Computational Modeling of Scene-selective Visual Neurons in area LIP (Lateral Intraparietal)

Jung Uk Kang

A Thesis Submitted to
Brown University in Partial Fulfillment
of the Requirements for the Honors Degree
of Bachelor of Science

Department of Computer Science
Brown University

May 2, 2013

1 Introduction

Processing of visual information in the neural circuits of the brain has been a fundamental question to neuroscientists. Researchers now understand that a visual stimulus first arrives at the neural system via the retina and is transmitted to lateral geniculate nucleus (LGN) of the thalamus to the early visual areas of the brain. Yet, during the visual attention process, different regions of the brain contribute to make decisions on which to attend, and additional regions also support object recognition. The interactions between these different regions of the brain contributing to the visual attention are still under investigation.

For computer scientists researching methods of computer vision, visual cognition performance of humans has been regarded as an upper-bound for any kind of computer vision system. Several models of biological vision have yielded promising results in visual perception, and the methods can be even better if computer scientists can have a more accurate map of visual perception of the brain. On the other hand, it is also possible that certain computer vision methods contribute to a better understanding of the processing of visual information in the neural circuits. Performance of computational models predicting a neural response with the given input of computer vision features can be evaluated using statistical analysis methods.

Therefore, a synthesis of computer science and neuroscience will enable scientists in both of the disciplines to gain a better understanding of computer vision methods and the visual system of the brain. Especially, the question of how the brain performs object recognition and makes visual attention would not only give an answer to the fundamental

question to neuroscientists but also for computer scientists to build a more accurate object detection method.

Compared to computer vision, biological vision has a mechanism of visual attention to deploy resources for visual perception. Because there are resource constraints on visual processing, the visual system needs to optimize its perception pathway by selectively attending to salient locations. In the visual system of the brain, the frontal eye field (FEF) contributes to visual attention, and the inferotemporal (IT) region contributes to object recognition. According to Monosov *et al.* [7], neural activity in FEF precedes neural activity in IT, and therefore visual attention precedes object recognition. Therefore, visual attention is an important process in which the visual system increases its efficiency.

One of the brain regions known to contribute to the visual attention process is the Lateral Intraparietal (LIP) area. Bisley and Goldberg [1] proposed that area LIP constructs a map where objects are represented by their "behavioral priority" [1] or their saliency information. The "priority map" [1] is later used by the visual system to make rapid eye movements and visual attention. Previously, computational modeling of visual areas such as V1, V2, and V3 has been completed to investigate their neural activity triggered from different visual stimuli, but a computational model that is able to accurately predict both neural activity and visual attention process of the brain has not been completed yet.

Kay *et al.* [6] developed a decoding method based on quantitative receptive-field models that characterize the relationship between visual stimuli and fMRI (functional Magnetic Resonance Imaging) activity in V1, V2, and V3. fMRI method measures change in blood oxygenation in the brain, and the localized hemodynamic response is

relevant to neural activity of the brain region. Kay *et al.* [6] first estimate a receptive-field model for each voxel using a set of images and its corresponding neural activities and identify or predict the image presented by their given voxel activity pattern.

Serre *et al.* [11], on the other hand, developed a feature detector that is tolerant to scale and position of a visual scene. This model is based upon the ventral stream of visual cortex of the brain. According to Pinto *et al.* [9], the model performs better object detection than conventional computer vision models such as SIFT, SLF, PHOG, and PHOW. Though the final goal of computer vision techniques has been to train computers to visually recognize scenes as humans do, Serre *et al.* [11] developed the first model that is fundamentally based upon the visual system of the human brain.

Another direction that both computer scientists and neuroscientists have investigated is to compute salient locations of an input image to predict where humans will give visual attention of a certain image or video. Itti and Koch [4] developed a visual saliency model based on intensity, color, and orientation, and a diagram of the model is illustrated in Figure 1. Feature maps are extracted from the input image at several spatial scales and are combined into three separate conspicuity maps. Judd *et al.* [5] used eye-tracking device to compare the prediction of the saliency model and the actual human eye fixations. Judd *et al.* [5] produced successful prediction of human fixation points through using low-level (intensity, orientation, and color contrast), mid-level features (gist), and high-level features (face detector and person detector).

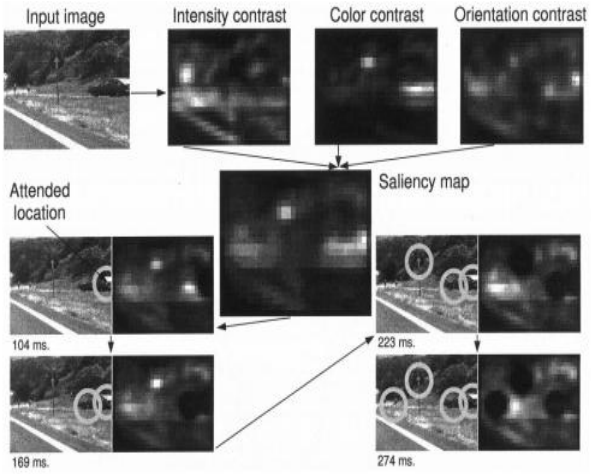


Figure 1. The visual saliency model by Itti and Koch [4] predicts attended locations based on color, orientation, and intensity.

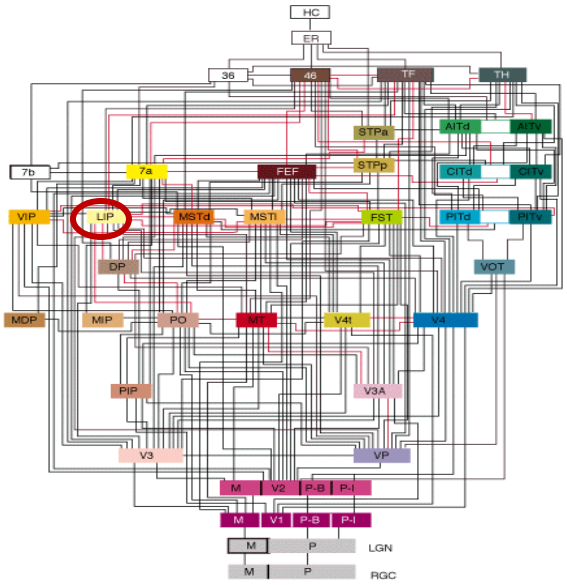


Figure 2. Connectivity diagram of the primate visual system by Felleman and Van Essen [3]. Neural activity from area LIP (circled in red) will be collected to investigate the visual attention process.

