Toward A Cognitive Computer Vision System

D. Paul Benjamin

Pace University, 1 Pace Plaza, New York, New York 10038, 212-346-1012
benjamin@pace.edu

Damian Lyons

Fordham University, 340 JMH, 441 E. Fordham Rd., Bronx, NY 10458, 718-817-4485
dlyons@fordham.edu

ABSTRACT

A truly cognitive computer vision system has not yet been built. Current vision system designs are based on mathematical and engineering considerations rather than on cognitive principles. The goal of our research project is to create a vision system that is both cognitively and neurally plausible. We are designing a human-like vision system to better understand how the human visual system works and to understand any engineering advantages it possesses.

We are building a cognitive vision system that is suitable for a mobile robot. This system works in a manner similar to the human vision system, using saccadic, vergence and pursuit movements to extract information from visual input. At each fixation, the system builds a 3D model of a small region, combining information about distance, shape, texture and motion to create a local dynamic spatial model. These local 3D models are composed to create an overall 3D model of the robot and its environment. This approach turns the computer vision problem into a search problem whose goal is the acquisition of sufficient spatial understanding for the robot to succeed at its tasks.

This paper gives an overview of our work, beginning with the cognitive principles and structure of our existing system. We give an example showing how our system renders a bouncing ball. Then we describe our current and future work, which includes integrating our cognitive approach with existing work on neural visual structures.

1. INTRODUCTION

The current generation of behavior-based robots is programmed directly for each task. The programs are written in a way that uses as few built-in cognitive assumptions as possible, and as much sensory information as possible. The lack of cognitive assumptions gives them a certain robustness and generality in dealing with unstructured environments. However it is proving a challenge to extend the competence of such systems beyond navigation and some simple tasks [11]. Complex tasks that involve reasoning about spatial and temporal relationships require robots to possess more advanced mechanisms for planning, reasoning, learning and representation.

The ADAPT project (**A**daptive **D**ynamics and **A**ctive **P**erception for **T**hought) is a collaboration of three university research groups at Pace University, Brigham Young University, and Fordham University to produce a robot cognitive architecture that integrates the structures designed by cognitive scientists with those developed by robotics researchers for real-time perception and control [1-4]. Our goal is to create a new kind of robot architecture capable of robust behavior in unstructured environments, exhibiting problem solving and planning skills, learning from experience, novel methods of perception, comprehension of natural language and speech generation.

Our approach is fundamentally different from other hybrid architectures, which typically attempt to build a comprehensive system by connecting modules for each different capability: learning, vision, natural language, etc. Instead, we are building a *complete cognitive robotic architecture* by merging RS [9,10], which provides a model for building and reasoning about sensory-motor schemas, with Soar [5,6], a cognitive architecture that is under development at a number of universities. RS possesses a sophisticated formal language for reasoning about networks of port automata and has been successfully applied to robot planning [8]. Soar is a unified cognitive architecture [45] that has been successfully applied to a wide range of tasks including tactical air warfare [12].

One of the most unique and important aspects of our architecture is its treatment of perception and language and their relationship to knowledge representation. Our view of perception is that it is an "active" process [7] that is goal-directed and task-dependent, i.e. it is a cognitive problem-solving process rather than a peripheral activity separate from cognitive processing. Furthermore, we view perception as intimately linked with the formation and modification of representations, so that perception's main purpose is to identify, build and modify representational structures that make the robot's goals easier to achieve. In this view, the robot solves problems primarily by searching among different ways of perceiving the world and the task. This is in contrast to the usual approach of searching among sequences of actions in one or a few fixed representations. Each way of perceiving the world and task leads to a distinct symbolic formulation.

This cognitive approach to vision casts visual perception as a search process. ADAPT searches among ways to build a model of itself and the environment; its goal is a model that is sufficiently accurate to use for planning. The next section describes this search process in detail.

2. A COGNITIVE APPROACH TO VISION

One of the distinctive features of ADAPT is its virtual world. ADAPT’s virtual world is a multimedia simulation platform capable of realistic simulations of physical phenomena. It combines the various forms of map information found in most robots: topological, metric and conceptual information. In addition, this virtual world has a sophisticated physics plugin, giving it the ability to predict dynamics. ADAPT completely controls this virtual world, and can create arbitrary objects and behaviors in it. Central to ADAPT’s use of its virtual world is its ability to view these constructions from any point.

