Generalization of Figure-Ground Segmentation from Binocular to Monocular Vision in an Embodied Biological Brain Model

Brian Mingus, Trent Kriete, Seth Herd, Dean Wyatte,
Kenneth Latimer, and Randy O'Reilly

Computational Cognitive Neuroscience Lab
Department of Psychology
University of Colorado at Boulder
Muenzinger D244, 345 UCB
Boulder, Co, 80309, USA
{brian.mingus,trent.kriete,seth.herd,dean.wyatte,
kenneth.latimer,randy.oreilly}@colorado.edu
http://grey.colorado.edu/ccnlab

Abstract. Humans have the remarkable ability to generalize from binocular to monocular figure-ground segmentation of complex scenes. This is clearly evident anytime we look at a photograph, computer monitor or simply close one eye. We hypothesized that this skill is due to of the ability of our brains to use rich embodied signals, such as disparity, to train up depth perception when only the information from one eye is available. In order to test this hypothesis we enhanced our virtual robot, Emer, who is already capable of performing robust, state-of-the-art, invariant 3D object recognition [1], with the ability to learn figure-ground segmentation, allowing him to recognize objects against complex backgrounds. Continued development of this skill holds great promise for efforts, like Emer, that aim to create an Artificial General Intelligence (AGI). For example, it promises to unlock vast sets of training data, such as Google Images, which have previously been inaccessible to AGI models due to their lack of embodied, deep learning. More immediately practical implications, such as achieving human performance on the Caltech101 object recognition dataset [2], are discussed. [1]

Keywords: figure-ground, neural network, object recognition, embodiment, vision, depth perception

[1] Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the U.S. Army Research Office contract number W911NF-10-C-0064. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, the U.S. Army Research Office, or the U.S. Government.

1 Introduction

When we look at a photograph our mind perceives the objects as if they existed right in front of us in three dimensional space. This is surprising since each eye conveys the same image of the photograph with no useful disparity signals. One can demonstrate this to themselves by looking at a photograph with one eye closed and noting the rich perception of depth. So too is our depth perception intact when we perceive the world more generally with only one eye open. In normal binocular viewing conditions the disparity between objects in the two eyes helps us to compute their depth, but it is rather remarkable that we can continue to do this in lieu of this cue.

1.1 Genetic vs. Learned Monocular Figure-Ground Segmentation

How does the brain perform this remarkable feat? One possibility is that the skill evolved and the expression of the requisite neural circuitry is hard-wired into our genome. This seems rather unlikely as it is hard to imagine an environment that consistently caused humans (or their ancestors) to lose an eye, constituting a strong and consistent evolutionary selection pressure.

Experiments performed with Emer, our virtual robot, have led to the more parsimonious hypothesis that this skill is learned. Evidence comes from a variety of sources, such as [3], who found that in the case of isoluminance across figure and ground the luminance cue breaks down and gestalt contours can fail to pop out. In this case we rely on color, which, having weak stereopsis, is a monocular cue. We hypothesize that the brain learns how to map this and other cues from only one eye onto the standard binocular 3D depth map.

Additional evidence comes from [4], which reviews the evidence for innate vs. learned depth perception in a large number of species. Although depth perception is innate in some species, such as the rat, in others, such as humans, it is very poor at birth and advances along with the infant's sensorimotor coordination and, perhaps, conceptual knowledge of the world. Notably, many species in which vision develops quickly or is innate, such as the rat, have very poor visual acuity, while human vision is rather sharp. This is potentially due, in part, to the longer learning curve human vision undergoes. We hypothesize that this longer learning trajectory in humans is due to the need to slowly incorporate a vast array of signals into vision, and likewise, to incorporate vision into an extremely complex brain. Indeed, as we have recently shown, human representations of objects in IT cortex are physically changed by conceptual knowledge, allowing us to make certain conceptual generalizations at the speed of visual object recognition [1]. The need to incorporate conceptual knowledge in such a deep way seems to slow the progression of visual learning of all kinds, including depth perception and, by extension, monocular figure-ground segmentation.

1.2 Idealized Training of Monocular Figure-Ground Segmentation

An idealized method of training such a network follows directly from the description of the problem. An infant looks at an object with one eye open, predicts

its depth, and then opens its other eye. Learning occurs as a function of the difference between the predicted and actual depths.