Because area LIP is known to contribute to contructing visual saliency map, neural acitivities collected after presenting different saliency map could give us a better understanding of the area. One important question is what kind of features does a LIP neuron respond to. The visual saliency model developed by Itti and Koch [4] is largely

based upon the COI channels: color, orientation, and intensity. Yet, visual attention can also possibily result from top-down information which contain higher-level features.

In our research, neural activities from area LIP of a primate are collected after presenting different natural scenes. Computer vision experiments conventionally use thousands of images to train a certain model, but, in single-cell neuron recording, stability of the neuron quickly decreases. As such, each dataset contains at most 200 images and their corresponding neural activities and saccade points. The final goal our research is to be able to predict both neural activity and saccade points of a certain natural scene.

2 Neural Activity in area LIP

An electrode to collect neural activity from area LIP is implanted through craniotomy after verifying the accurate location of the area by MRI scan. After implanting the electrode, the location was reaffirmed after testing the physiological characteristics of the cells. For visual stimuli, static natural scenes from the SUN Database by Xiao *et al.* [12] are presented to primates for 250ms. Neural data was collected from 25ms to 75ms. Both neural activity and saccadic eye movement positions are collected in each trial. To minimize the noise of the data collected, we repeated the neural measurements for 10 times and computed the mean value to use for our analysis.

3 Methods

3.1 Computer Vision

To train computers to recognize visual characteristics of scenes based on neural activity from area LIP, we used conventional computer vision methods: GIST, spatial

pyramid, and SIFT. In our experiments, neural activity is our new scene category to be used for the models above.

GIST descriptor developed by Oliva and Torralba [8] represents an energy function of an input image, and it computes the function across 32 points (8 directions & 4 distances) in the image. Resolution of the descriptor can be changed by users, and for our initial steps we used a global representation as our descriptor to train the computer.

After producing GIST descriptors of the images used in our data, we used K-nearest neighbor method to find K nearest neighbors or images that have a similar GIST descriptor of the default image. Schematic of finding the three nearest neighbors of a certain image is provided in Figure 3.

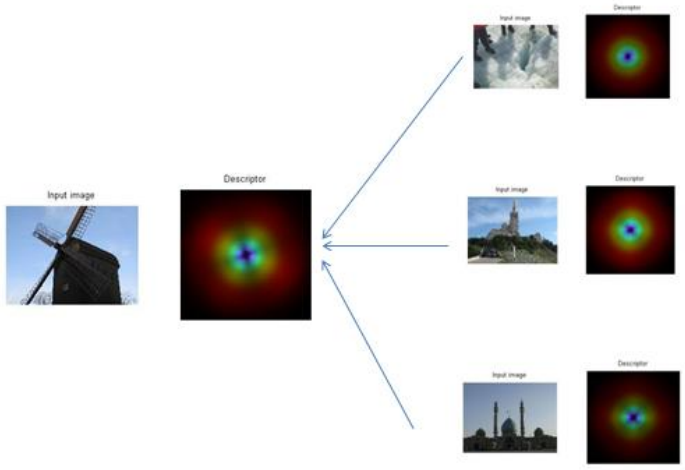


Figure 3. Three nearest neighbors of a certain image are found by calculating the distance from the default descriptor to the descriptors from the dataset

K-nearest neighbor method, however, did not yield GIST descriptor neighbors that are near to each other. An example of K-nearest method output of a neural dataset containing 90 images and their neural activity is shown in Table 1.

Input Image	K=3 Nearest Neighbors
Rank 1	Rank 29, 33, 50
Rank 30	Rank 53, 25, 83
Rank 60	Rank 79, 8, 74
Rank 90	Rank 23, 69, 11

Table 1. K=3 nearest neighbors of the rank 1, 30, 60, and 90 images from a dataset containing 90 images

K=3 nearest neighbors of the three images of the first dataset containing 90 images are shown in Table 1. Rank 1 is the image that produced the largest neural response. Rank 30 produced the 30^{th} largest neural response. The same rule applies for the other images on Table 1. Far distances between the K-nearest neighbors are the obstacles for predicting the better or more accurate neural responses in later steps. To evaluate the accuracy of the prediction by using K-nearest neighbor method is shown below. The error metric Error percentage is defined as the following.

$$\text{Error Percentage (\%)} = \frac{|\text{Prediction} - \text{Neural Response}|}{\text{Neural Response}} \times 100$$

Four experiments were completed to evaluate the accuracy of neural response prediction by using GIST descriptor.

1. K=3 nearest neighbor method (Global GIST Representation)
2. K=3 nearest neighbor method (4×4 GIST Representation)
3. Random match prediction (Randomly predict the response from the response after presenting another image)

4. Mean rate prediction (Always predict that the response is the average of the whole response in the dataset)

Overall, K=3 nearest neighbor method did not yield a statistically promising output. The prediction was not as near to the collected neural data as expected. Table 2 summarizes the results of experiments using K-nearest neighbor method.

	N = 90 Images	N = 148 Images
Nearest Neighbor (K=3)	43.03 %	42.03 %
Random prediction	46.65 %	43.73 %
Mean rate	38.33 %	39.40 %

Table 2. K=3 nearest neighbors method performance. The results are not statistically promising.

To confirm our observation from the experiments above, we implement two statistical testing: permutation test and rank correlation.

1) Permutation Test

We compare the error metric from K-nearest neighbor method and that from random permutation. For permutation test, we have two variables: D_i and D_i^b.

- D_i: Difference of neural activity data and means computed from 10 nearest neighbors.
- D_i^b: Difference of neural activity data and means computed from 10 permuted neural spikes.

- i: Total number of images in a dataset.
- b: Total number of permutations
- $S = \sum|D_i|$
- $S^b = \sum|D_i^b|$

The null hypothesis of the permutation test is the following.

$$H_0: \text{pdf}(S) = \text{pdf}(S^b)$$

The distribution of S^b and S are shown in Figure 4 as blue and red respectively. Because the p-value is significantly high, we conclude that the nearest neighbor method does not yield a statistically significant prediction.

