Peripheral pooling is tuned to the localization task

Sven Eberhardt
Cognitive Neuroinformatics, University of Bremen, Bremen, Germany

Christoph Zetzsche
Cognitive Neuroinformatics, University of Bremen, Bremen, Germany

Kerstin Schill
Cognitive Neuroinformatics, University of Bremen, Bremen, Germany

The human visual system exhibits substantially different properties between foveal and peripheral vision. Peripheral vision is special in that it has to compress data onto fewer units by reduced visual acuity and larger receptive fields, yielding greatly reduced performance on many tasks such as object recognition. However, here we show that the pooling operations implemented by peripheral vision provide exactly the invariance properties required by a self-localization task. We test the effect of different pooling sizes, as well as acuity reduction, on localization, object recognition, and scene categorization tasks. We find that peripheral pooling, but not reduced acuity, affects localization performance positively, whereas it is detrimental to object recognition performance.

Introduction

The distribution of visual processing power is not uniform across the visual field (Daniel & Whitteridge, 1961; Polyak, 1941; Rodieck, 1998; Weymouth, 1958; Wilson, Levi, Maffei, Rovamo, & DeValois, 1990). The best visual processing properties are restricted to only a small central area, the fovea, whereas at larger eccentricities only coarse-grained and distorted information is available.

Two reasons are regarded to cause this inhomogeneity. First, there is a significant drop of spatial resolution with increasing eccentricity (Weymouth, 1958), limited by the resolution of the discrete neural image carried by the human optic nerve (Wilkinson, Anderson, Bradley & Thibos, 2016).

Fine details can only be discriminated efficiently in the fovea, whereas only coarse structures can be resolved at larger eccentricities. Consequently, peripheral vision is not well-suited for object or letter recognition, since these tasks require the identification of small spatial features and of their relative spatial positions. Since both types of features are disturbed by low-pass filtering, the accurate perception of objects is severely impaired in the periphery.

Although peripheral vision exhibits reduced acuity, measures exist to compensate for this—at least under laboratory testing conditions: An isolated object, if scaled appropriately such that its individual features become larger than the peripheral resolution limit, can again be perfectly recognized (Anstis, 1974; Virsu, Näsänen, & Osmoviita, 1987; Virsu & Rovamo, 1979). This seems to indicate that the performance is not substantially different between peripheral and foveal vision.

However, there is an important second reason that leads to inferior processing in the visual periphery. Not only does the spatial resolution drop, but an additional destructive effect due to a specific kind of spatial interference can be observed. This effect arises between neighboring objects, even if they are sufficiently scaled to surpass the peripheral resolution limit. If the objects are not clearly separated by a critical distance from each other, recognition will be substantially reduced, a phenomenon known as "crowding" (for review, see, e.g., Levi, 2008; Pelli & Tillman, 2008; Strasburger, Rentschler, & Jüttner, 2011; Whitney & Levi, 2011).

The currently prevailing interpretation of crowding is in terms of some sort of spatial pooling of visual information (Balas, Nakano, & Rosenholtz, 2009; Freeman, Chakravarthi, & Pelli, 2012; Freeman & Simoncelli, 2011; Levi, 2008; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli, Palomares, & Majaj, 2004; Van den Berg, Roerdink, & Cornelissen, 2010; Wilkinson, Wilson, & Ellemberg, 1997). In particular, it has been proposed that the peripheral representations

Citation: Eberhardt, S., Zetzsche, C., & Schill, K. (2016). Peripheral pooling is tuned to the localization task. *Journal of Vision*, *16*(2):14, 1–13, doi:10.1167/16.2.14.

doi: 10.1167/16.2.14 Received June 6, 2015; published November 30, 2016 ISSN 1534-7362

in visual cortex provide only local statistics of the visual image (Balas et al., 2009; Freeman & Simoncelli, 2011; Parkes et al., 2001; Van den Berg et al., 2010). However, a recent study suggests that not all elements of the peripheral descriptor are captured in local statistics (Wallis, Bethge, & Wichmann, 2016). Since this statistical pooling cannot be avoided, it is designated as mandatory pooling or compulsory averaging (Parkes et al., 2001). Crowding in this perspective is seen to be caused by a two-stage process in which individual features are first detected independently and are then spatially pooled or integrated in some fashion on a subsequent stage of processing. Only this pooled information is available for subsequent recognition processes (but see, e.g., Chaney, Fischer, & Whitney, 2014; Herzog, Sayim, Chicherov, & Manassi, 2015). We thus observe spatial pooling on two stages: In addition to the reduced spatial acuity (which is a pooling across space in the sense of a low-pass filtering of the luminance function by optical and neural means) there is an additional spatial pooling of elementary features (for example V1 units) on a higher cortical processing stage.

Like the reduced spatial resolution of the periphery, crowding represents an information bottleneck for peripheral vision (Levi, 2008; Rosenholtz, Huang, & Ehinger, 2012a). However, while the resolution bottleneck can be overcome by scaling, the limitations incurred by crowding are much more severe. In natural scenes and in our cultural environments objects rarely appear in isolation, or are clearly separated from each other. Rather, these environments are often heavily cluttered such that under natural viewing conditions peripheral performance may be much worse than observed in many traditional laboratory tests of visual perception. During real-world vision, most of the visual field will be crowded by a large number of objects and consequently most of these objects will not be correctly recognized by the periphery.

What is the reason that forces the system to accept such a severe bottleneck? The main reason is that it is simply impossible to afford full foveal processing quality across the entire visual field. Our eyes would have to be as large as our heads, and our brain would weigh several thousand kilograms (Anstis, 1998; Schwartz, 1994). The system seems to deal with this pressure on complexity by using large integration fields. Larger receptive fields are cheaper than small ones because fewer units are required to densely cover the peripheral visual field (Pelli et al., 2004).