ADAPT uses this virtual world in a novel way. Typical robotics architectures connect their sensory mechanisms to their world models, so that sensory data is processed and modeled in the world model. The reasoning engine then operates on the world model to plan the robot’s behaviors. This type of architecture treats perception as a separate process from the central reasoning, and typically the implementation reflects this, e.g. a computer vision module processes the vision data and puts symbolic representations of the recognized objects and their relationships in the world model, and the reasoning engine then manipulates these symbols to plan and learn. The reasoning engine does not process the sensory data.

In contrast, ADAPT’s virtual world is not connected to its sensory processes. ADAPT’s sensory data is placed directly in the reasoning engine (after some low-level processing). ***The reasoning engine’s principal task in ADAPT is to reason about how to model the data***. It does this using the following loop:

> It compares visual data from the camera with visual data from the corresponding virtual camera, using a least-squares measure to find areas of disagreement. Each disagreement causes a Soar operator to be proposed to attend to that disagreement.
>
> One Soar operator is selected, based on the robot’s current goals. For example, if the goal is navigation, operators will be preferred that are for visual disagreements in the robot’s current path. The selected operator fires, causing the cameras to saccade to the area of disagreement and fixate on it.
>
> Stereo disparity, color segmentation, and optical flow are computed only in the small region of focus. Restricting the computation to this small region permits the use of highly accurate but computationally expensive optimization algorithms for these computations.
>
> This information is input to the object recognition database, and a mesh model of the best match is rendered into the virtual world. If the information indicates that a current mesh model is inaccurate, that model is modified to incorporate the new information. Optical flow is used to create or update the RS process model associated with each object.

The reasoning engine searches alternative combinations of virtual entities and behaviors to attempt to minimize the measured disagreement. In this way, *perception becomes a problem-solving process*. This enables all the knowledge of the system to be brought to bear on perception, and unifies the reasoning and learning processes of problem solving with those of perception.

The research hypothesis of this work is that the movements of the robot's cameras are only those that are necessary to build a sufficiently accurate world model for the robot's current goals. For example, if the goal is to navigate through a room, the model needs to contain any obstacles that would be encountered, giving their approximate positions and sizes. Other information does not need to be rendered into the virtual world, so this approach trades model accuracy for speed.

In this way, ADAPT prunes the information at the perception stage, using its knowledge about objects and dynamics. This is in contrast to the usual approach gathering lots of sensory information, processing it all and rendering it into a world model, then deciding which information is necessary for decision making. This latter approach wastes a great deal of time and processing effort to refine information that is discarded in the decision making process.

3. EXAMPLE

A robot wishes to predict the path of a bouncing ball so that it can intercept the ball as quickly as possible. The ball will bounce off walls that will alter its path. The robot needs to perceive the objects that the ball will hit and also perceive the ball's motion, then combine this information to produce a predicted path.

In Figure 1 below, we see two boards placed on the floor. Our vision system detects keypoints and lines in the images, then selects a region for initial focus. The density of keypoints at the bottom of the images causes this to be selected as the region of focus. This region is denoted by the dark boxes in Figure 2.