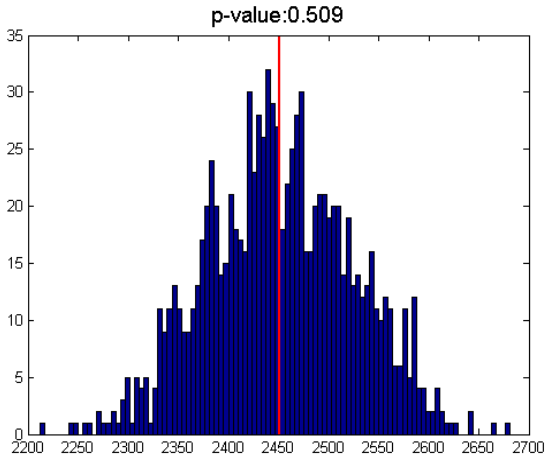


Figure 4. Result of Permutation Testing (K-nearest neighbor method)
Blue: S^b Red: S

2) Rank Test

To more robustly confirm whether there is a statistical relationship between the neural activities and the rank of the average from 10 nearest neighbors, we compute the correlation between the two rankings. The correlation value was 0.323.

From the permutation test and the rank test, we conclude that there is not a significant statistical relationship between the average from the nearest neighbors computed from GIST descriptor and the collected neural activity data.

3.2 Machine Learning

1) Non-liner SVM

K-nearest neighbor method was not robust enough to yield statistically significant prediction of neural activities. Another direction that we consider is implementing machine learning algorithm. Yet, while many machine learning algorithms aim to train computers to recognize categories, this research aims to predict neural activity in real numbers.

For SVM method, we set every integer from 60 to 100 as threshold and complete the training process with all images that produced neural spikes higher than the threshold. Instead of linear SVM, we use non-linear SVM that uses Gaussian radial basis function kernels. This decision was made empirically after using different methods of SVM for this project.

For testing process, we again use all images in the neural data and let the program classify the images that trigger neural spikes higher than the threshold. We use the same data for both training and testing because there are a limited number of images and neural spikes in this project. The results of classification after using non-linear SVM are provided in Figure 5.

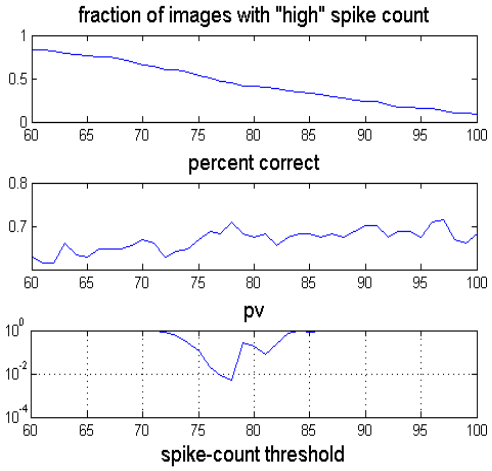


Figure 5. Result after implementing non-linear SVM method with Gaussian radial basis function kernel

A) Fraction of images with "high" spike count: Fraction if images that are higher than the threshold value from the neural spike 60 to 100

B) Percent correct: Percent of correct classification of images triggering "high" spike counts

C) pv = 1 - binomial cumulative distribution (number of correct classifications, number of total images in the data, percent correct). Minimum of pv is the error metric that will be used for statistical testing.

To confirm whether we achieved a statistically significant result, we implement another permutation test here. For permutation testing, we repeat the identical process of SVM with permuted spikes. Therefore, each image in the original dataset is now assigned with a different neural activity values. After comparing the error metric distribution from the permutation testing and that from the SVM method we used, we can conclude

whether we yielded a statistically significant prediction of neural spikes. The result after running permutation testing is shown in Figure 6.

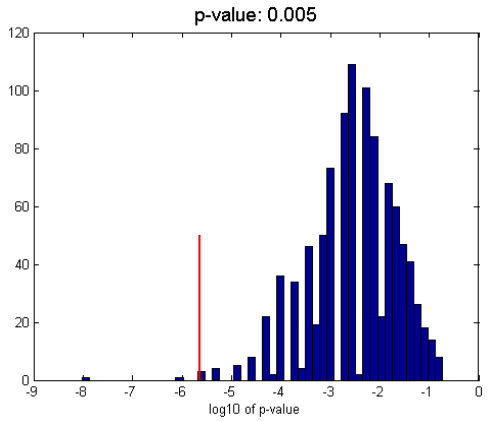


Figure 6. Blue: Distribution of error metric (Permutation Test) Red: SVM Method

Because the p-value of this test is 0.005, we can conclude that we reached a statistically significant prediction of neural spikes by using non-linear SVM with Gaussian radial basis function kernel.

2) Regression

As mentioned before, SVM method is usually used when classification of data is needed. In this research, however, binary classification such as “high neural activity” and “low neural activity” is not enough. We want to have a prediction that is as accurate as possible. Therefore, we perform a regression to predict the neural spike after presenting the images in the data. For regression, we regard all 32 elements of a global GIST descriptor as independent variables and simultaneously perform regression in 32 dimension space.

R^2 value, the error metric of regression, after running 32-dimension regression was 0.3091. To confirm whether this result is statistically relevant, we again perform permutation testing with the same procedures outlined before with regression method.

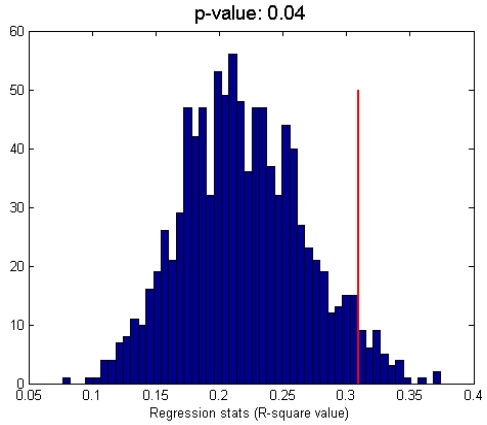


Figure 7. Blue: Distribution from permutation testing. Red: Error metric from our SVM method

The R^2 value from regression is higher than most of the R^2 values from permutation. Therefore, the regression method after using non-linear SVM with Gaussian radial basis function kernel does not yield a statistically significant prediction for neural activity.

