Spatial Understanding as a Common Basis for Human-Robot Collaboration

D. Paul Benjamin[1], Tianyu Li[1], Peiyi Shen[1], Hong Yue[1], Zhenkang Zhao[1] and Damian Lyons[2]

[1] Pace University, 1 Pace Plaza, New York, New York, 10038, United States of America
{dbenjamin, yh19243n}@pace.edu
[2] Fordham University, 340 JMH, 441 E. Fordham Rd. Bronx, New York, 10458, United States of America
dlyons@fordham.edu

Abstract. We are developing a robotic cognitive architecture to be embedded in autonomous robots that can safely interact and collaborate with people on a wide range of physical tasks. Achieving true autonomy requires increasing the robot's understanding of the dynamics of its world (physical understanding), and particularly the actions of people (cognitive understanding). Our system's cognitive understanding arises from the Soar cognitive architecture, which constitutes the reasoning and planning component. The system's physical understanding stems from its central representation, which is a 3D virtual world that the architecture synchronizes with the environment in real time. The virtual world provides a common representation between the robot and humans, thus improving trust between them and promoting effective collaboration.

Keywords: Virtual World · Soar Cognitive Architecture

1 Introduction: Mental Models

Computer vision has had a difficult time reproducing the human ability to understand visual scene information across a wide range of applications domains and environmental conditions. There is evidence from cognitive psychology [1] that effectively leveraging context is a key aspect of this human facility. However, while there has a strong bottom-up Marr-based stream of vision research [2], the use of context has also been recognized in computer vision for a long time: at least from the Univ. of Mass. VISIONS project [3] and more recently to the linguistic-inspired Bag of Words approaches (e.g., [4]) global extensions of scale-invariant features (e.g., [5]) and others [6]. But in general these approaches still view scene recognition as a 'recognize the snapshot' problem, with little input from ongoing, long term objectives and tasks of the system. The scene understanding problem for a human is one of an embedded system leveraging sensing to fulfill its goals: sensing is strongly biased in the service of task and how the agent's and other agents' actions are expected to play out in the physical world.

Research in cognitive psychology indicates that the use of 3D models in spatial comprehension is fundamental, even in people who have been blind since birth [7].

Recent evidence in cognitive psychology [8] and neuroscience [9] supports the proposition that simulation, the "re-enactment of perceptual, motor and introspective states" is a central cognitive mechanism that helps to provide context for planning. Shanahan [8] proposes a large-scale neurologically plausible architecture that allows for direct action (similar to a behavior-based approach) and also "higher-order" or "internally looped" actions that correspond to the rehearsal or simulation of action without overt motion.

People have evolved a set of sophisticated strategies for using limited short term memory and limited processing speed to solve extremely difficult problems, including visual comprehension. These strategies reflect an engineered structure that avoids the computationally expensive algorithms of modern computer vision. Our vision system architecture is directly inspired by this cognitive and neurobiological structure, and our goal is to try to replicate it at a functional level. One of our unfunded collaborators is a professor of neurobiology at Fordham University, who will advise us on the cognitive and neurobiological plausibility of aspects of our system and also compare our system's performance with that of the human visual system.

The human vision system does not apply equal computational resources everywhere in its visual field, but instead focuses on and analyzes just a small portion of the visual field at each moment; this is called a *fixation* [10]. After extracting the needed information from that region of the visual field, the vision system rapidly moves the eyes to a new region of the visual field for the next fixation. These rapid movements are *saccades*, which are quick movements across the visual field, and *vergences*, which change the depth of focus [10]. The effect of this organizational structure is to permit efficient use of limited computational resources. Instead of fully processing all of the sensory input and then discarding everything that is not relevant to the goals, this organization applies computational resources only to the parts of the sensory input that are likely to be relevant to the agent's goals. The key is to organize the search of the visual field in a manner that effectively gathers useful information.

Much work has been done on measuring the functioning of the human vision system and of its system of saccades and vergences [10], but there has not been a computational implementation that connects the actions of the vision system to the goals of the agent.

The research hypothesis of our work is that the movements of the vision system are those that are necessary to build a sufficiently accurate 3D 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. The vision system needs to obtain this information; other information does not need to be rendered into the virtual world.