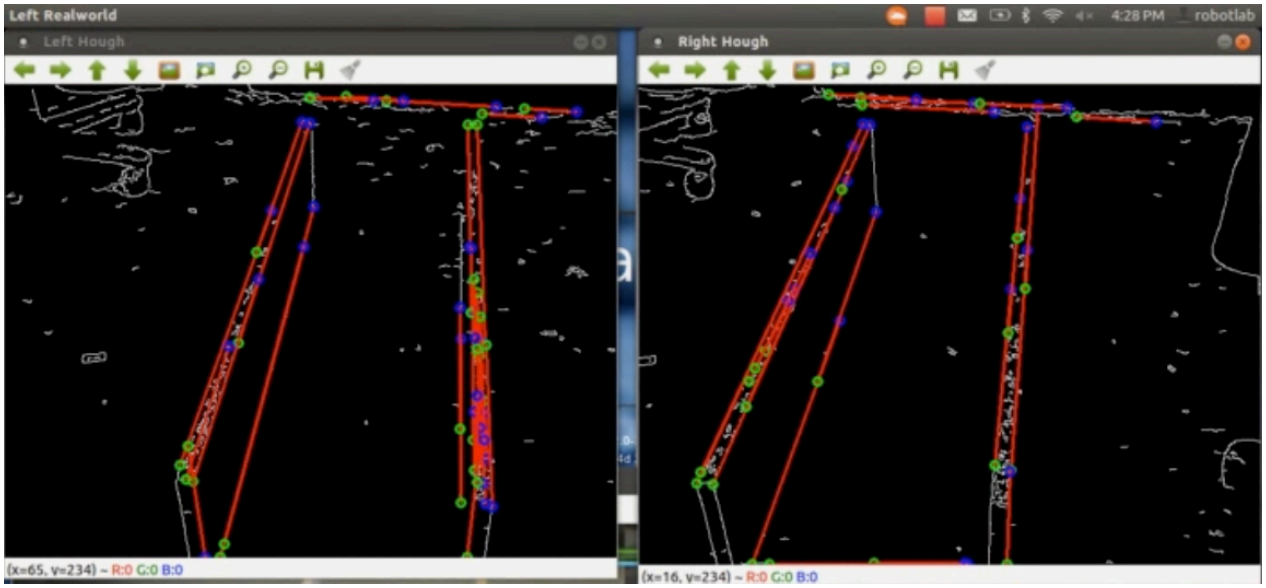
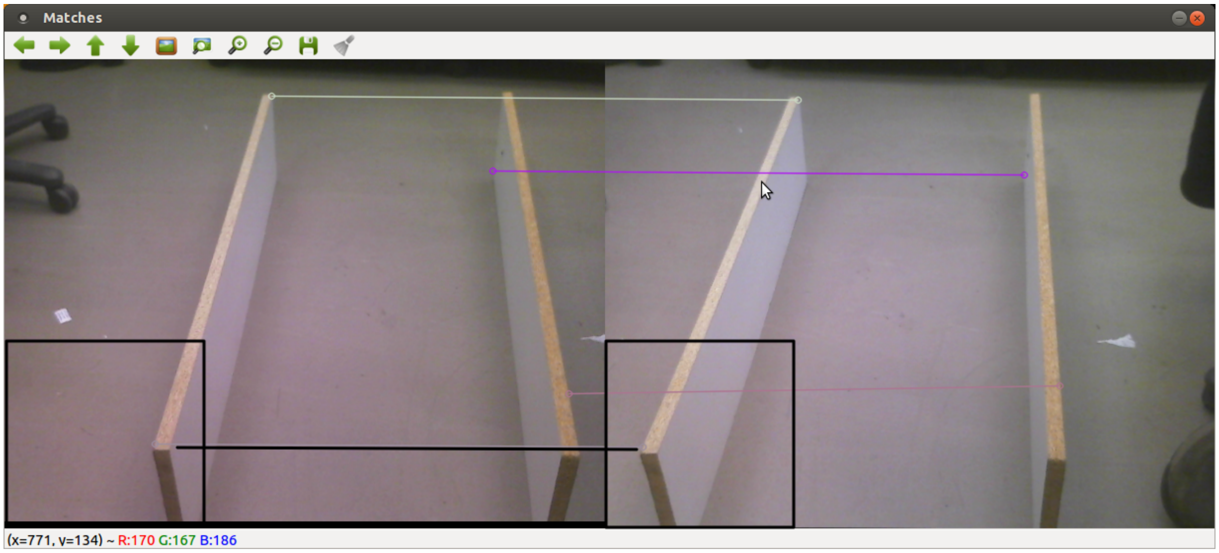


Figure 1. Initial processing of two boards on the floor, showing keypoints and lines.

The upper corners of the boards are detected, and registration of the left and right images produces an initial correspondence, together with position and orientation data.

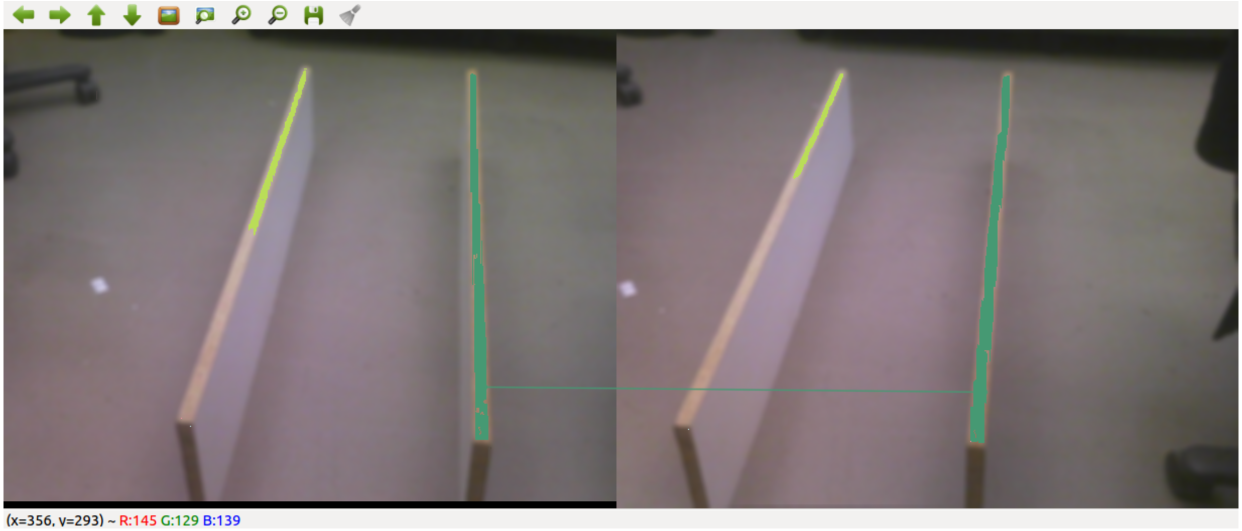


```
(x=771, y=134) ~ R:170 G:167 B:186
```