3) Bag of Words Model

Li and Perona developed a Bayesian hierarchical model for recognizing natural scenes. The model creates a bag of "codewords "of certain computer vision features, and employs a probabilistic approach to detect a category of the natural scene such as "suburb," "highway," or "building." Because this model is known to be able to train a computer to recognize a given scene based on detailed representation of certain features, we decided to use this model to predict a neural activity of an input image. Initially, to

make our experiment similar to traditional computer vision experiments, we divided the data into 10 categories after sorting them by neural activity. An example of the classification experiment results is provided in figure 8.

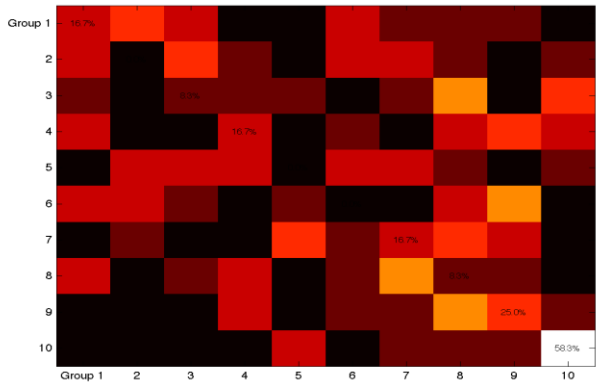


Figure 8. An example of classification experiment. (The data contains 90 images)
Average accuracy=0.15

Different from natural scene categories, neural activity is unstable because it contains biological noise in it. The classification accuracy was not high enough to conclude that this approach could give us a better prediction of neural activity. Additionally, to confirm whether the classification is accurate enough, we compared the classification results of randomized data after we ran the bag of words model using the GIST. The classification results of two datasets are the following.

	Dataset1 (Sorted by neural activity)	Dataset2 (Sorted by neural activity)	Dataset1 (Randomly sorted)	Dataset2 (Randomly sorted)
Accuracy (Using Bag of Words model with GIST)	5%	15%	7.5%	18.3%

Because the classification accuracy was lower than the classification results of randomized data, instead of dividing our data into 10 categories, we made the

classification task to be binary. The scheme for dividing the data into binary groups is the following. We label the two binary groups as “High neural activity” and “Low neural activity.”

1. Sort the images by their corresponding neural activity.
2. Divide the images into three groups by neural activity order.
3. Exclude the image group in the middle.
4. Use 90% of images for training and 10% of images for testing for the bag of words model.

A subset of the binary groups used for testing is shown below.

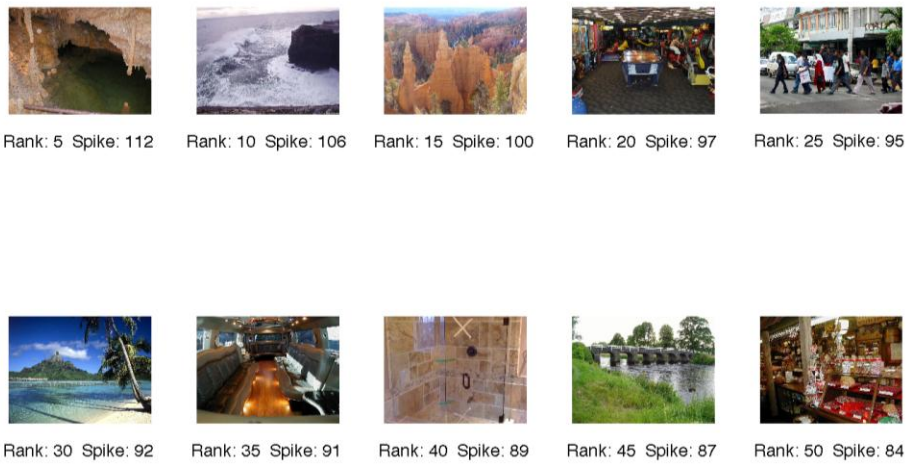


Figure 9. A subset of images that triggered “High neural activity”

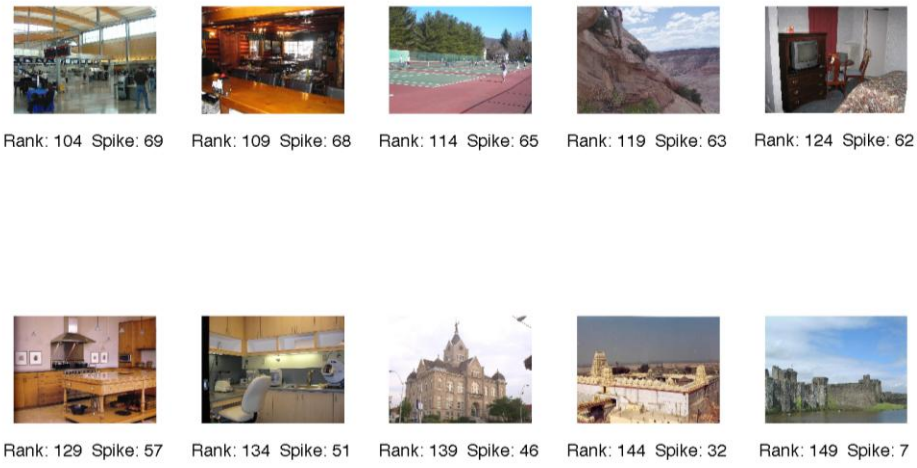


Figure 10. A subset of images that triggered "Low neural activity"

After using 90% of the images in the data for training, we tested the performance of the bag of words model after presenting 10% of images. The following figures contain the classification results.

Figure 11. Classification result of five "High neural activity" images

Figure 12. Classification result of five "High neural activity" images

We also ran additional tests with the identical procedures using two other methods: dense SIFT and spatial pyramid. Because the number of images in the datasets is smaller than those in conventional computer vision experiments, we repeated our testing 100 times and computed mean value to have more accurate classification rate. Table 3 contains the summary of our classification experiments.

	Dataset1 60 Images	Dataset2 100 Images
Dense SIFT	59%	63.5%
Spatial Pyramid	62.7%	60.%%
GIST (Linear SVM)	55.8%	73.7%

Table 3. Mean of classification accuracy after running classification tests for 100 times

3.3 Visual Clutter & Visual Saliency

1) Visual Clutter (Rosenholtz *et al.* [10])

Rosenholtz *et al.* [10] developed a model that computes visual clutter of a natural scene based on three factors: Feature Congestion, Subband Entropy, and Edge Density. The Feature Congestion model is based on the idea that it becomes harder to add a salient feature on a given scene if the scene is already cluttered. The Subband Entropy is relevant to an idea that entropy of the scene is inversely proportional to navigability of the scene. The Edge Density computes density of edges in the scene to infer the number of objects in the scene. An example of the visual clutter model output is provided below.