In terms of a neural network, there are two input layers representing the primary visual cortex (V1) neurons for the left eye and right eye, respectively. These map onto a layer which computes focal disparity, that is, the zero-disparity region of foveation. The computation of disparity in the case of binocular vision follows from straightforward geometry (see the horopter in [5]). During training the information from one eye is removed and the network is asked to predict the depth map of the scene. After making a guess based on monocular cues, the the other eye is returned and the weights are changed based on the difference between the predicted depth map and the actual depth map. While such a simple network only provides marginal figure-ground segmentation ability, it clearly demonstrates the point that we hope to make with Emer: that rich 3D signals can serve as a training signal for figure-ground segmentation with 2D cues.

1.3 The Relevance of Embodiment

A notable demand of this disparity-based training paradigm is that it requires an embodied agent. It would not be possible to compute the target 3D depth map if the agent did not have two offset eyes. Just having eyes is not always sufficient for depth perception, however. For example, human skill at depth perception correlates with the ability to crawl [4]. Infants make predictions about how far away something is and translate that guess into a goal in motor coordinates. After the infant crawls to the object it knows the actual distance and can translate that motor error back into visual error by proxy of parietal cortex, thus improving depth perception. While we would eventually like to bootstap the learning of the binocular 3D depth map itself in such a manner, which would then bootstrap monocular depth perception, for now we assume the existence of trained binocular depth perception and focus on generalizing it to the monocular case.

2 Materials and Methods

Experiments were conducted using the emergent Neural Network Simulation System [6]. The Leabra neural network architecture and learning rule was used for all simulations [7].

2.1 Emer

Our simulated robot, Emer (Fig. 1), is implemented using the Open Dynamics Engine rigid body physics simulator [8] and the Coin3D 3D Graphics Developer Kit [9]. Each of his eyes is a camera, and their offset positions on his head give him slightly different views of objects, facilitating stereo vision. The eye-beams projecting from his eyes give a sense of where he is foveating. Emer's full brain includes a superior colliculus and a cerebellum, with which he can learn to turn

Fig. 1. Our virtual robot, Emer. His name is based on "emergent", our neural network simulator. Seen here are his torso, head, eyes, eye-beams, and the fish that he is foveating in prepration for object recognition.

his head and foveate on an object. In this simulation, however, we have disabled these parts of his brain and have instead hard coded foveation in order to make the simulation faster.

2.2 CU3D-100 dataset

To test the sufficiency of our model on a realistic, challenging version of the object recognition problem, we used our dataset of nearly 1,000 3D object models from the Google SketchUp warehouse (the *CU3D-100* dataset [10]) organized into 100 categories with an average of 9.42 exemplars per category (Fig. 2a-d). Two exemplars per category were reserved for testing, and the rest were used for training. Objects were rendered to 20 bitmap images per object with random $\pm 20°$ depth rotations (including a random 180° left-right flip for objects that are asymmetric along this dimension) and overhead lighting positioned uniformly randomly along an 80° overhead arc. These images were then presented to the model with planar (2D) transformations of 30% translation, 20% size scaling, and 14° in-plane rotations. The CU3D-100 dataset avoids the significant problems with other widely-used benchmarks such as the Caltech101 [12], by ensuring that recognition is truly robust to significant amounts of invariance, and the 3D rendering approach provides full parameterization over problem difficulty.

2.3 Structure of the LVis Model

The LVis (Leabra Vision) model [1] (Fig. 3) preprocessed bitmap images via two stages of mathematical filtering that capture the qualitative processing thought to occur in the mammalian visual pathways from retina to LGN (lateral geniculate nucleus of the thalamus) to primary visual cortex (V1). The output of this filtering provided the input to the Leabra network, which then learned over a sequence of layers to categorize the inputs according to object categories. Although we have shown that the early stages of visual processing (through V1) can be

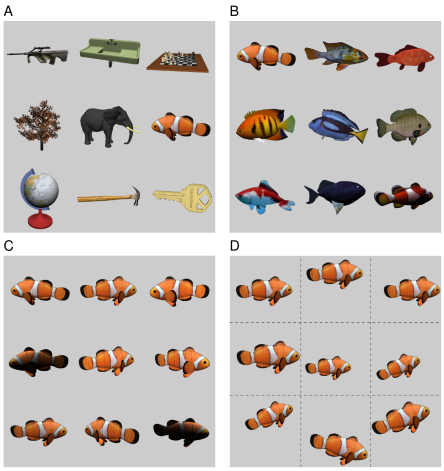

Fig. 2. The CU3D-100 dataset. **a)** 9 example objects from the 100 CU3D categories. **b)** Each category is further composed of multiple, diverse exemplars (average of 9.42 exemplars per category). **c)** Each exemplar is rendered with 3D (depth) rotations and variability in lighting. **d)** The 2D images are subject to 2D transformations (translation, scale, planar rotation), with ranges generally around 20%.