Large integration fields are a reasonable way to deal with limited resources (as compared with using only a small subset of the features). The statistical pooling can be related to lossy compression (Rosenholtz et al., 2012a) and the specific peripheral processing then is a result of trying to achieve the best information possible under the restriction of overall limited complexity. The cortex thereby can form an economic representation without needing to represent each individual member of a set of objects (Ariely, 2001). Pooled statistical information enables us to rapidly asses the general properties and layout of a scene, or its *gist* (Oliva, 2005). This in turn can provide contextual information, which helps in central object recognition as well as in motor control (Torralba, Oliva, Castelhano, & Henderson, 2006).

However, although the system seems to make the best use of its limited resources, for many tasks the extended feature pooling is massively destructive. These range from simple tasks like Vernier acuity and orientation discrimination to more complex tasks like letter and object recognition (Levi, 2008; Strasburger et al., 2011). On the phenomenal level, flanking objects "squash" the target object, features become "jumbled," or seem to be located at wrong positions (Pelli et al., 2004). Most investigations so far suggest that perceptual tasks can always be better solved with the fovea than with the periphery. Compared with the performance level of the fovea, peripheral processing seems to be hampered by a deficit.

But is this really true for all tasks? Are there basic behavioral tasks that can be solved even *better* by peripheral vision? Is there any beneficial byproduct of the spatial feature pooling associated with crowding? Currently this is regarded as an open question (Bulakowski, Post, & Whitney, 2011).

Local pooling is an important component of visual recognition models because it provides invariance to translation and small local rotations (Dalal & Triggs, 2005; Pinto, Barhomi, Cox, & DiCarlo, 2011). The established way to employ local pooling for feature recognition is in a hierarchical fashion, where pooling layers with small pooling ranges are alternated with feature matching layers to build set features that are increasingly more specific to more complex patterns, while at the same time being increasingly more invariant to location and scale. This kind of hierarchy has been suggested by Fukushima, Miyake, and Ito (1983), Marko and Giebel (1970), Riesenhuber and Poggio (1999), and was later refined to form the HMAX-model by Serre, Oliva, and Poggio (2007) and also laid the foundation for deep learning architectures for image classification (Krizhevsky, Sutskever, & Hinton, 2012).

But can pooling also be helpful when applied with a large range in a single step, such that it destructs local information? In the HMAX model, this has been tried by Isik, Leibo, Mutch, Lee, and Poggio (2011) to model peripheral pooling behavior, but has not been used to show improvement on any task. Classic candidates of models employing large pooling ranges that come to mind would be the recognition of the *gist* of a scene

(Oliva, 2005), or the fast parallel processing of ensemble properties (Ariely, 2001). However, regarding these two perceptual tasks, evidence so far seems inconclusive. Bulakowski et al. (2011) argued that dissociations between crowding effects and ensemble representations make it difficult to interpret the latter as a positive byproduct of the feature pooling that causes crowding. Gist recognition is surprisingly good in the periphery but central presentation of wide-field images always yields the best results (Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001). Larson and Loschky (2009) found in a detailed analysis that despite a seeming advantage of the periphery (if defined as 5°), central vision is in fact more efficient at processing gist.

We thus consider here an alternative task, vision-based *localization*. It is of substantially different nature than the classic perceptual tasks, since it is closer related to mobility and navigation than to the recognition of objects or scenes. Furthermore, it differs from perceptual tasks by the substantially different invariance properties (Eberhardt & Zetzsche, 2013; Wolter, Reineking, Zetzsche, & Schill, 2009).

Two components of visual matching contribute to the vision-based localization task. The first component is to match known objects at previously learned locations, a landmark-based approach that can be driven by an object recognition mechanism. A second component is not to match individual objects, but to match scene material properties such as the type of foliage in natural or building styles in urban scenes. An example would be to recognize a street scene as being taken in Paris from its typical house façades (Doersch, Singh, & Gupta, 2012).

Localization based on the second component does not make assumptions about the particular location of certain features and can therefore be driven by a texture-based approach (Eberhardt & Zetzsche, 2013) that exhibits large pooling properties. We therefore presume that localization could be a candidate for a genuine advantage of the periphery over foveal vision. We test this hypothesis by using both model simulations and behavioral experiments. In particular, we will consider the following questions:

- What influence does spatial pooling of features, as performed in the visual periphery, have on different tasks? Of particular interest here is the influence on object recognition, as opposed to localization.
- Which kind of features and architectural properties are most suited for vision-based localization? Here we will consider different existing model architectures (GIST, Textons, Spatial Pyramids, Statistical Periphery Model).
- Which kind of pooling is better suited for the different tasks: input pooling or feature pooling?
- Do human subjects show a genuine advantage of peripheral over foveal vision in a localization task?
- Does the human periphery with its inherent pooling capabilities profit from the recognition of scene material properties as opposed to recognizing concrete objects to identify a location?

The paper is organized as follows. In the Peripheral pooling model section, we present a peripheral pooling model in which we are able to control for the input acuity as well as pooling sizes. We also introduce a database for localization and test the effect of pooling and acuity variations on the localization as well as on control datasets. In the Model comparison section, we evaluate the performance of different visual feature descriptors on localization as well as on other control tasks to find which kind of descriptor is best suited to solve each problem. In the Human experiments section, we present an experiment on human subjects that shows how much peripheral vision contributes to a localization task compared with foveal vision. Finally, results are summarized and discussed in the Discussion section.

Peripheral pooling model

Methods

Model description

To test the influence of pooling regions on different tasks, we use a model derived from early stages of the HMAX implementation by Serre et al. (2007).