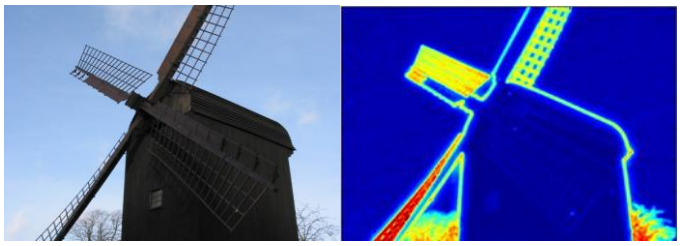


Figure 13. Input image and its clutter map and saliency map

According to Rosenholtz *et al.* [10], the degree of clutter of a scene is relevant to visual search and attention. To evaluate whether the clutter information is related to neural activity in area LIP, we employed the following scheme to analyze our data by the visual clutter model. This test is conducted to confirm our hypothesis that more cluttered scene triggers higher neural activity in area LIP.

1. Compute clutter maps for the images in the data
2. Divide the clutter maps into a 3 by 4 grid and compute mean values of each grid.
3. For each mean value of the grids, compute correlation with neural activity.

Figure 14. Image in default setting Image divided into 3 by 4 grids

	Col 1	Col 2	Col 3	Col 4
Row 1	.2430	.2298	.3744	.3743
Row 2	.0949	.1112	.1697	.2332
Row 3	.1112	.0652	.0801	.0689

Table 4-1. Correlation between visual clutter and neural activity (Dataset 1: 148 images)

	Col 1	Col 2	Col 3	Col 4
Row 1	.2246	.1793	.2818	.3212
Row 2	.1510	.2077	.3024	.3375
Row 3	.1162	.2602	.3764	.3865

Table 4-2. Correlation between visual clutter and neural activity (Dataset 2: 90 images)

Additionally, because the visual clutter model is based on edge detection method, we conducted the same analysis method with edge detection method. Results of the edge detection analysis is provided below.

	Col 1	Col 2	Col 3	Col 4
Row 1	.2866	.2430	.4686	.4801
Row 2	.1588	.0920	.2331	.2313
Row 3	-.1118	-.1031	-.0746	-.1269

Table 5-1. Correlation between visual saliency and neural activity (Dataset 1: 148 images)

	Col 1	Col 2	Col 3	Col 4
Row 1	-.0306	.0876	-.0233	-.0572
Row 2	-.0957	.1290	.1156	.0167
Row 3	.0835	.1257	.0978	.0423

Table 5-2. Correlation between edge detection and neural activity (Dataset 2: 90 images)

2) Visual Saliency (Itti and Koch [4])

After computing the correlation values between the visual clutter and neural activity, we observed that the portions of images producing higher correlation values have similar locations to those of receptive fields of LIP neurons that were used for the experiments. When a stimulus relevant to a function of a certain neuron is presented, neural activity of the neuron is altered. To investigate the location of the image yielding higher correlation with neural activity, we computed correlation values not by the grid but by pixel resolution. If we could detect regions that directly affect an LIP neuron, it will enable us to optimize our analysis further and infer characteristics of image features relevant to area LIP. A receptive field can have various sizes and shapes and densities. Therefore, computational optimization to detect a receptive field of a neuron involves several parameters. Initially, we used three methods to evaluate the detection of receptive fields of LIP neuron cells: visual saliency model developed by Itti and Koch [4], visual clutter model by Rosenholtz *et al*. [10], and the canny edge detection method [2].

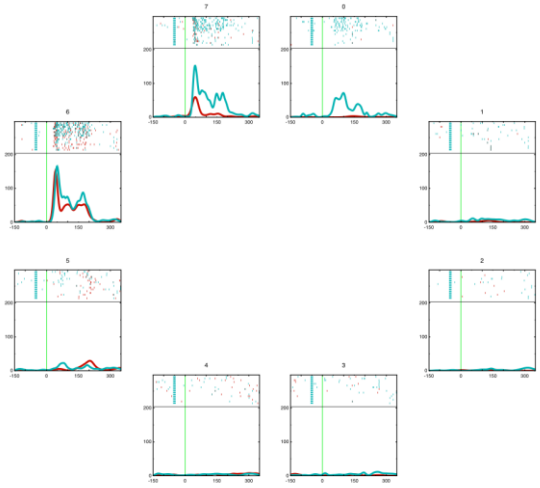
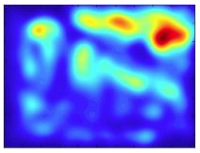
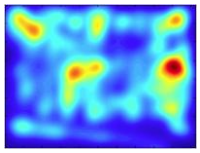
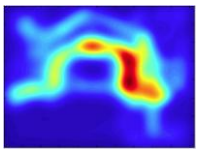


Figure 15. Neural measurement of receptive field location of a neuron in LIP

A location of receptive field is first measured by directly recording neural responses from eight positions of our stimuli space. Figure 15 illustrates this particular neuron has a receptive field on upper left corner. From this particular neuron, we recorded neural activity after presenting natural scenes and computed correlation values with the three methods mentioned above. The visual saliency model by Itti and Koch [4] produced a correlation plot that most resembles the receptive location of this neuron. This observation is important because it computationally shows the validity of the view that an LIP neuron contributes to contruct a visual saliency map.

We compute the correlation maps by the following algorithm.

Image			
Neural Activity	29.6	25.6	30.4
Model Output			
Output at (x,y) = (10,10)	0.0068	0.0851	0.0167

We compute the correlation between the two vectors.

1. Neural Activity = [29.6 25.6 30.4 …]

2. Output at (10,10) = [0.0068 0.0851 0.0167 …]

Correlation at (x,y) = (10,10): 0.1246

We iteration across all pixels in the image. For each pixel, we have a different vector for the model output and the same vector for neural activity. The correlation plots

after we complete the iterations is provided in Figures 16 and 17. For Figure 17, we used output from the clutter model and the canny edge detection method [2] instead of visual saliency for comparison. The saliency model produces the correlation map that has the highest correlation around the receptive field of an LIP neuron.

