Neurobiologically realistic model of statistical pooling in peripheral vision

Noshaba Cheema,* Lex Fridman, Ruth Rosenholtz, and Christoph Zetzsche

University of Applied Sciences Bremen, Germany
Cognitive Neuroinformatics, University of Bremen, Germany
CSAIL, Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA

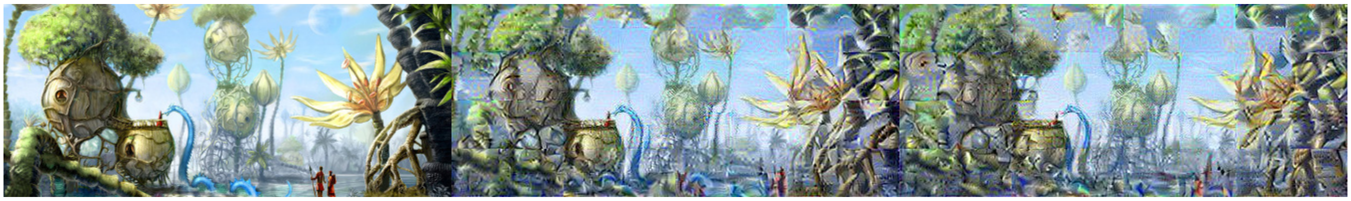

Figure 1: *Reconstruction results for the mathematical and the neurobiological model. From left to right: original image, mathematical correlation (Gram matrix), neurobiological AND-like interactions. Overall reconstruction quality is similar for both models. Detailed visual comparisons can be made by zooming into the visible individual tiles. Image: ©Fantasy Landscape by Deevad*

Keywords: peripheral vision, visual crowding, neuro-biologically motivated statistics, neural networks, image compression

Concepts: •**Mathematics of computing** → Probability and statistics;

1 Introduction

Our senses can process only a limited amount of the incoming sensory information. In the human visual system this becomes apparent in the reduced performance in the peripheral field of view, as compared to the central fovea. Recent research shows that this loss of information cannot solely be attributed to a spatially coarser resolution but is essentially caused by statistical pooling operations which lead to a texture-like representation. This so called "crowding effect" can be seen as a strategy of the visual system to deal with the information bottleneck in sensory processing. As such, it may also be of interest for image compression, feature extraction and computer vision. A recent approach towards modelling this effect makes direct use of formal statistical computations [Balas et al. 2009; Freeman and Simoncelli 2011], and thus is not completely convincing with respect to its neurobiological plausibility. Here we investigate whether a more plausible model can be developed and whether it achieves the desirable properties.

2 Summary statistics model

Our model is illustrated in Fig. 2. It differs from the existing crowding model [Balas et al. 2009; Freeman and Simoncelli 2011] in two aspects, in the use of a deep network architecture and in the provision of neurobiologically plausible AND-like operations as basic nonlinearities. We first apply a spatial tiling operation to an image x in which the tiles get larger with increasing eccentricity, roughly similar to the assumed layout of the spatial overlapping pooling regions in human vision, cf. [Balas et al. 2009; Freeman and Simoncelli 2011]. To aid visual inspection, no smoothing is applied to the tiles, such that borders remain visible. We then compute the activations in layer "pool3" of each vectorized pooling region $\vec{p}$ with a normalized version [Gatys et al. 2015] of the VGG-19 net [Simonyan and Zisserman 2014] with average pooling. These feature maps are then stored in a matrix $F^p \in \mathbb{R}^{NxM}$, where F^p_{jk} is the activation of the filter j at k in pooling region p, N the number of filter kernels and M the size of each vectorized feature map.

In order to obtain a texture-like representation within each pooling region, the spatial information needs to be discarded. A summary statistic that does this, is given by the correlations $S^p = \sum_k F^p_{ik} F^p_{jk}$ based on the feature map matrix F of a pooling region p. Like the crowding model of [Balas et al. 2009; Freeman and Simoncelli 2011] the present model version makes explicit use of the *multiplication* of two variables. However, researchers have long debated whether such multiplication operations are biologically plausible [Koch 2004]. Furthermore, computation of the formal statistical correlations may not be necessary, as much of the functionality may be preserved if multiplication is replaced by a neurobiologically more plausible operation. We thus replace the multiplications by neurophysiologically plausible AND-like operations [Zetzsche and Barth 1990] . It is well known that biological hardware can easily realize an ON/OFF rectification and nonlinear transducer functions with sigmoid shape. With these ingredients, one can make use of an old Babylonian trick to derive the suitable AND-like computations: $AND(a, b) = N[a + b] - N[a - b]$ where N is a suitable nonlinear transducer function. These AND operations are characterized by the property that they attain their maximum (for the sum $a + b$ constrained) if a and b have the same size. If a or b is decreased, the response is systematically reduced, until it vanishes if either a or b equals zero. If N is a sigmoid nonlinearity the resulting AND will have a threshold-like behavior for small input values and will go into saturation for large input values. For simplicity, the model version which is based on the formal statistical computations is henceforth designated as "mathematical model" (in spite of its other neurobiology-related components) and the model version with the AND operations is designated as "neurobiological model".

