

# Boosting Vision through Embodiment and Situatedness

Artur M. Arsenio

Artificial Intelligence Laboratory  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

<http://www.ai.mit.edu>



**The Problem:** We pretend to boost an agent's vision capabilities by fully exploiting the concepts of an embodied agent situated in the world. Thus, instead of using only active control of the perceptual mechanism to solve a visual problem, we argue for actively changing the environment so that a visual task might be easily solved.

**Motivation:** Although biological organisms are particularly good at solving visual tasks, this is still a challenge in computer vision.

In active vision the percepts can be acquired in a purposive way by action of a camera. However, it is often necessary to change features of an environment in order to solve a visual task. Some of the features that may be modified by a robotic agent (using for example a robotic arm) are the edges / corners contained in the image (spatial content), the motion (the optical flow computed from the image), the frequency content, the color content, or even the structure of a scene in an image.

For instance, a robotic manipulator may be used to change the structural content of the image, and this information can be used to memorize the location of an object, or the location where an object is highly probable to be present. We are interested in localizing objects of interest on an image wherever they are present [1], and also in estimating probable locations for an object on a scene. This way, whenever an agent needs to store an object, it will do it in a place where it usual belongs, and where the agent can find it latter.

**Previous Work:** Although some researchers exploited the fact that a robot is situated in the world, and information is also stored in that world, often this information is not exploited or even considered during visual tasks, and more specifically in object recognition. However, some previous work in our lab [2] dealt with the problem of image segmentation by poking, which, although not exploiting explicitly situatedness, makes a useful use of embodiment of an agent.

We have exploited Embodiment of an agent in different ways - for camera calibration, by using the robotic grip or the M4 nose to calibrate the cameras, instead of using a calibrated grid; and for visual guided search, through the use of the arm end-effector to memorize the location of the objects. Since the robotic arm has always the same size, the ambiguity between depth and size on an image can also be solved. This information might then be used to guide the estimation of an object's size located at the arm's end-effector.

**Approach:** To classify motion in varied dynamic systems, we intend to apply a supervised learning strategy operating on two spatial and one temporal dimensions, therefore boosting a functional object recognition scheme. A robotic arm manipulator is used to actuate the system and create the motion that feeds the learning scheme.

The grip of Cog's robotic arm (see Figure) is being used to automatically calibrate the cameras, and we are investigating its use to infer an object's size, depth or pose in a monocular image, since these features are known for the robotic arm.

**Impact:** Computer vision is having a hough application in industry, services, security, only to name a few. However, several problems remain to be solved, such as object recognition, and thus computer vision applications are still too limited due to the lack of good solutions. In addition, the approach suggested applies equally well to non-robotic agents, since a disembodied social agent might lead an embodied one to act on the environment so that the former can solve a visual task.

**Future Work:** Most of this work is currently being implemented or tested, and soon the motion recognition scheme will be implemented. We are also integrating this work with an approach for the automatic identification and control of dynamic systems, which will allow the robot to learn to control different kind of objects with different



Figure 1: Humanoid robot Cog executing a task - turning a crank (left). Probable locations for finding a object in the image - training set consisting of object with large horizontal and small vertical oscillations (Right). We used a mixture of gaussians, and the expectation-maximization algorithm EM. An iterative probabilistic clustering estimates the number of clusters based on the Rissanen order identification criteria (MDL).

degrees of functionality.

**Research Support:** This work is funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102.

#### References:

- [1] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, pages 145–175, 2001.
- [2] P. Paulfitz and G. Metta. Better vision through manipulation. *Epigenetic Robotics Workshop*, 2002.