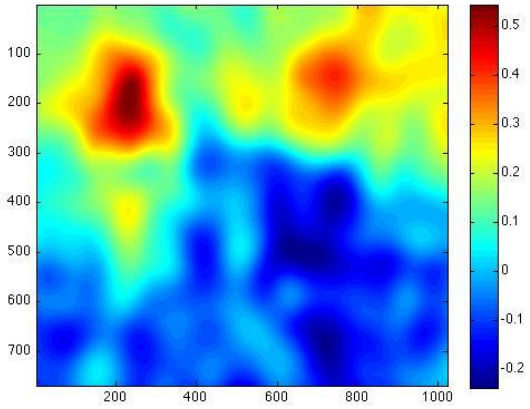


Figure 16. Correlation plot (Visual saliency and neural activity)

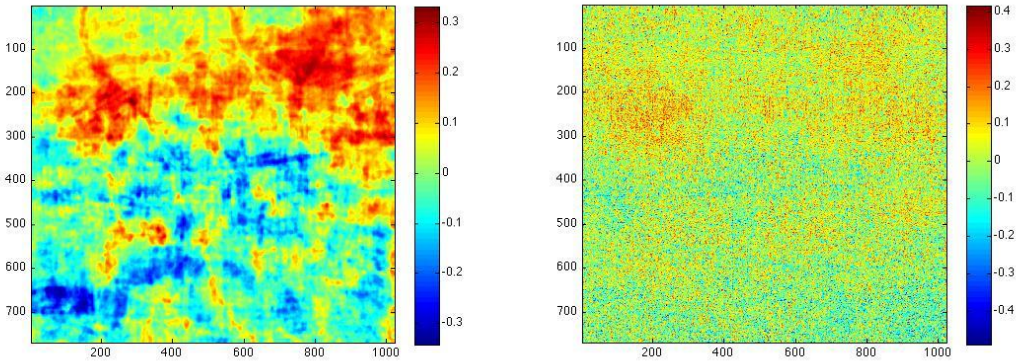


Figure 17. Correlation plot (Visual clutter and neural activity) Correlation plot (Edge detection and neural activity)

Because the visual saliency model is based on color, orientation, and intensity of image pixels of the data, we wanted to improve the correlation plot better by setting the three channels as parameters. A neuron does not always respond equally to color,

orientation, and intensity information, and detecting the pattern that the neuron receives the three inputs let us better produce the correlation plot. After we input the parameters for the three channels, the model computes relative weights of the channel and produce a saliency map. For instance, "COI 124" denotes that the color, orientation, and intensity channels have weights one, two, and four respectively. Different sets of parameters for the three channels did not significantly alter correlation output. Two examples of different correlation plots by differently setting the parameters of the three channels are provided below.

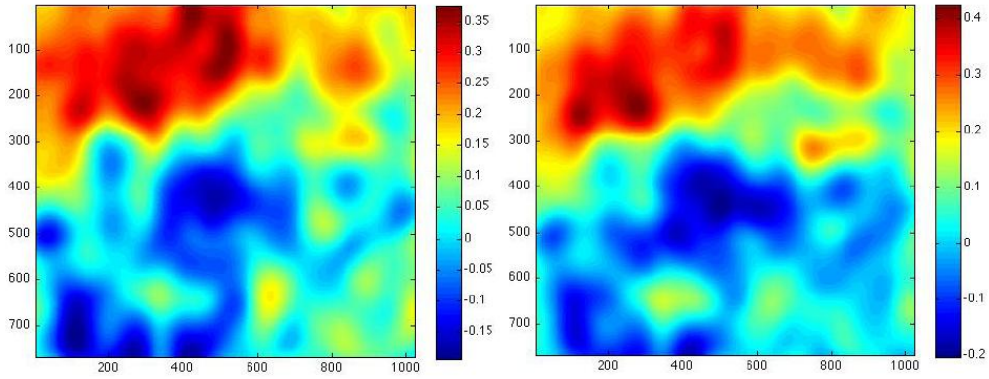


Figure 18. Corelation plot (COI 114) Correlation plot (COI 414)

Also, because the neural recording procedure is 250 ms, the correlation plot can be different by the time window we use for computation. Currently, we use data from 25 ms (stimulus on) to 75ms (stimulus off), and temporal duration of our data is 50ms. However, if we plot correlation value by continuously moving the time window across 250ms, we can detect the specific time point that a neuron is highly responsive to incoming visual stimuli. In this approach, we again divide the visual stimuli space into 3 by 4 grid and compute the correlation value by continuously moving the time window (20

ms). Through this approach, we can detect spatio-temporal location of a receptive field of an LIP neuron. Also, to test whether the correlation value is stastitically significant, we ran permutation tests by permuting the order of images and the collected neural activity data 1000 times.

