal on Computer Vision* 2:333–356.

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

Arsenio, A. 2002. *Boosting Vision through Embodiment and Situatedness*. MIT CSAIL research abstracts.

Arsenio, A. 2003. Embodied vision - perceiving objects from actions. *Int. Conf. on Human-Robot Interactions*.

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

Breazeal, C. 2000. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Ph.D. Dissertation, MIT, Cambridge, MA.

Comaniciu, D., and Meer, P. 1997. Robust analysis of feature spaces: Color image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fitzpatrick, P. 2003. *From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot*. Ph.D. Dissertation, MIT, Cambridge, MA.

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

Krotkov, E.; Klatzky, R.; and Zumel, N. 1996. *Robotic perception of material: Experiments with shape-invariant acoustic measures of material type*. O. Khatib and K. Salisbury, editors, Experimental Robotics IV. Springer-Verlag.

Luenberger, D. 1991. *Linear and Nonlinear Programming*. Addison-Wesley.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision* 145–175.

Perrett, D. I.; Mistlin, A. J.; Harries, M. H.; and Chitty, A. J. 1990. Understanding the visual appearance and consequence of hand action. In *Vision and action: the control of grasping*. Norwood, NJ: Ablex. 163–180.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (22):888–905.

Torralba, A., and Sinha, P. 2001. *Indoor scene recognition*. MIT AI Memo 2001-015, CBCL Memo 202.

Wolfson, H., and Rigoutsos, I. 1997. Geometric hashing: an overview. *IEEE Computational Science and Engineering* 4:10–21.