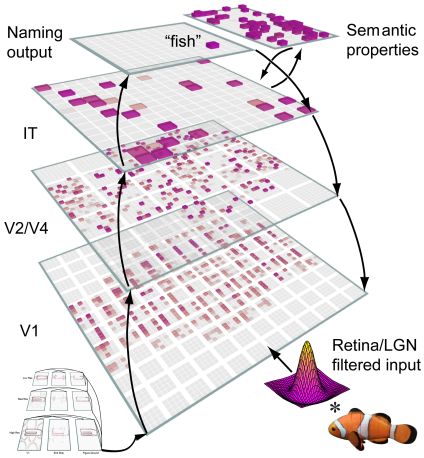

Fig. 3. The architecture of the LVis model [1]. LVis is based on the anatomy of the ventral pathway of the brain, from primary visual cortex (V1) through extrastriate areas (V2, V4) to inferotemporal (IT) cortex. V1 reflects filters that model the response properties of V1 neurons (both simple and complex subtypes). In higher levels, receptive fields become more spatially invariant and complex. All layers are bidirectionally connected, allowing higher-level information to influence bottom-up processing.

learned via the self-organizing learning mechanisms in Leabra [6], it was more computationally efficient to implement these steps directly in optimized C++ code. This optimized implementation retained the k-winners-take-all (kWTA) inhibitory competition dynamics from Leabra, which we have found important to for successful recognition performance. Thus, the implementation can be functionally viewed as a single Leabra network.

2.4 Details of the Figure-Ground Model

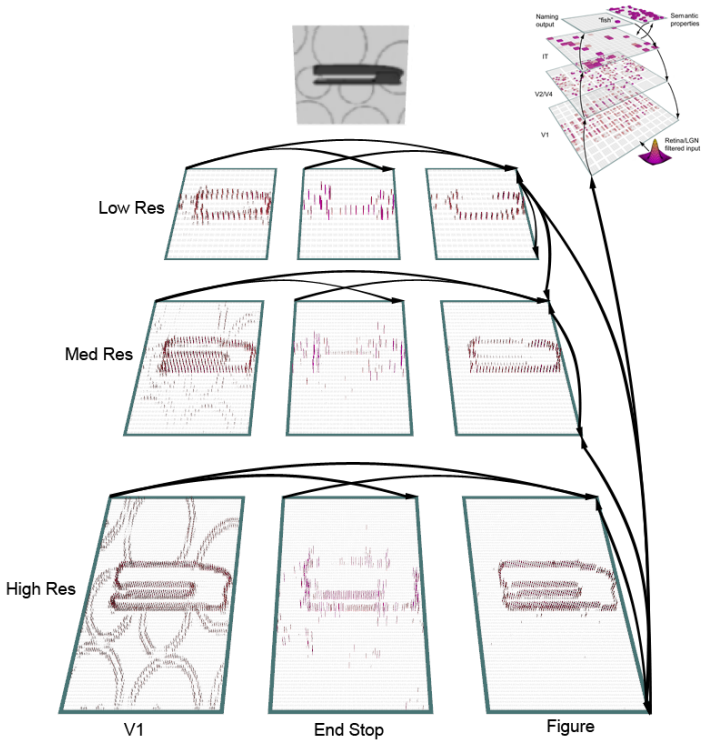

Fig. 4. The figure-ground segmentation model. There are three sets of layers at three interacting spatial resolutions. The first set corresponds to V1, the second set to V1C end-stop cells [11], and the third set learns to extract the figure from the background. The network is connected in a feedforward fashion from left to right. All of the figure layers have both recurrent and bidirectional connectivity, allowing the network to learn how to combine the information from each resolution so as to reproduce just the figure in the high resolution figure layer. The high resolution figure layer then projects to the V1 layer of the LVis model, helping the figure to pop-out in object recognition.

The figure-ground segmentation model (Fig. 4) has three feedforward streams, going from left-to-right, that each operate at a different resolution. The rightmost figure column comprises the output stages of the network, with the high resolution figure layer being the final output. The principle of operation of the network can be clearly seen by comparing the low and high resolution V1 layers with the high resolution figure layer. At low resolution the background falls out out completely but much of the fine detail of the stapler is lost. At high resolution the fine detail is retained but so too is the background. In the high resolution figure output layer it is evident that the network has learned to be constrained by both sources of information and produces a stapler with fine detail and no background.