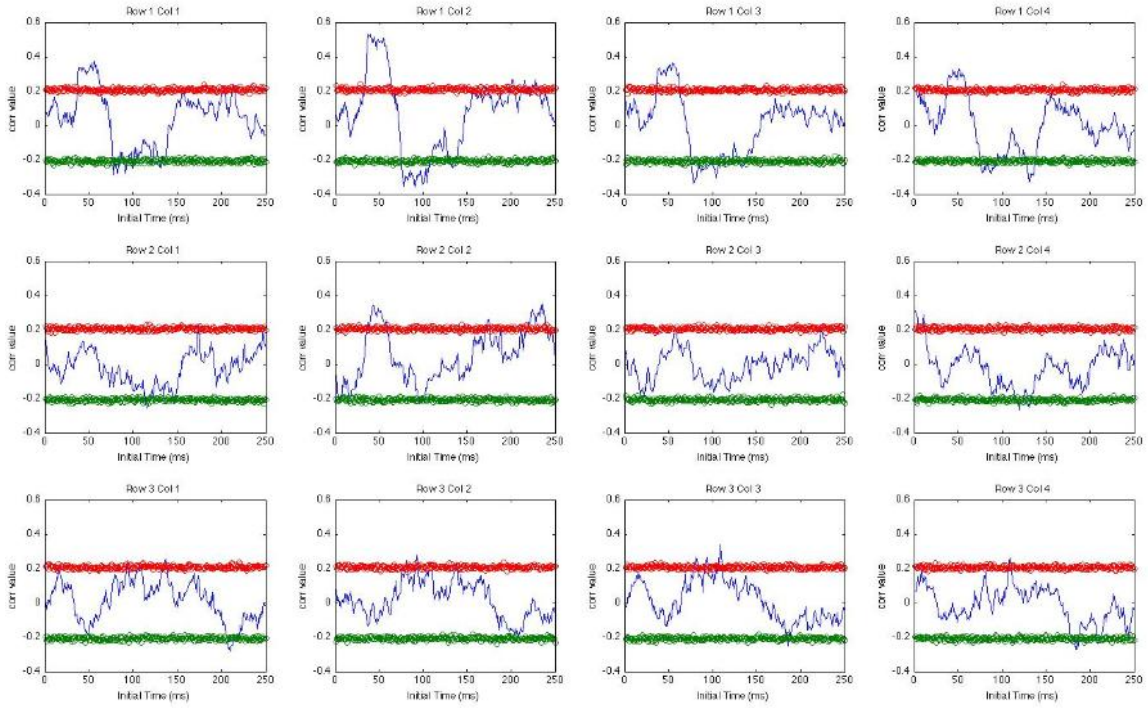


Figure 19. Spatio-temporal correlation (Visual saliency and neural activity)
Red: Upper 2.5% of the correlation value from permutation tests (1000 times)
Green: Lower 2.5% of the correlation value from permutation tests (1000 times)

Receptive field of a neuron, however, does not necessarily correspond to one of the spatial windows above. To better detect spatial region of the receptive field of a neuron, we also continuously move the circular spatial window (256 pixel × 256 pixel) and plot the correlation value. We observed that moving the spatial window by pixel takes long computation time. To reduce computation time, we then moved the window by

32 pixels in each time and plot the correlation values. Examples of correlation maps from moving the spatial window by the two different methods are provided below.

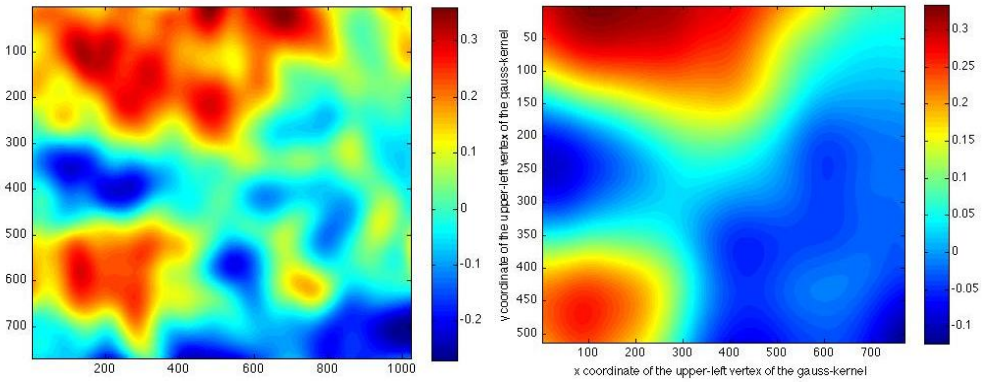


Figure 20. Pixel resolution correlation Spatial window correlation
(Move by 1 pixel after each iteration)

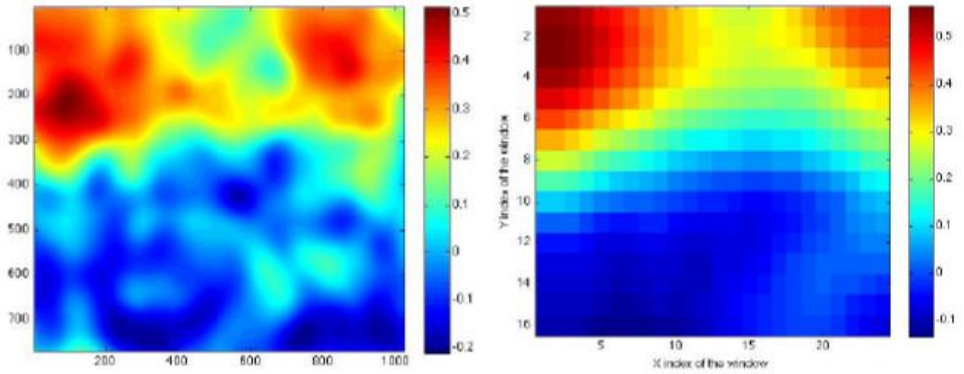


Figure 21. Pixel resolution correlation Spatial window correlation
(Move by 32 pixel after each iteration)

For the spatial window approach, we use a circle that has a radius of 128 pixels. Though computer vision experiments conventionally use retangular regions for detection experiments, we used circle to model biological vision by using a gaussian kernel. After moving the spatial window by 1 pixel and 32 pixels respectively, we observed that the correlation value does not increase significantly compared to that from pixel resolution.

However, we were able to detect a region that is biologically comparable to a receptive field of a neuron by choosing the region producing the maximum correlation value.

During our experiments with computer vision methods, one of the reaons that the methods were not able to classify the images based on neural activity is because number of images in our experiments is limited. With the visual saliency model, we tested whether the model is robust enough to function with small number of images. The following is our testing scheme to test the robustness of the model and also to evaluate how many percentages of the total images in the data are necessary to achieve the correlation value from the default data.

- Experiment 1: Use 10% of the total images in the neural data
- Experiment 2: Use 20% of the total images in the neural data
- Experiment 3: Use 40% of the total images in the neural data
- Experiment 4: Use 60% of the total images in the neural data
- Experiment 5: Use 80% of the total images in the neural data

All experiments are conducted 1000 times. For computing the correlation value, we employed the method of moving the circular window by 32 pixels and set the maximum value of the correlation as our static.