Grayscale input images are rescaled to 256 × 256 input units and padded with black if needed. On inputs $I_{x,y}$, edge detection using Gabor filters $G_{x,y}$ is applied using a normalized dot product to produce outputs $S^1_{x,y,\lambda}$:

$$G_{x,y} = \exp\left(\frac{\bar{x}^2 + \alpha\bar{y}^2}{-2\sigma^2}\right) \cdot \cos\left(\frac{\bar{x} \cdot 2\pi}{\theta} - \phi\right) \quad (1)$$

$$\bar{x} = \cos(\lambda) \cdot x - \sin(\lambda) \cdot y \quad (2)$$

$$\bar{y} = \cos(\lambda) \cdot y + \sin(\lambda) \cdot x \quad (3)$$

$$\overline{S}^1_{x,y,\lambda} = \sum_{x'=-r}^{+r} \sum_{y'=-r}^{+r} I_{x,y} \cdot G_{x+x',y+y'} \quad (4)$$

$$S^1_{x,y,\lambda} = \frac{\overline{S}^1_{x,y,\lambda} + \epsilon}{\sqrt{\sum_{x'=-r}^{+r} \sum_{y'=-r}^{+r} I^2_{x+x',y+y'} + \epsilon}} \quad (5)$$

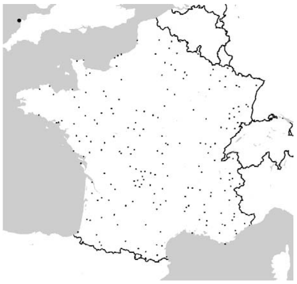

Figure 1. Randomly sampled locations from the Google Street View data source. Each dot represents one of 204 locations, i.e., classes in the dataset. Instances in the class are 36 images taken at different yaw rotations.

Six orientations $\lambda \in \{0, \pi/6, \pi/3, \pi/2, \pi \times 2/3, \pi \times 5/6\}$ are sampled with $\alpha = 0.3$, $\theta = 3.5$, $\sigma = 2.8$, and $r = 5$.

A pattern matching layer S^2 computes radial basis function distances ($\sigma_2 = 1/3$) of inputs $S^2_{x,y,f}$ to a simple feature dictionary $D^2_{x,y,\lambda,f}$ with 2,000 features f of size 3 × 3 ($r_2 = 1$) in the dictionary. The dictionary is sampled from random locations in natural images and sparsified such that only the orientation $\lambda^{x,y}$ with the strongest activation in the sampling source is stored in the template and the remaining activations set to zero.

$$S^2_{x,y,f} = \exp\left(\frac{-\left(\sum_{x'=-r_2}^{+r_2}\sum_{y'=-r_2}^{+r_2}\sum_{\lambda} S^1_{x+x\prime,y+y\prime,\lambda} - D_{x\prime}, y\prime, \lambda, f\right)\cdot 1}{\sigma_{2^2}}\right) \tag{6}$$

A local inhibition layer $L^2_{x,y,f}$ on top of $S^2_{x,y,f}$ provides inhibition between features at each location. It zeroes all feature activations that lie below a threshold of $\Theta = 0.9$ of the maximum activation over all features at each location. A pooling layer C^2 is put on top of L^2, which pools over a range of inputs using a simple averaging operation.

Two variations are tested here: The pooling range λ is varied to model pooling over peripheral features, while input images are rescaled to a smaller size to model the effect of reducing acuity. When the pooling range is modified, the stride or sampling density of features is kept constant, resulting in additional overlap between features on the larger pooling conditions. The model software is implemented in MATLAB using Cortical Network Simulator (Mutch, Knoblich, & Poggio, 2010) and built in part from the HMAX package.

Model testbed

For the evaluation diagnosticity of feature descriptors, we follow a classification approach. We evaluate performance of a classifier on different features vectors and varying tasks.

For the object recognition and scene classification comparison tasks, this is the most natural approach as established datasets exist and are in widespread use. For object recognition, Caltech 101 by Fei-Fei, Fergus, and Perona (2007) provides a well-tested database for classification among a large number of object categories. For Scene Categorization, the Scene-15 dataset by Lazebnik, Schmid, and Ponce (2006) is used.

The localization task cannot be transferred to the classification concept naturally, because vision-based localization is typically assessed by a procedure that derives coordinates form images and calculates error as distance between determined and true position. However, a purely class-based approach is chosen here. Images are sampled randomly from Google Street View, where each GPS location constitutes one class, and samples of each class are generated by testing different views. That is, geometric information between the classes is discarded entirely and performance is measured in percent correct assigned classes to yield comparability with the other classification tasks.

Street view images are provided through the Google Street View Panorama API, which yields equirectangular (plate carre) projection, extracted at 90° field of view for the model runs. A total of 204 random locations are sampled, constrained within France to ensure that camera recording settings are as uniform as possible (see Figure 1). Because all images from one location are taken at the same time, there are some external conditions not related to localization that classifiers may pick up on that cannot be controlled such as time of the day and weather condition. However, since the number of locations is large and most images are recorded at moderate weather conditions during daytime, this effect should be small.

To control for features that just match image overlap between views of a location, an *opposite* condition is tested, in which it is ensured that training and testing samples never have overlapping image regions. In this condition test samples are guaranteed to be rotated at least 90° away from all training samples (see Figure 2).

All data was reduced to 256 dimensions using principal component analysis (PCA), and classification was executed using linear regression with leave-one-out-cross-validation for the regularization parameter, 10 training samples per class and the rest used for testing. Classifier training was performed using the GURLS software package (Tacchetti, Mallapragada, Santoro, & Rosasco, 2011) for MATLAB.