Learning, Training and Testing During training Emer is presented with 3D objects against complex backgrounds from the CU3D-100 and is asked to identify them by name. Generalization performance is then assessed on the remaining two objects per category. The training signal is derived from our disparity matching algorithm which is written in optimized C++ code. In order to compute the training signal Emer first foveates an object, resulting in stereo images. Our disparity matching algorithm then compares the images and computes the zero-disparity region that is used as a training signal. During training the network predicts what the figure is in each of the figure layers at its corresponding resolution. Next, the weights are changed as a function of the difference between this prediction and the correct answer as output by our disparity matching algorithm. The network is then tested on its ability to predict this training signal for objects it has never seen before.

V1 and End-stop Cells The middle column of Fig. 4 contains end-stop cells, which are essential to solving the figure-ground problem [11]. End-stop cells detect when one oriented edge in V1 terminates into another, enabling the detection of T-junctions and contours, which are excellent cues for figure-ground segmentation. This can be seen visually by noting that the end-stop layers are active for the figure with little activity for the background. These weights are not learned, but are rather fixed, with each end-stop cell integrating over several rows of V1 neurons.

Both V1 and end-stop cells have feedforward projections to the figure layers. Each neuron in V1 projects to its corresponding neuron in the figure layer. Each group of neurons in the end-stop layer connects to every other neuron in the corresponding group of neurons in the figure layer. The weights in these projections are learned over training.

Recurrent and Bidirectional Connectivity Between Figure Layers Recurrent projections connect each figure layer to itself. These projections are intended to support the continuation of contours, by helping to make edges robust to differences in bottom-up strength. This principle of good continuation is implemented by having each neuron connect to neurons of similar orientation in

neighboring unit groups. The weights are initially randomized, but are then tuned over learning.

Bidirectional projections exist between all figure layers, which allows the network to learn about coarse-to-fine and fine-to-coarse interactions. This is implemented by having each neuron in the figure layers receive from a small patch of neurons in the other figure layers at approximately the same spatial location.

Interactions with the LVis Object Recognition Model The figure-ground model and the LVis object recognition model interact by having the high-resolution figure layer influence the V1 layer of the object recognition model. This is achieved by comparing the activations of each unit in the figure layer with the corresponding unit in LVis V1 and applying the smaller of the two activations to LVis V1. Because the LVis V1 layer sees the figure against a background, and the figure layer learns to ignore the background, this has the effect of damping the background in the object recognition pathway, improving performance in object recognition against a complex background. This "min" operation occurs after the first five cycles, and then periodically throughout settling if the activity in the figure layer has changed by a sufficient amount.

3 Results

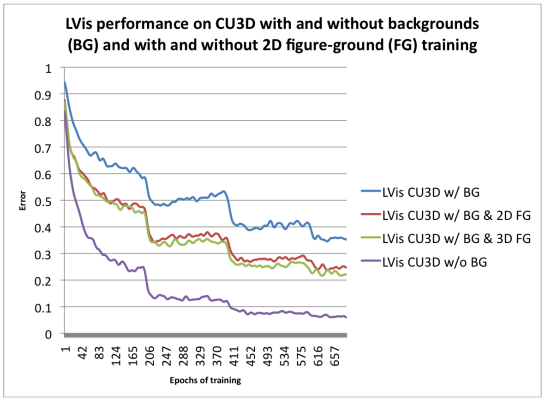

Fig. 5. Generalization performance of LVis in four object recognition conditions. **w/ BG:** With backgrounds and without figure-ground training error asymptotes at 35.3%. **w/ BG & 2D FG:** With backgrounds and with the learning figure-ground front-end intact performance asymptotes at 24.7% error. **w/ BG & 3D FG:** With backgrounds and using the target depth map as input into LVis (i.e., no 3D to 2D generalization - this is the best possible case for the previous condition) performance asymptotes at 22.2% error. **w/o BG:** Without backgrounds using just the standard LVis model performance asymptotes at 6% error.

All of the conditions in Fig. 5 have the same basic task, which is invariant object recognition on the CU3D-100 dataset. The model is trained on approximately eight exemplars per category and then generalization performance is tested on the remaining two objects from each category. Two objects for each of the 100 categories are held out for testing and performance is computed as the number of errors divided by 200 (as seen in Fig. 5).

The performance of the learned monocular figure-ground segmentation is compared to several other conditions in Fig. 5. The key comparison conditions are standard LVis with no backgrounds, LVis with backgrounds and without figure-ground segmentation and LVis with the best 3D figure-ground segmentation that our disparity matching system can compute.

The main condition being tested is object recognition against a background (such as the one seen in the picture in Fig. 4) with the learned monocular figure-ground segmentation model in place. To demonstrate that this is a hard problem, note that the difference in performance between LVis with and without backgrounds (and without figure-ground segmentation) is 29.2% error, which is a dramatic decrease in performance. The other key comparison is between the model that uses the computed disparity signal (and thus does not need to generalize from 3D to 2D) versus the learned monocular figure-ground segmentation model. The monocular model has only 2.4% more error, a relatively slight difference.