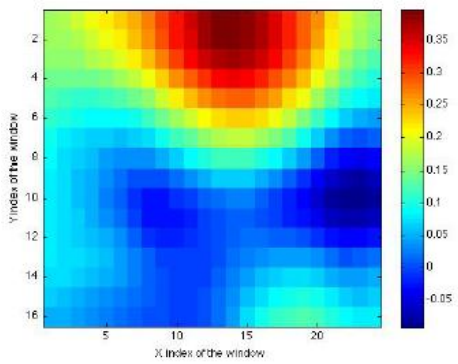


Figure 22. Correlation map from all the images in one dataset (150 images)

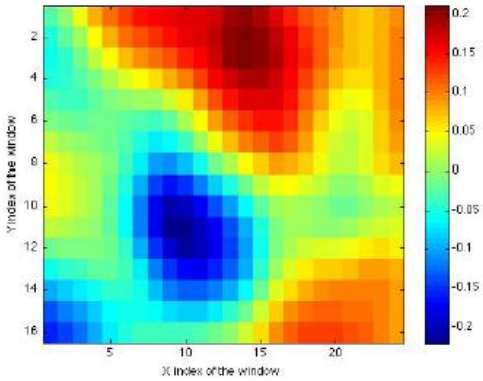


Figure 23. Correlation map from 60 images (40%)

We compute the correlation between the two maps in figures 22 and 23, and we use the computed correlation value as our static for our evaluation. We repeated the experiments 1, 2, 3, 4, and 5 for 1000 times and observed the distribution of the correlation value. The following is one test result of a dataset containing 90 images.

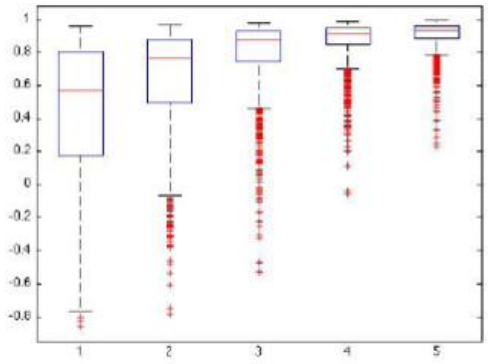


Figure 24. Random subsampling experiment result (90 images total)

Overall, the visual saliency map is able to produce a correlation plot that is highly correlated with the correlation map using all the images in a dataset if we have at least 60% of the total images. This result illustrates that the visal saliency model can be robust

enough even with small number of images and a better method to study the visual system compared to other computer vision methods.

4 Discussion

Though computer vision system aims to train the computers recognize visual scenes as humans do, conventional computer vision methods were not able to simulate visual attention process of scene-selective visual neurons in area LIP. Binary classification results with methods such as GIST, Dense SIFT, and Spatial Pyramid did not yield a high accuracy rate. If we want to have a better model of human visual perception, current research directions in computer vision largely centered on obejct detection, segmentation, and classification should be reconsidered. Human visual perception is still an upper-bound for all computer vision systems, and neurobiological approach to build a new computer vision system can give us a solution to better the computer vision models.

The visual clutter model by Rosenholtz *et al.* [10] and the visual saliency model by Itti and Koch [4], however, were able to detect neural receptive field of a neuron in LIP both spatially and temporally. The visual clutter model is the first model for us to detect a spatial receptive field of an LIP neuron because the correlation map showed us higher values at regions that correspond to neural receptive field. However, the correlation map did not always correspond with the receptive field. The visual saliency map, which itself correspond to the function of LIP neurons contructing a saliency map, produced a correlation map that better matches to that of neural receptive field of neuron cells. Also, the saliency map showed us better robustness in producing the correlation map with small number of images compared to conventional computer vision methods.

The visual saliency model is largely based on color, intensity, and orientation of an input image. Neurons in LIP, however, can construct salience map both in top-down and bottom-up direction, and our results were not able to test our hypothesis about the top-down direction visual attention. Saccadic eye movements were collected simultaneously during the neural recording, but the eye movements are not used in our experiments. Implementing methods to use the eye movements would let us be able to test our hypothesis about the top-down direction attention. By comparing the eye movements and the visual saliency map by Itti and Koch [4], we can be able to differentiate the regions that are salient to the model and to the primate.

Visual attention usually involves learning, and therefore neural activity in LIP can change if we insert a specific target in a scene and train the primate to search the target. Neural activity after this type of learning can be different from our data, which does not involve any kind of specific task, and can give us another answer about the function of LIP neurons. To computationally model visual saliency after this type of learning has not been successful, and observation of neural data during the task can give us clue about building the computational model.

LIP is not the only brain region that participates in constructing the saliency map. Another prominent region that contributes to object detection and is known to provide a pathway of visual information is area IT (Inferotemporal). Though Serre *et al*. [11] developed a computational model of object detection, no modeling has been completed to illustrate the interaction between LIP and IT. Interactions between several brain regions during the visual attention process is highly complex, and the model which can simulate these interactions would be able to give us a better computer vision model.

References

[1] Bisley J. W. and Goldberg M. E. (2010) Attention, Intention, and Priority in the Parietal Lobe. *Annual Review Neuroscience* **33:**1-21

[2] Canny, J., A Computational Approach To Edge Detection (1986) *IEEE Trans. Pattern Analysis and Machine Intelligence*, **8**(6):679–698

[3] Felleman, D. J., Van Essen, D. C. (1991) Distributed Hierarchical Processing in the Primate. *Cerebral Cortex* **1**(1):1-47

[4] Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* **40** 1489-1506

[5] Judd, T., Ehinger, K., Durand, F., Torralba, A. (2009) Learning to Predict Where Humans Look. *International Conference on Computer Vision (ICCV)*

[6] Kay, K. N., Naselaris, T., Prenger, R. J., Gallant, J. L. (2008) Identifying natural images from human brain activity. *Nature* **452** 352-355

[7] Monosov I. E., Sheinberg, D. L., Thompson K. G. (2010) Paired neuron recordings in the prefrontal and inferotemporal cortices reveal that spatial selection precedes object identification during visual search. Proceedings of the National Academy of Sciences

[8] Oliva, A., Torralba, A. (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision* **42** (3): 145-175

[9] Pinto, N., Barhomi, Y., Cox, D. D., DiCarlo, J. J. (2011) Comparing State-of-the-Art Visual Features on Invariant Object Recognition Tasks. *IEEE Workshop on Applications of Computer Vision (WACV)*

[10] Rosenholtz, R., Li, Y., and Nakano, L. (2007). Measuring visual clutter. *Journal of Vision* **7** (2):17, 1-22

[11] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust Object Recognition with Cortex-Like Mechanisms. *PAMI* **29** (3), 411-426

[12] Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A. (2010) SUN Database: Large-scale Scene Recognition from Abbey to Zoo. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*