The information provided by the different model representations can be visualized and evaluated by reconstructed images, designated as *mongrels* [Balas et al. 2009; Rosenholtz et al. 2012] and as *metamers* [Freeman and Simoncelli 2011]. They can provide important hints on which information is being preserved and which is discarded by the representation. A reconstruction can be obtained by using gradient descent to generate a new image that has the same local summary statistics as the original one. In our work we used the L-BFGS-B solver, which is a reasonable choice for such a high dimensional optimization problem. The loss-function

*e-mail:ncheema@mpi-inf.mpg.de

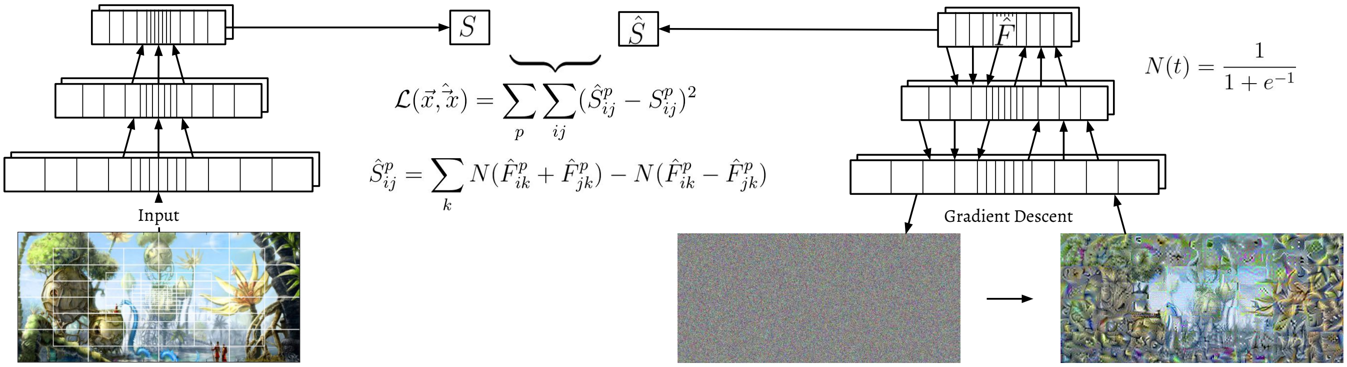
S
$\hat{S}$
$\hat{F}$
$\mathcal{L}(\vec{x},\hat{\vec{x}}) = \sum_p \sum_{ij} (\hat{S}^p_{ij} - S^p_{ij})^2$
$\hat{S}^p_{ij} = \sum_k N(\hat{F}^p_{ik} + \hat{F}^p_{jk}) - N(\hat{F}^p_{ik} - \hat{F}^p_{jk})$
$N(t) = \frac{1}{1+e^{-1}}$
Input
Gradient Descent

Figure 2: *Model architecture and reconstruction algorithm.*

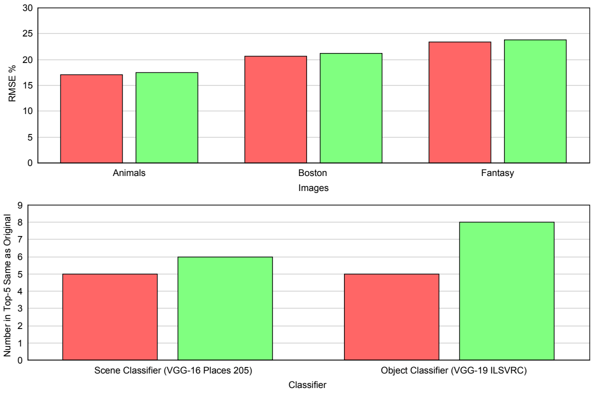

Figure 3: *Comparisons between mathematical (red) and neurobiological model (green). Upper figure: Similarity of reconstruction results as measured in terms of root-mean-square distance. Lower figure: Similarity of the mathematical and the neurobiological representation as measured in terms of classification performance. Left: place recognition. Right: object recognition. Classification performance is measured in how many of the top-5 classification results were the same as the top-5 results of the original 3 images.*

that is used for this optimization strategy is defined by $\mathcal{L}(\vec{x},\hat{\vec{x}}) = \sum_p \sum_{ij} (\hat{S}^p_{ij} - S^p_{ij})^2$.

3 Model Comparison

We compare the mathematical and the neurobiological with respect to different criteria. First, we consider the reconstructed images as such. The neurobiological model yields reconstruction results which are at least as good, and in some aspects even superior to those obtained with the mathematical correlation statistics (Fig. 1). Further measures of the similarity between the reconstructions obtained from the two types of models are the simple root-mean-square deviation from the original image, and the classification performance that can be obtained with the respective reconstruction images. Fig. 3 shows that the root-mean-square distance betwwen the reconstructed images and the original images is approximately of the the same order for the mathematical models and the neurobiological models and that the classification performance is also similar. The images had been classified using the VGG16 net trained on Places205 for scene recognition and the VGG19 net trained on object recognition using the data from ImageNet2012. A more detailed table can be found here.

In conclusion, our investigations indicate that the information content being represented in the neurobiological model is comparable or even superior to that of the the mathematical model. This becomes evident in the reconstructed "Mongrel" images, in the distance measures, and in the classification performance. These results may thus be considered as one further step towards a fully plausible model of the neural information processing being performed in the visual periphery.

References

BALAS, B., NAKANO, L., AND ROSENHOLTZ, R. 2009. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision 9*, 12, 13–13.

FREEMAN, J., AND SIMONCELLI, E. P. 2011. Metamers of the ventral stream. *Nature neuroscience 14*, 9, 1195–1201.

GATYS, L., ECKER, A. S., AND BETHGE, M. 2015. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 262–270.

KOCH, C. 2004. *Biophysics of computation: information processing in single neurons*. Oxford university press.

PORTILLA, J., AND SIMONCELLI, E. P. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision 40*, 1, 49–70.

ROSENHOLTZ, R., HUANG, J., RAJ, A., BALAS, B. J., AND ILIE, L. 2012. A summary statistic representation in peripheral vision explains visual search. *Journal of vision 12*, 4, 14–14.

SIMONYAN, K., AND ZISSERMAN, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

ZETZSCHE, C., AND BARTH, E. 1990. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision research 30*, 7, 1111–1117.