Biologically Inspired Foveated Robot Vision

Wallace Lawson, Keith Sullivan, Chris Bradel, Olu Roy, Esube Bekele
Naval Research Laboratory, Washington DC

Abstract
One of the biggest challenges in robot vision is how to efficiently make decisions from a stream of high resolution images. There is simply too much information for a robot to examine, and the robot may not have the available resources (e.g. GPUs) to analyze such a large stream of data. We propose to solve this problem using a foveated approach to vision. We show preliminary results on terrain classification using a small convolutional neural network.

Introduction
Consider a robot autonomously navigating outdoor paths primarily intended for people. In this environment, a robot must solve many interrelated problems from reasoning to perception. Here, we focus on a particular part of this problem: terrain classification. It’s a challenging problem: the pixel throughput makes it difficult to process all of the data that is available. Visually, there’s too much to focus on all at once, and yet this is a commonly used approach in deep learning based robot navigation systems (Traver and Bernardino 2010; Sullivan and Lawson 2017). These networks use complex attention mechanisms, where they look at every pixel in the image using a uniform receptor field before making a decision. In practice, this has several important implications. First, it places a significant processing burden on a resource limited robot. To process in real-time, it is necessary for these robots to use power-hungry GPUs. Second, it also makes little sense to require the robot to process the entire scene with a uniform receptor field.

Here, we propose to solve this problem using an approach that is inspired by nature. In this paper, we foveate on select parts of the image using a log-polar representation that is motivated by the retina. The retina has a high concentration of photoreceptors in the center of the field of view, with a lesser density farther away. This complex logarithmic retinotopic mapping (Schwartz 1977) results in a region of 1-5° at the center of the visual field observed at a higher resolution, with the remainder of the visual field seen with much less detail. We use a log-polar representation that mimics the retina.

Log Polar Representation
We briefly summarize the log-polar representation here, and refer the interested reader to (Araujo and Dias 1996; Traver and Bernardino 2010) for a thorough discussion of the log-polar representation as well as a discussion of other similar foveation strategies. The log-polar representation maps the retinal plane \((\rho, \theta)\) onto a cortical plane \((\log(\rho), \theta)\), such that points that are closer to the center are sampled more densely than points further than the center (visually, see figure 1). The image is sampled using a set of concentric circles that have been divided into a fixed number \(N_{ang}\) of angles. The image is divided both by the the angles and by the circles, and the pixels inside each region are averaged together. This produces a higher resolution in the middle of the image and a lower resolution outside.

Classification
We train a deep convolutional neural network to classify terrain types The network itself is simple enough to permit deployment to a SWAP constrained robot: it is composed of 2 convolutional layers, 2 fully connected layers, and a classification layer. The convolutional layers are \(5 \times 5\) with a stride of 2; the first convolutional layer has 6 output channels and the second has 16 output channels. The fully connected layers have 120 neurons, 84 neurons, and finally 3 neurons for the classification layer. Prior to processing by the neural network, we resize the input to a size of \(32 \times 32\) using a bilinear interpolation sampling strategy.
First, the human visual system has a foveation that we are not addressing in this paper. There are several other important points to note. Substantially improve our ability to classify different types of terrain. The roads are painted and older paint can be faded. There are also imperfections such as cracks in the asphalt. Older roads tend to have a different color than newer roads. The increasing availability of very efficient neuromorphic hardware platforms makes this a very appealing option for resource constrained mobile robots. The primary challenge is to retain a similar level of accuracy as what is available in a convolutional neural network.

Second, foveation is highly coupled with fixation and saccades. It is difficult to discuss one without the other. We have used a very simple fixation strategy, where we systematically fixate on all points across the image. This is not at all similar to the human visual system (Bandera and Scott 1989). The human approach to fixation is driven both by task and by saliency. That is to say, the regions where we fixate will vary depending on what we are doing and what is present. In this task, for example, we may be looking mostly in front of us, although this region may expand as we look further away from the camera. The robot likely should alternate between these fixation points as it navigates. Fixation should also be driven by highly salient regions, such as areas with large contrast or with motion. This may be highly advantageous as these may indicate regions that are important and areas that should be avoided (e.g., a person that is present).

Finally, in future efforts we propose to use a spiking neural network with a similar structure to analyze the different types of terrain. The input data was collected on a small robot that was manually controlled; the robot moved on surfaces that were concrete, asphalt and grass. Each terrain was labeled manually. For this experiment, the robot systematically fixates on regions on the image at regular steps. We compare two strategies: cropping small regions around the fixation point and foveating around the fixation point using a log-polar representation. That is, in the former we consider a uniform receptor field and in the later we considered the non-uniform log-polar receptor field. Training and testing data were collected on two separate days. We trained the each network 4 times, table 1 shows the average of testing each of these networks.

These results tell a very interesting story. It seems that for some surface types (e.g., grass vs concrete), the problem is simple enough that cropping is sufficient. However, the network has a difficult time on asphalt. One interesting thing to note is that asphalt has a significant amount of variation. Older roads tend to have a different color than newer roads. The roads are painted and older paint can be faded. There are other features such as crosswalks. There are also imperfections such as cracks in the asphalt.

One hypothesis for why foveation performs better on asphalt is due to the re-sized cropped image. Texture is an important feature and a simple way to classify asphalt. When an image is resized smaller, small details will become harder to see. It’s difficult to see the texture in the asphalt image, particularly when the image has been downsized uniformly. By foveating on the center of the image, we produce a sufficient amount of texture that the network is able to properly determine the material type. Images of foveated and cropped asphalt regions are in figure 2.

## Discussion

Although the results are preliminary, foveation appears to substantially improve our ability to classify different types of terrain. There are several other important points to note about foveation that we are not addressing in this paper. First, the human visual system has a $1 – 5^\circ$ fovea, which means that we foveate on a larger region when we are looking at things that are close to us vs. when we are looking at things that are farther away. Depth information, therefore, would be a valuable addition as it would permit us to foveate on regions differently depending on how far away they are from the camera.

Figure 2: Cropped image (left) and foveated image (right)

### Preliminary Experimental Results

The increasing availability of very efficient neuromorphic hardware platforms makes this a very appealing option for resource constrained mobile robots. The primary challenge is to retain a similar level of accuracy as what is available in a convolutional neural network.

### Acknowledgements

Wallace Lawson and Keith Sullivan were supported by the Office of Naval Research. Esube Bekele was supported by the National Research Council.

### References