In this way, our system prunes the information at the perception stage, using its knowledge about the agent's goals and about objects in the world and their dynamics to decide where to look and what type of information to obtain. This is in contrast to the usual approach of gathering lots of sensory information, processing it all and rendering it into a world model in a goal-independent manner, then deciding which information is necessary for decision making. This latter approach wastes a great deal of processing time processing information that is discarded in the decision-making process. Our goal is to design a fast, inexpensive vision system by emulating the functional organization of the human vision system.

This approach to spatial comprehension and modeling is based on the functional structure of the human visual system. The usual approach to the use of vision in robotics is to attempt to solve two problems [11]:

(a) Process visual data to extract all the objects and motions in the environment,
(b) Identify the results from (a) that are important and relevant to the current task.

Unfortunately, both of these steps are very expensive computationally. The first step requires processing an enormous amount of visual data, especially when the environment is very dynamic. The second step is a difficult data mining problem.

Our approach to this complexity issue is to leverage goal-directed rendering: the robot first decides which aspects of its environment are relevant, based on its task goals. This information is used to focus the cameras on specific regions of the environment and extract only the information needed for the goals. This approach is important because it has the potential to be faster and less expensive than current approaches. Our current system runs on a laptop in real time.

2 A Basic Example: Tracking a Rolling Ball

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.

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.

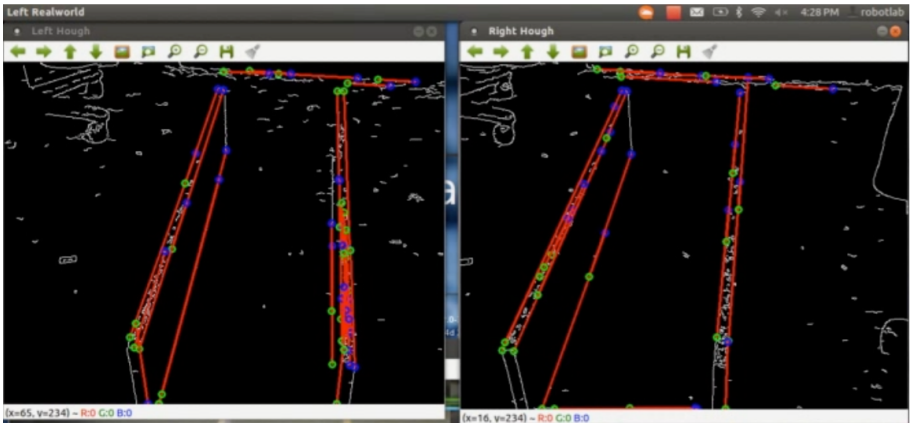

Fig. 1. Initial processing of two boards on the floor, showing keypoints and lines.

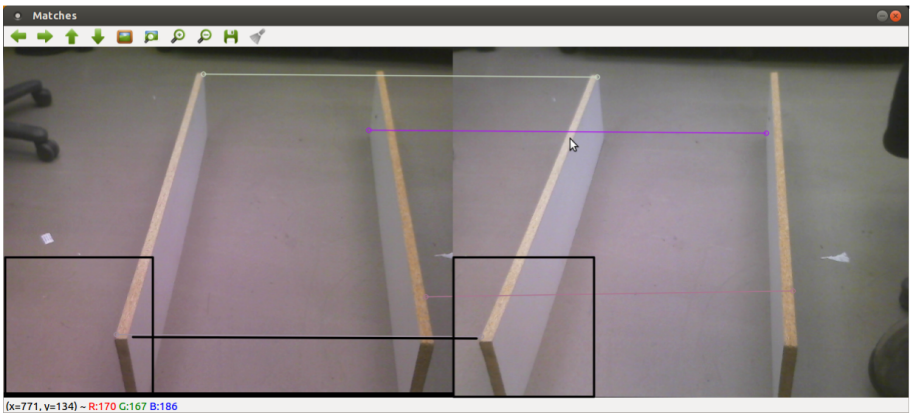

Fig. 2. The region of focus is the boxes at the lower left, yielding the correspondence between the corners.

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.