Figure 2. The region of focus is the boxes at the lower left, yielding the correspondence between the corners.

A saccade is performed to the next region of focus. In this case, the closest region is chosen, which is a correspondence between the right boards. This process is repeated four times until the top corners of the boards are reached.

Segmentation information is added, producing additional correspondences. This is combined with the correspondences from the keypoints, yielding a small set of best correspondences. In Figure 3, we see the tops of the boards correspond.



```
(x=356, y=293) ~ R:145 G:129 B:139
```

Figure 3. Segmentation correspondence between the boards.

Finally, the boards are rendered in PhysX, and a ball is added. The ball is rolled from right to left.

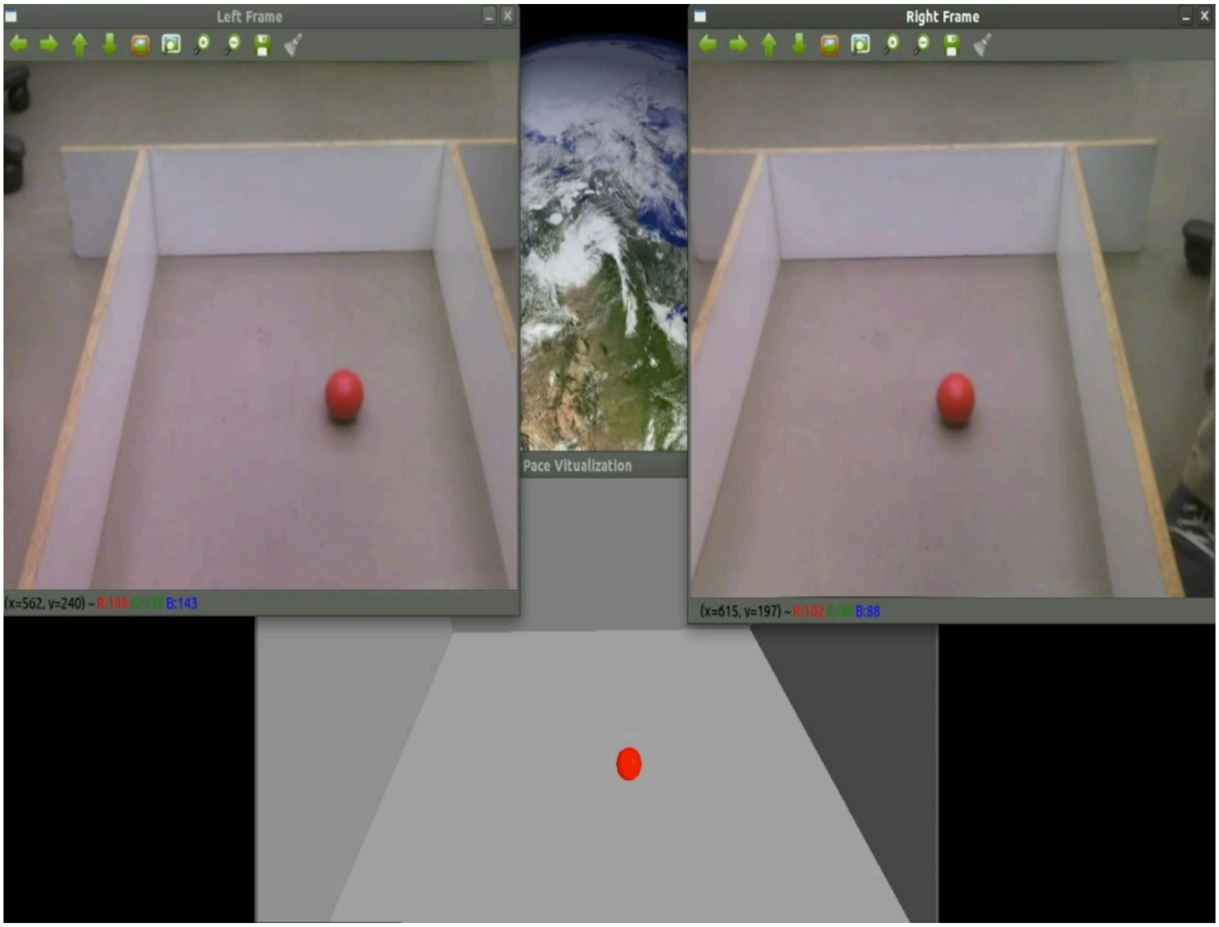

Left Frame
Pace Vitualization
Right Frame
(x=562, y=240) ~ R:145 G:135 B:143
(x=615, y=197) ~ R:102 G:96 B:88

Figure 4. A ball is rolled between two boards. Left and right images are at top. The virtual world is at bottom.

The direction and velocity of the ball are computed over a small interval then duplicated in the virtual world. The physics engine is then run much faster than real-time, producing a predicted path for the ball. A mobile robot can use this prediction to intercept the ball efficiently.

A number of videos showing this process in various scenarios are available at http://csis.pace.edu/robotlab/videos.html