4 Discussion

We chose to start with relatively simple backgrounds that nonetheless resulted in a dramatic detriment to performance in object recognition. The monocular figure-ground segmentation system had only 2.4% more error than it possibly could have, demonstrating that the model does indeed learn how to segment figure from ground.

We believe this result merits continued effort and has great promise for the field of AGI. For example, it has recently been argued that the field of computer vision has become stuck using narrow AI methods to do what is effectively a sophisticated form of pattern recognition in lieu of true invariant object recognition [12]. A main crux of this argument is that many state-of-the-art object recognition datasets, such as the Caltech101 [2], do not provide a rich enough training signal to truly develop invariance over scale, translation, and 3D orientation. Therefore, models which attempt to do object recognition on this dataset are effectively limited to doing an advanced form of pattern recognition because the training data needed for true invariant object recognition is simply not there.

This has led our lab [13] to develop a new dataset of nearly 1000 objects in 100 categories based on 3D models derived from the Google Sketchup 3D Warehouse. We have recently used this dataset, known as the CU3D-100, to test that our model of biologically plausible object recognition [1] solves the hard problem of invariant object recognition as opposed to pattern recognition.

However, we believe that the Caltech101 is not in itself intrinsically flawed. Rather, researchers are going about solving the problem in the wrong way and avoiding the hard problem. Because the images in the Caltech101 are derived from Google Images they have arbitrary backgrounds, requiring figure-ground segmentation at some level. We believe the correct way to solve the Caltech101 problem, and to solve the problem of training on the entirety of Google Images in general, is to first create an embodied agent that can learn to do invariant object recognition based on 3D models. Next, the agent should learn to do invariant object recognition against complex backgrounds using its ability to do monocular figure-ground segmentation. Finally, the embodied agent can be tested on the Caltech101 and will stand a chance of getting closer to human performance on the task.

5 Future Work

Achieving human performance on the Caltech101 is a challenging problem. In the context of the model presented here, we have only used disparity signals as cues to depth. The 2D monocular system can only learn to be as good as its training signal, and so a richer signal is needed to get to the performance of LVis without backgrounds. Our next step is therefore to investigate ways to improve the 3D disparity signal.

One well-known mechanism that animals use is disparity-from-motion. It is possible that motion not only acts as a depth cue, but can train the 3D disparity system as to the depth in the scene, which can then in turn train the 2D disparity system. Our lab has already begun initial investigations into this mechanism and the results are promising. Ultimately, it seems likely that the brain uses a multitude of signals to converge on monocular figure-ground segmentation that is nearly as good as binocular, and it will be some time before our model reaches that level.

References

1. O'Reilly, R., Wyatte, D., Herd, S., Mingus, B., Jilk, D.: Bidirectional Biologically Plausible Object Recognitino. In Press. (2011)
2. Caltech101, `http://www.vision.caltech.edu/Image_Datasets/Caltech101/`
3. Troscianko, T., Montagnon, R., Le Clerc, J., Malbert, E., Chanteau, P.: The role of colour as a monocular depth cue. Vision research. 1923–1929 (1991)
4. Walk, D.: The Development of Depth Perception in Animals and Human Infants. Concept of Development: A Report of a Conference Commemorating the Fortieth Anniversary of the Institute of Child Development. (1966)
5. Wheatstone, C.: Contributions to the physiology of vision.–Part the First. On some remarkable, and hitherto unobserved, phenomena of binocular vision. Philosophical Magazine Series 4. (1852)
6. Aisa, B., Mingus., B, O'Reilly, R.: The emergent neural modeling system. Neural Networks. 1045–1212 (2008)

7. O'Reilly, R.: The Leabra Model of Neural Interactions and Learning in the Neocortex. PhD Thesis. (1996)
8. Open Dynamics Engine, http://www.ode.org/
9. Coin3D 3D Graphics Engine Developer Kit http://www.coin3d.org/
10. CU3D dataset http://grey.colorado.edu/CompCogNeuro/index.php/CU3D
11. Yazdanbakhsh, A., Livingstone, M.: End stopping in V1 is sensitive to contrast. Nature Neuroscience. (2006)
12. Pinto, N., Cox, D., DiCarlo, J. Why is real-world object recognition hard? PLoS Computational Biology. (2008)
13. Computational Cognitive Neuroscience Lab http://grey.colorado.edu/ccnlab