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

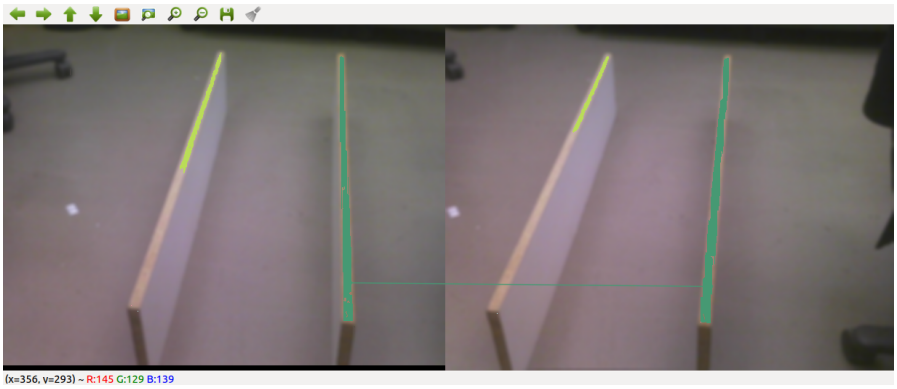

Fig. 3. Segmentation correspondence between the boards.

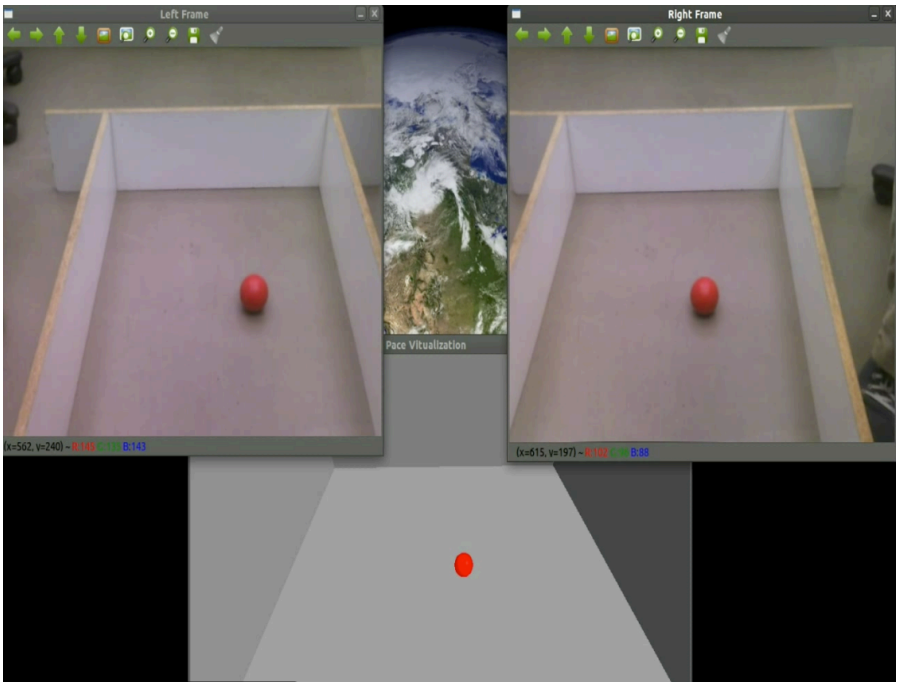

Fig. 4. A ball is rolled between two boards. Left and right images are at top. The virtual world is at bottom.

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.

3 Example: Tracking a Moving Car

Our current work is to monitor and predict the motion of a car so that we can drive another car beside it without any accidents. We are patterning our work on the KITTI project because of its excellent data, and our goal is to render actual traffic into the 3D model in real time. Our initial step is to use the PhysX vehicle demo for both worlds.