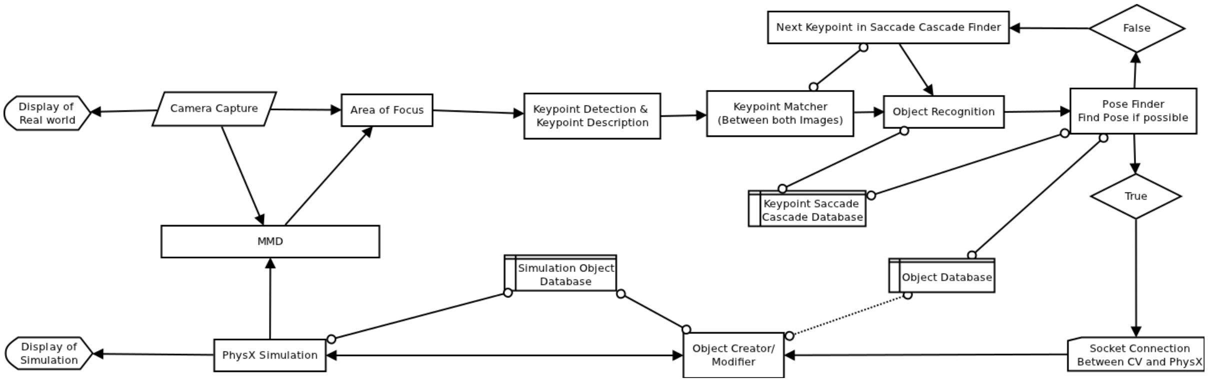

Next Keypoint in Saccade Cascade Finder
False
Display of Real world
Camera Capture
Area of Focus
Keypoint Detection & Keypoint Description
Keypoint Matcher (Between both Images)
Object Recognition
Pose Finder Find Pose if possible
Keypoint Saccade Cascade Database
True
MMD
Simulation Object Database
Object Database
Display of Simulation
PhysX Simulation
Object Creator/ Modifier
Socket Connection Between CV and PhysX

Figure 5. Schematic of the information flow in the system.

4. INTEGRATING COGNITION WITH NEURAL IMPLEMENTATION

A key goal of our project is to connect cognitive architectures with neural substrates to produce a neuroscientifically complete implementation. The convergence between behavioural and physiological findings in understanding the generation of saccades is an outstanding accomplishment of neuroscience. A large body of work has investigated the neural structure of the vision system, and especially that of saccades. For example, Buschman and Miller [14, 15] investigated the structure of serial search strategies in saccades, finding that scenes were serially scanned, driven by a fixed clock of 25 times per second. Aks [13] surveys the work on search strategies in saccades, with the goal of understanding and integrating dynamic neural patterns of saccades in search, and especially the role of self-organized criticality.