For the performance measurement, many multiclass classification measures exist (Sokolova & Lapalme, 2009). Since datasets of varying class counts are compared here, classification accuracy was evaluated as mean recall over all classes. That is, with tp_i and fn_i

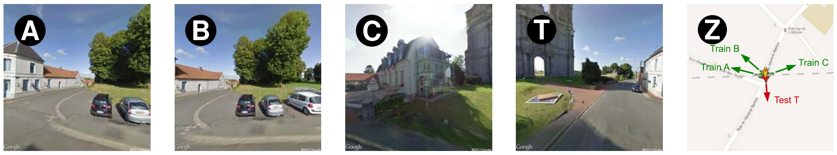

Figure 2. Opposite testing condition: (A, B, C) are training samples for the classifier, which include no visual overlap with testing sample T, as seen on bird's-eye view map Z. Images (c) by Google Inc., used under Fair Use policy (http://www.google.com/permissions/geoguidelines.html).

being true positives and false negatives of class i respectively, performance p over n classes is:

$$p = \frac{1}{n}\sum_{i=1}^{n}\frac{tp_i}{tp_i + fn_i} \quad (7)$$

All classification runs have been repeated 50 times with different train and test sets to yield a more stable average. Mean errors were calculated for the error margins given on the percent correct.

Results

Performance results of the model classifier are shown in Figure 3. Both pooling regions and input acuity changes have a strong effect on the resulting performance. For the Street View dataset, the localization task, performance increases drastically as a function of pooling range from 1.04 ± 0.02% ($d' = 0.29$) for no pooling ($\lambda = 1$) up to 66.01 ± 0.12% ($d' = 3.19$) for maximum pooling over the whole image range ($\lambda = 240$; see Figure 3A) The *opposite* condition follows the same pattern at a much lower performance from 0.68 ± 0.07% ($d' = 0.12$) at $\lambda = 1$ and 10.03 ± 0.26% ($d' = 1.38$) at $\lambda = 240$. The scene categorization dataset shows similar dependency, going from 17.96 ± 0.31% ($d' = 0.66$) at $\lambda = 1$ up to 51.07 ± 0.21% ($d' = 1.72$) at $\lambda = 30$, but then saturates, indicating that larger pooling ranges are no longer advantageous to that task at some point. Although localization and scene categorization benefit from pooling, the opposite effect can be observed for the object categorization test using Caltech 101. Here, the result goes down slightly from 31.09 ± 0.20% ($d' = 1.97$) at $\lambda = 240$ to 21.37 ± 0.17% ($d' = 1.64$) at $\lambda = 1$.

Reduction of image resolution as a simulation of reduced visual acuity in the periphery does not result in the same performance gains (see Figure 3B). Unlike pooling on the higher level, performance is affected negatively by the reduced acuity for all datasets. Reduction of performance is gradual for the scene categorization task, falling from 54.52 ± 0.26% (d' = 1.82) at full resolution down to 45.74 ± 0.25% ($d' = 1.57$) at 1/8 acuity. The localization tests, both with the regular and with the *opposite* condition, show similar, but not quite as stable behavior.

Although the performance decay is also apparent on the object categorization dataset, it remains mostly stable between full resolution (30.46 ± 0.16%, $d' = 1.95$) and reduction to 1/4 (30.44 ± 0.16%, $d' = 1.95$), but then falls off to 25.99 ± 0.22% ($d' = 1.80$) as the resolution is halved again.

Together, these results show that reducing dimensionality by pooling on midlevel simple features (i.e., S^2 features) leads to better performance than performing the pooling earlier, at the visual acuity level. They also show that such pooling is beneficial to the execution of holistic, contextual recognition tasks such as localization and scene categorization while it hinders performance in object classification.

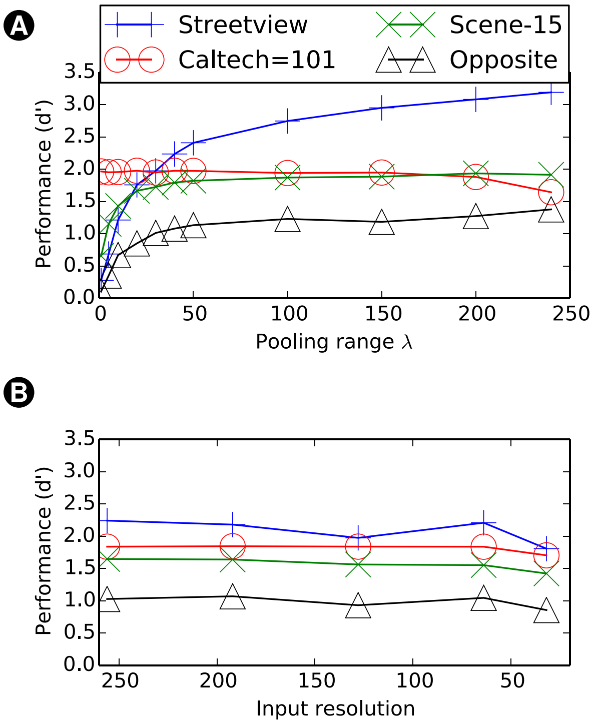

Figure 3. Classifier performance as d' on different datasets using the model described in Methods of the Peripheral pooling model section. (A) Variation of pooling range λ on constant input resolution of 256 × 256. (B) Variation of input resolution using constant $\lambda = 50$. Datasets test for location (Street View), object category (Caltech 101), place category (Scene 15 and location from 90° view angle (Opposite).

Model comparison

Methods

To test more widely which classes of visual features are discriminative for the location categorization tasks and put the achieved model performance into context, we also selected and tested bio-inspired low-level feature descriptors that cover different semantic aspects of the input image and implement different pooling schemes on top of densely sampled filters:

- Global pooling (as done in the *Texton*, *SIFT,* and parametric texture descriptor in this study) builds statistics over the whole image and discards all position information.
- Image partitioning (as done by the *Spatial Pyramid* and *GIST*) descriptors pool over segments of the image, resulting in a larger feature vector in which the feature index encodes the image partition.
- Local pooling (as done in the parametric V2 model from section 2.1) works like image partitioning, but allows neighboring pooling regions to overlap.