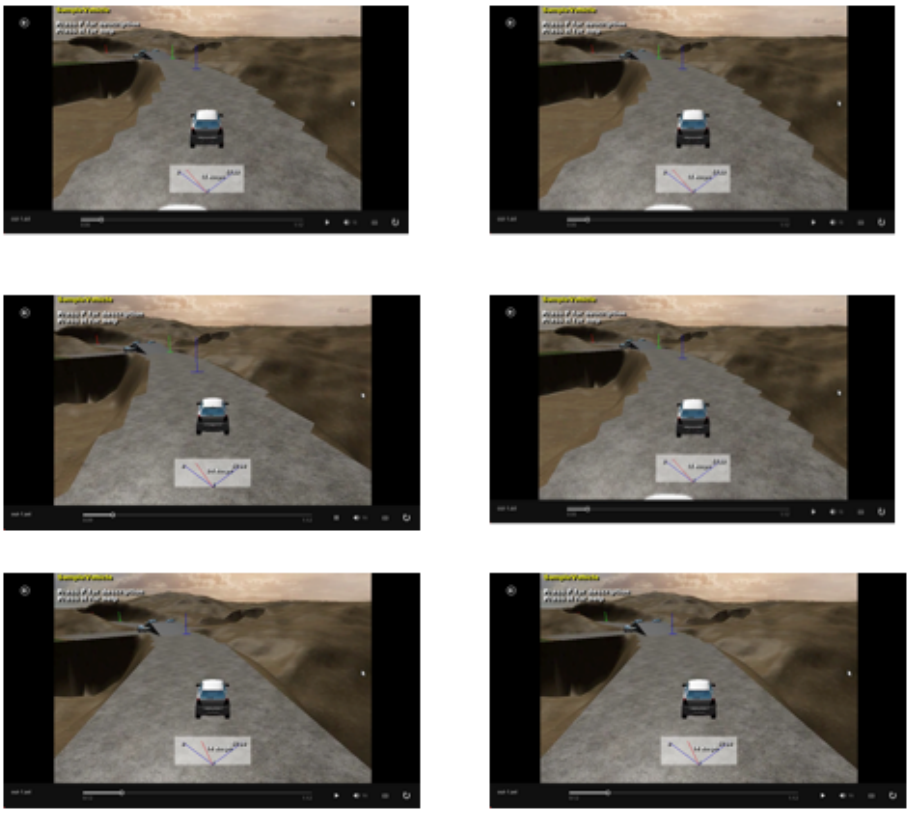

Fig. 5. The PhysX demo is used as both the real world and the virtual world, to develop the tracking capability.

Figure 5 shows the use of the PhysX demo. The left column is the real world and the right column is the rendered world. In the first pair of images, the cars start synchronized. In the second pair of images, the real car begins to move, and differences are created between the images. In this demo, the virtual camera follows the car, so the differences include differently shaped regions in the background of the images and along the borders of the road.

The differences are not centered on the car itself; this forces the system to reason about how the differences have been caused, and to attempt to register the backgrounds.

The larger dark brown area on the left of the real image indicates that it is closer, and the car has moved towards it, and the system registers the backgrounds and adjusts the velocity of the virtual car to match the real car, shown in the bottom image.