This body of work provides a detailed account of the neural implementations of saccades and fixations in the human visual system, and how they can be used to search. However, the cognitive processing of the system (the goals) are not as well understood, and this literature does not connect them to saccades. We are currently investigating plausible goal structures and search methods for saccades. Our overall hypothesis is that one of the main purposes of the visual system is to build and maintain the 3D spatial model of the environment, and that saccades and fixations implement the search strategy used for this purpose. Our approach is to integrate various neural implementations with our existing cognitive system, and test the combined system on navigational tasks with our mobile robots, including tasks that involve interaction with people.

5. SUMMARY

We have sketched the overall design of a cognitive computer vision system based on the structure and behavior of the human visual system. Our system builds a 3D model of a dynamic environment, updating it in real time as the world changes. Stereo cameras are moved and refocused by a cognitive architecture to build and update this model. We are integrating our system with neural implementations of saccade structures derived from the published literature, and evaluating numerous tradeoffs in designing such a system. The vision system will be used on a mobile robot that performs a variety of tasks and interacts with humans.

Further information on this work, including video clips showing the robot moving under the control of schemas and the use of the world model, can be downloaded from the website for the Pace University Robotics Lab: http://csis.pace.edu/robotlab

REFERENCES

[1] Benjamin, D. Paul, Damian Lyons and Deryle Lonsdale, "Embodying a Cognitive Model in a Mobile Robot", Proceedings of the SPIE Conference on Intelligent Robots and Computer Vision, Boston, October, 2006.

[2] Benjamin, D. Paul, Damian Lyons and Thomas Achtemichuk, "Obstacle Avoidance using Predictive Vision based on a Dynamic 3D World Model", Proceedings of the SPIE Conference on Intelligent Robots and Computer Vision, Boston, October, 2006.

[3] Benjamin, D. Paul, Damian Lyons and Deryle Lonsdale, "Designing a Robot Cognitive Architecture with Concurrency and Active Perception", Proceedings of the AAAI Fall Symposium on the Intersection of Cognitive Science and Robotics, Washington, D.C., October, 2004.

[4] Benjamin, D. Paul, Damian Lyons and Deryle Lonsdale, " Cognitive Robots: Integrating Perception, Action and Problem Solving in Behavior-Based Robots", AAMAS-2004 Proceedings, pp. 1308-1309, 2004.

[5] Laird, J.E., Newell, A. and Rosenbloom, P.S., "Soar: An Architecture for General Intelligence", *Artificial Intelligence* **33**, pp.1-64, 1987.

[6] Newell, Allen, *Unified Theories of Cognition*, Harvard University Press, Cambridge, Massachusetts, 1990.

[7] A. Blake and A. Yuille, eds, *Active Vision* , MIT Press, Cambridge, MA, 1992.

[8] Lyons, D.M. and Hendriks, A., “Exploiting Patterns of Interaction to Select Reactions”, Special Issue on Computational Theories of Interaction, Artificial Intelligence **73**, 1995, pp.117-148.

[9] Lyons, D.M., “Representing and Analysing Action Plans as Networks of Concurrent Processes”, IEEE Transactions on Robotics and Automation, June 1993.

[10] Lyons, D.M. and Arbib, M.A., “A Formal Model of Computation for Sensory-based Robotics”, IEEE Transactions on Robotics and Automation **5**(3), Jun. 1989.

[11] M. Nicolescu and M. Mataric, “Extending Behavior-based System Capabilities Using an Abstract Behavior Representation”, *Working Notes of the AAAI Fall Symposium on Parallel Cognition*, pages 27-34, North Falmouth, MA, November 3-5, 2000.

[12] Rosenbloom, P.S., Johnson, W.L., Jones, R.M., Koss, F., Laird, J.E., Lehman, J.F., Rubinoff, R., Schwamb, K.B., and Tambe, M., “Intelligent Automated Agents for Tactical Air Simulation: A Progress Report”, Proceedings of the Fourth Conference on Computer Generated Forces and Behavioral Representation, pp.69-78, 1994.

[13] Aks, Deborah J. (2005). 1/f Dynamic in Complex Visual Search: Evidence for Self-Organized Criticality in Human Perception. In M. A. Riley & G. C. Van Orden (Eds.), Tutorials in contemporary nonlinear methods for the behavioral sciences (pp. 353-400). Retrieved March 1, 2005, from http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp

[14] Buschman, T.J., Miller, E.K., “Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations”, Neuron 63 (3), 386-396 (2009).

[15] Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., Miller, E.K., “Synchronous oscillatory neural ensembles for rules in the prefrontal cortex”, Neuron 76 (4), 838-846 (2012).