The *Texton* descriptor as described by Leung and Malik (2001) calculates a set of linear filter edge detection responses for each pixel, assigns clusters, and builds a histogram over the whole image. The *SIFT*, short for Scale Invariant Feature Transform descriptor by Lowe (1999), is a histogram of local orientations. We sample densely with a spacing of eight pixels and patch size of 16 pixels and use k-means clustering data to find 200 cluster centers from 100,000 random samples from 50 training images from each tested dataset. The *Spatial Pyramid* descriptor, introduced by Lazebnik et al. (2006), modifies this local histogram by pooling over a pyramid of image regions from global pooling down to a grid of 4 × 4 segments. The pyramid is run on the densely sampled SIFT descriptors. A second texture descriptor is the parametric texture model by Portilla and Simoncelli (2000) calculates joint statistics of complex wavelet coefficients and had been applied to the full image for this study. Although the descriptor has been used for texture description and synthesis, it has also been suggested as a model for peripheral vision (Freeman & Simoncelli, 2011; Rosenholtz, 2011). Additionally, it has been validated experimentally for peripheral tasks such as visual search (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012b). However, it has also been found to be an incomplete representation in a temporal three-alternative oddity task (Wallis et al., 2016), suggesting that the peripheral descriptor includes additional features not captured by local statistics alone.

The GIST descriptor (Oliva & Torralba, 2001) is a holistic image description designed for scene categorization, which calculates the first principal components of spectral image contents on a very coarse grid as well as on the whole image. It is calculated using the published MATLAB code from Oliva and Torralba (2001) in default settings on grayscale images (four scales, eight orientations per scale) without projection into the category space of the paper.

All descriptors are analyzed on the same datasets in the testbed, described in the Methods section of Peripheral pooling model, including the PCA where the feature dimensionality was larger than 256, to yield a performance in percent correct. To allow better comparison between datasets of different class count, performances are converted to d' by treating classification between m classes as a m-alternative-forced choice decision problem using the *pysdt* python toolbox function (Green & Dai, 1991). The V2 descriptor is using the model from the Methods section of Peripheral pooling model with maximum λ and resolution settings.

Results

Results of the model-dataset comparison are shown in Figure 4. For the localization task, all globally pooled texture-based approaches show strong performance. Best results are achieved by the Texton descriptor at 83.81 ± 0.11% ($d' = 3.81$). Even the low-dimensional parametric texture model (ParTex) reaches 71.97 ± 0.13% ($d' = 3.37$) here. On the other hand, the GIST descriptor performs worst of all the compared models at only 36.86 ± 0.10% ($d' = 2.38$). The Spatial Pyramid descriptor at 66.63 ± 0.14% ($d' = 3.21$) yields no improvement over its associated global pooling vector (SIFT) at 74.14 ± 0.11% ($d' = 3.44$). Generally, image partitioning methods as performed by the Spatial Pyramids as well as the GIST descriptor are not useful for the localization task. This is true for both the regular and the *opposite* condition.

An orthogonal effect can be found on the object categorization test on Caltech 101, where both spatially segmented descriptors show the strongest performance at 58.72 ± 0.11% ($d' = 2.74$, GIST) and 53.53 ± 0.17% ($d' = 2.60$, Spatial Pyramids).

The results of scene categorization are less diverse. Segmented descriptors perform well at 57.03 ± 0.21% ($d' = 1.90$) for Spatial Pyramids and 51.86 ± 0.26% ($d' = 1.75$) for GIST, but the global pooling vector presented in this paper (V2) reaches similar performance at 57.63 ± 0.18% ($d' = 1.91$).

In summary, we find that holistic pooling methods over texture descriptors perform better than spatially partitioned features on the localization task, whereas this effect is less pronounced on the place categorization problem and performance ordering is reversed for object classification.

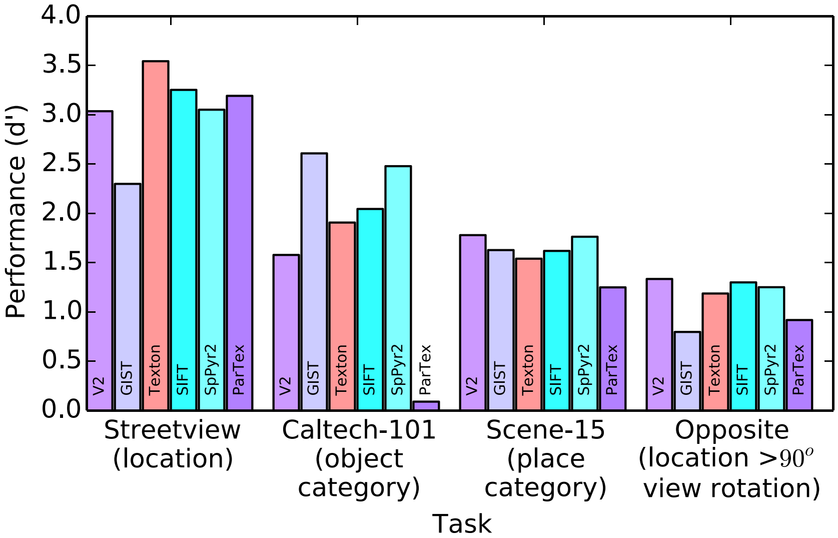

Figure 4. Comparison model performance as d' on different datasets. V2: Our pooling model. SpPyr2: Spatial Pyramid level 2 over SIFT descriptors. ParTex: Parametric texture model by Portilla and Simoncelli (2000). Percent correct overview in supplementary materials.