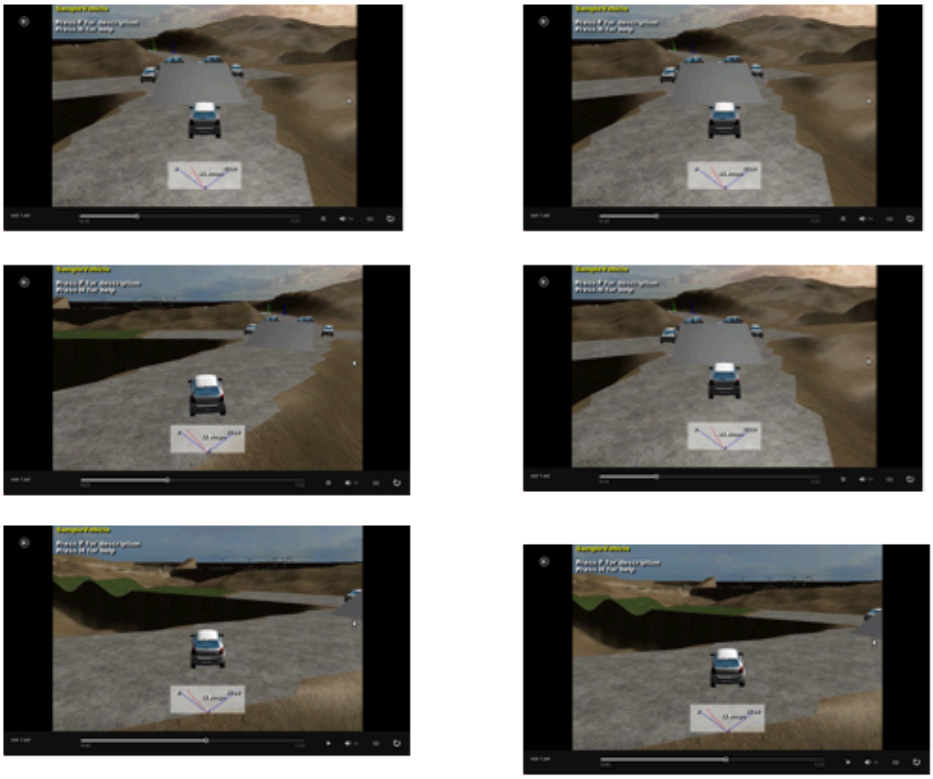

Fig. 6. Another example showing the car turning.

In Figure 6, we see another example, in which the real car turns left. In the middle pair of images, the differences caused by the turn are many. Once again, the reasoned searches to register the scenes, this time matching the cars in the background to derive the turn. The virtual car is repositioned and given the appropriate turning radius to match the real car at bottom.

Errors accumulate as the real car follows a path that is not exactly straight and the virtual car goes straight (similar to dead reckoning error). We ran a number of simulations and found that the vehicle's position needs to be updated about every 4 seconds on average. The computational effort required to keep the worlds registered was very small, less than 5% of total processing time.

This is not comparable to other data because nobody else seems to be doing this kind of activity modeling/comprehension, e.g. the KITTI database has no data on this.

4 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.

Further information on this work, including implementation details, can be found in [12,13]. 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. Oliva, A., and Torralba, A., The role of context in object recognition, TRENDS in Cognitive Sciences, **11**, No. 12 (2008).
2. Marr, D., Vision. W. H. Freeman, San Francisco (1982).
3. Hanson, A. and Riseman, E., VISIONS: A computer System for Interpreting Scenes, In Hanson, A., Riseman, E. (eds.) Computer Vision, New York: Academic Press (1978).
4. Csurka, G. et al. Visual categorization with bags of keypoints, In: Workshop on statistical learning in computer vision, ECCV. Vol. 1, (2004).
5. Mortensen, E., Deng, H., Shapiro, L., A SIFT Descriptor with Global Context, In: Int. Conf. on Computer Vision and Pattern Recognition, (2005).
6. Marques, O., Barenholtz, E., and Charvillat, V., Context modelling in computer vision: techniques, implications and applications, In: Multimedia Tools and Applications **51**:303-339, (2011).
7. Ungar, S., Cognitive mapping without visual experience. In Kitchin, R. & Freundschuh, S. (eds), Cognitive Mapping: Past Present and Future, London: Routledge.
8. Shanahan, M.P., A Cognitive Architecture that Combines Internal Simulation with a Global Workspace, In: Consciousness and Cognition, **15**, pp. 433-449, (2006).
9. Pezzulo, G., et al., The mechanics of embodiment: a dialog on embodiment and computational modeling, In: Frontiers in Psychology, **2,** A5, January (2011).
10. Rayner, K., Eye movements and cognitive processes in reading, visual search, and scene perception, In J. M. Findlay, R. Walker, & R. W. Kentridge (eds.), Eye Movement Research: Mechanisms, Processes, and Applications, (pp. 3-21. New York: Elsevier, (1995).
11. N. Barnes and Zhi-Qiang Liu, Embodied computer vision for mobile robots, In: ICIPS '97. pp. 1395 – 1399, 1997, DOI: 10.1109/ICIPS.1997.669238 (1997).
12. Benjamin, D. P., Lyons, D., Monaco, J. V., Lin, Y., and Funk, C., Using a Virtual World for Robot Planning, In: SPIE Conference on Multisensor, Multisource Information Fusion, http://csis.pace.edu/robotlab/pubs/SPIE2012.pdf (2012).
13. Lyons, D.., Nirmal, P., and Benjamin, D. P., Navigation of Uncertain Terrain by Fusion of Information from Real and Synthetic Imagery, SPIE Conference on Multisensor, Multisource Information Fusion. http://csis.pace.edu/robotlab/pubs/LyonsNirmalBenjamin2012.pdf (2012).