Human experiments

Methods

Participants

A test for how much peripheral vision contributes to human performance was run on 15 subjects (nine men, six women) aged 18 to 35 years per study requirement. Individual age was not recorded. All subjects were undergraduate students in digital media studies who took the study voluntarily for course credit. The study was acknowledged by the Institutional Review Board of Bremen University, and subjects gave informed consent.

Stimuli

Stimuli were shown in a match-to-sample paradigm. Subjects were placed in front of a projection wall (eye-wall distance of 186 cm, projection height and image size 194 cm) and instructed to decide which two of three presented images are from the same location. Images were flashed briefly in sequence (see Figure 5) with a reference image first and two test images following. Subjects answered which of the two test images was taken from the same location, but facing a different direction as the reference image. Answers were recorded without feedback.

Images were sampled randomly from the dataset described in the Peripheral pooling model section, but only a circular region up to 55° field of view is shown. A random location and view direction was picked as the reference image, then a correct answer was sampled from the remaining view directions ensuring a minimum view rotation of 10° yaw angle. The false answer was picked as a random sample from the remaining locations with a random view angle.

For the foveal presentation, image content reached from zero to b degrees and the peripheral presentation showed the range from b to $c = 27.5°$. The split angle b between foveal and peripheral presentation was chosen to approximate equal cortical coverage of the image projection into V1.

In order to find the split angle between peripheral and foveal stimulus display, we try to match cortical information between the two conditions. Images were 480 × 480 pixels in size and presented at a maximum

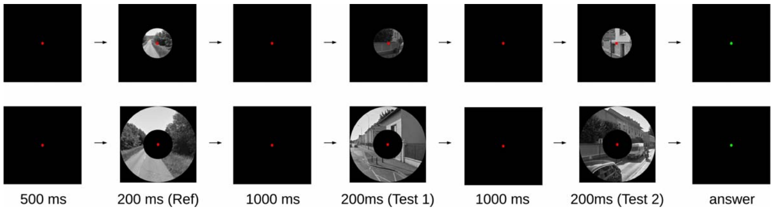

Figure 5. Trial structure for foveal (top) and peripheral (bottom) trial. Subjects first saw a reference image, then two test images. They had to answer which of the two test images was taken from the same location, but viewing into a different direction, as the reference image.

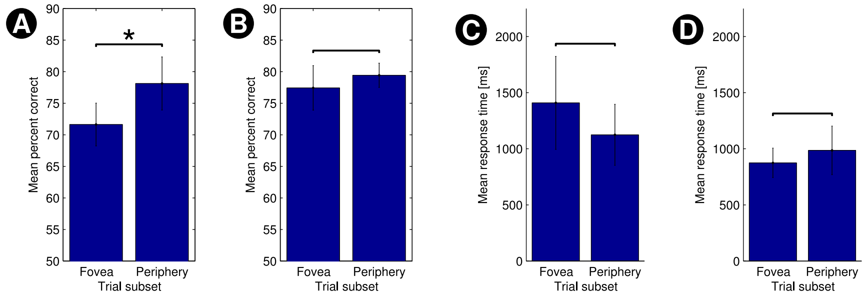

Figure 6. Human subjects results on localization study on peripheral versus foveal vision. (A) Mean percent correct answers for equal cortical mapping ($b = 10.1°$) condition. (B) Mean percent correct answers for equal radial coverage ($b = 13.8°$) condition. (C) Mean response times for equal cortical mapping condition. (D) Mean response times for equal radial coverage condition. Error bars are mean error. Significance is tested using a paired, two-tailed t test at $p < 0.05$.

eccentricity of $c = 27.5°$; this leads to a spatial display resolution of $\Delta x = 7'$. Because spatial resolution of the visual system is higher than Δx in the fovea, but the low image resolution limits the amount of information available to the fovea, constant cortical magnification $M(r) = k$ was assumed up to a radius $r = a$, where $a = 10°$ was determined by mapping spatial image resolution to the vernier discrimination thresholds found by Strasburger et al. (2011). Vernier discrimination thresholds were picked as the most conservative condition possible, assuming that the vision system responsible for making the location assessment cannot possibly make use of visual information sampled more densely than this value. For $r > a$, an inverse proportional falloff $M(r) = k \cdot a/r$ was used. The split angle b must satisfy the condition:

$$\int_{r=0}^{b} M(r)\,dr = \int_{r=b}^{c} M(r)\,dr \quad (8)$$

Which, assuming $b > a$, can be solved trivially for b:

$$\int_{r=0}^{a} k \cdot dr + \int_{r=a}^{b} \frac{a \cdot k}{r} dr = \int_{r=b}^{c} \frac{a \cdot k}{r} dr \quad (9)$$

$$\Leftrightarrow 1 + \log(b) - \log(a) = \log(c) - \log(b) \quad (10)$$

$$\Leftrightarrow b = \exp\left(\frac{\log(c) + \log(a) - 1}{2}\right) = \frac{\sqrt{a \cdot c}}{\exp(1)} \quad (11)$$

For our setup, this yields $b = 10.1°$. In a more conservative control condition, $b_c = c/2 = 13.8°$ was also tested. The control condition equalizes horizontal visual angle coverage between peripheral and foveal display. Because the same amount of image content is captured along the horizon line in this condition, it tests whether horizon elements in the far distance alone are sufficient to perform the localization task. Eight of the subjects were tested on the first and the other seven on the second condition.

Procedure

For the presentation, 100 peripheral and 100 foveal trials were intermixed randomly. Trials were allocated to four blocks with short breaks. Subjects could resume from the break by pressing a button. The complete experiment lasted about 20 min per subject.

The study has been executed in adherence to the Declaration of Helsinki.

Results

Results of the psychophysics experiment are shown in Figure 6. In the test condition of equal cortical coverage, participants show significantly higher performance on peripherally presented stimuli (78.13 ± 4.20%) compared with foveal trials (71.62 ± 3.39%). When fovea and periphery are split by equal radius ($b_c = c/2 = 13.8°$), peripheral presentations are answered correctly at 79.4 ± 1.9% and foveal presentations at 77.4 ± 3.5%. The difference between the groups is not significant for this condition.

Reaction times, measured as the delay from onset of the second test image to the answer, average to 849 ms and show no significant difference between the conditions.

Results by image overlap

Performances by view yaw difference between the reference and the correct test image are displayed in Figure 7. Because the maximum eccentricity of images was 27.5°, no image overlap is possible between reference and correct test image for trials with a yaw difference larger than 55°. As expected, performance is significantly lower for the trials with large ($\geq 55°$) change in yaw at 71.32 ± 2.82 percent correct versus 84.00 ± 3.00 in the small change condition (Figure 7A). For trials with small yaw difference, there is no significant difference in performance between periph-

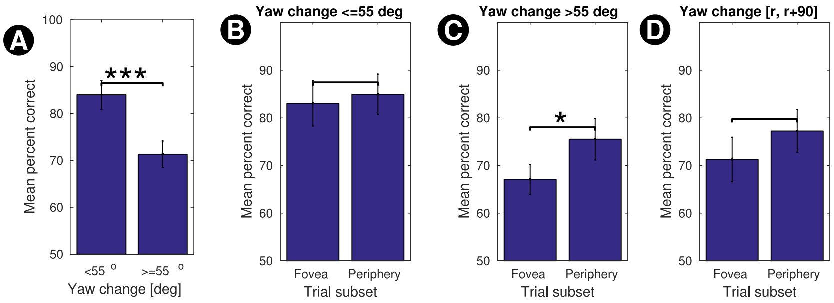

Figure 7. Evaluation by view overlap between reference image and correct test image for equal cortical mapping ($b = 10.1°$) condition. (A) Comparison of percent correct between trials that may have view overlap (view change $\leq 55°$) and trials that may not have overlap (view change $>55°$). (B) Comparison of percent correct between foveal and peripheral presentation for trials with small view change only. (C) The same comparison for trials with large view change. (D) Comparison of percent correct of all trials with a yaw difference between the outer radius of image content shown (r) and 90° outside of that value, i.e., $r = b = 10.5°$ for the foveal trials and $r = 55°$ for peripheral trials. Error bars are mean error. Significance is tested using a paired, two-tailed t test at $p < 0.05$.

eral and foveal presentation at 83.04 ± 4.74% for foveal and 84.96 ± 4.24% for peripheral display (Figure 7B). However, a strong difference can be found in the trials with large yaw difference. While performance for foveal trials is at 67.11 ± 3.15%, peripheral presentation leads to a significantly better value at 75.54 ± 4.37% (Figure 7C).

The performance comparison is also done over all trials in which the yaw difference lies between adjacent reference and test images, i.e., yaw difference $b = 10.5°$ in the foveal task and 55° in the peripheral task (Figure 7D). Despite the potential semantic advantage of visible image content being spatially closer to each other, the peripheral performance at 77.25 ± 4.46% exceeds the performance achieved on foveal trials at 71.28 ± 4.68%, but the margin is not significant.

Discussion

The main insights to be taken from the modeling and human subject experiments are:

- Peripheral vision plays an important role when humans match different views of locations.
- This role is stronger for large yaw differences, i.e., when there is no image overlap between matched views.
- Pooling over simple features leads to better recognition performance on holistic tasks than pooling earlier, i.e., directly over luminance values.
- The spatial pooling operation associated with peripheral vision contributes positively to the performance of a descriptor for localization, but not for object categorization tasks.

A common interpretation of peripheral vision would be that while specialization exists for detection of sudden object appearances or movements, it is just a "*crippled fovea*" for static image processing tasks. In this interpretation, information-consuming pooling is performed only to reduce the dimensionality to preserve processing energy and cortical space. However, our results suggest that this view should be refined in that the pooling operations performed by peripheral processing are actually tuned to perform certain tasks, i.e., the pooling operation performed in peripheral vision serves, in addition to dimensionality reduction, as an integral part of human visual processing. While foveal vision needs to identify individual elements, other important visual tasks include finding the context of the perceived scene (Oliva & Torralba, 2007), of which localization is certainly an integral part.

The performance gained by late pooling, i.e., using statistics over features, compared with early pooling modeled by image size reduction, is a result predicted by Rosenholtz (2011) because it preserves high spatial frequencies that are important for discrimination. Because receptor density is limited at high eccentricities, humans employ both early pooling, i.e., cutoff of spatial frequency resolution in the periphery (Loschky, McConkie, Yang, & Miller, 2005), as well as late pooling, resulting in crowding effects. Here, we show an advantage of the partial late pooling component over full early pooling.

The model comparison results (4) show that for a visual system that is capable of solving a variety of tasks including object and place recognition has to make trade-offs on architectural decisions such as pooling ranges. Note that there are more, often interdependent, decisions such as feature complexity, which nonlinearities to use, and which statistical properties to compute, which we have not controlled for all the models here. Because not all parameters could be tuned independently, other interpretations of the found results may also be considered. For example, the strong performance of GIST on the Caltech 101 set may be attributed to specialties in the dataset (Pinto, Cox, & DiCarlo, 2008). However, the reversal of performance ranks between the localization and the object recognition task reinforces the general idea that feature descriptors need to find a trade-off to be discriminative for individual tasks. Although the decrease in performance for larger pooling ranges on the object classification task found in this study is weaker than might be expected based on typical crowding results (Peripheral pooling model), it may underestimate the penalty that would be incurred by pooling in real-world object classification because of the cleaned-up nature and the low amount of clutter present in the Caltech 101 dataset. Because there is no single descriptor that solves all problems, a universal visual system may be implemented by including a collection of features tuned for distinct problem sets.

What are the features based on which models and human subjects solve the localization task? The large reduction in performance between the *Street View* and the *Opposite*-condition in the models (Model comparison section), as well as the performance loss induced by increasing the amount of yaw rotation between reference and test image (Figure 7A) show that a large portion of performance can be attributed to simple image matching. Whereas image matching is an important part of localization, it is not a unique or defining characteristic of this task. It is also expected that peripheral vision with its larger input range performs better because it can pool from a larger range of units in which the matched images may overlap.

The almost equal performance between periphery and fovea in the equal radius condition (Figure 6B) allows the hypothesis that localization can be solved well on environmental features appearing in the far distance along the horizon such as the skyline or distant foliage, since these conditions have equal horizontal coverage along the horizon line. But it's surprising that results would be equal despite the lowered spatial resolution in the periphery.

However, the experimental evaluation on the high yaw-difference condition (Figure 7C) shows that the peripheral advantage persists even when there is no image overlap between tested views. Because individual, highly discriminative objects cannot be matched in this case as they cannot be present in both views, recognition must be based on more general features that are typical for the location. Possible candidates are trees, foliage, and ground texture in natural scenes or construction styles in urban scenes. Peripheral vision, with its larger receptive fields, may be tuned to recognize such general features, which improves its role in the localization task. The model evaluation in the *Opposite*-condition (Result of Peripheral pooling model section) reinforces the idea that large pooling ranges on simple descriptors are useful features to describe a place.

When views of same adjacency are compared (Figure 7D), peripheral percent correct is higher than corresponding foveal performance, but the difference is no longer significant in the tested conditions. However, trials of much smaller yaw difference also contain a smaller amount of variation in content, which facilitates the task. Exactly how the effects of image content and different recognition methods between foveal and peripheral tasks interact cannot be derived from the study. Because peripheral vision naturally provides different semantic features like ground and sky elements, it could be argued that in addition to being tuned to the localization task, the periphery is also naturally exposed to more stimuli, which may be useful for this purpose. This reinforces the idea suggested in this paper that the periphery is also tuned functionally to perform well in the localization task.

For the sake of a controlled experiment, we define the localization task in a very artificial way by comparing only different views taken from the exact same location. But since we find that a peripheral advantage exists even on views without overlap, we assume that it should be general enough to also transfer to the more useful real-world task of recognizing places within a proximity radius.

A critical issue for the current investigation is what is regarded as central and as peripheral visual field. The central visual field is often seen to comprise the fovea and parafovea, and border between centrum and periphery would thus be located at 5°. Our split between central and foveal presentations (at 10.1°) in the experiment has been much more conservative toward a larger foveal display, so we would expect superiority over peripheral vision for the task using such a setting. We have decided here to use a different criterion and to look for approximately equal cortical areas for central and peripheral vision. We have assumed that the amount of cortical machinery does only reflect the local spatial resolution. Since the resolution of our images has been lower than the foveal resolution we have accordingly assumed a fictitious cortical area significantly smaller than the actual cortical area devoted to the processing of the central 10.1°. In this respect, our data are based on a bias for central vision.

Larson and Loschky (2009) performed a similar experiment with respect to gist recognition. For their setting, they calculated an area-balanced radius between 2.38° and 3.13°. Although the settings are not directly comparable, this can be seen as an indication for the conservatism of our estimate. They further found a crossover point for equal performance at a radius of 7.4°. It would be desirable for future investigations to measure such an equal-performance radius also for the localization task investigated here.

It is further interesting to consider the role of the gist computation in our model simulations. Contrary to expectation, the gist model yields only a weak performance in the localization tasks. Since human gist extraction works fine in peripheral vision (Larson & Loschky, 2009; Thorpe et al., 2001) this may point to a suboptimal implementation of the gist model. Using suitable texture features and pooling ranges instead of low-frequency information may enable the development of a more realistic gist model.

One reason why the gist descriptor works on the place categorization set but not on localization might be due to bias toward certain scene arrangements induced by the source of the photos, i.e., a dataset issue (Ponce et al., 2006).

The periphery exhibits spatial feature pooling over extended areas, which is thought to be the basis of crowding phenomena. For many perceptual processes this has been found to cause destructive effects. Here we found that this pooling can also have a beneficial effect. For the task of vision-based localization, we could demonstrate a clear advantage if basic features are pooled.

This seems to be a general principle that does not depend on the exact stage of processing on which this pooling takes place. However, it should be noted that if the full pooling would be performed early, e.g., as low resolution input to the cortex, then certain features will be destroyed and will be no longer available for pooling at higher stages. This alone suggests that a step-wise pooling on several subsequent stages would be more efficient than a one-step pooling on an early stage.

In conclusion, we have shown that the periphery is not always inferior to foveal vision, as is the case for a large number of perceptual tasks (Levi, 2008). For the special task of vision-based localization, the spatial pooling of features in peripheral vision can indeed cause a genuine advantage.

Keywords: peripheral vision, localization, modeling, psychophysics

Acknowledgments

Funding provided by DFG (SFB/TR 8 Spatial Cognition, project A5) and by the German Aerospace Center (DLR) with financial means of the German Federal Ministry for Economic Affairs and Energy (BMWi), project "KaNaRiA" (grant No. 50 NA 1318).

Commercial relationships: none.

Corresponding author: Sven Eberhardt.
Email: sven2@uni-bremen.de.
Address: Cognitive Neuroinformatics, University of Bremen, Bremen, Germany.

References

Anstis, S. (1998). Picturing peripheral acuity. *Perception*, *27*, 817–826.

Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision Research*, *14*(7), 589–592.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12):13, 1–18, doi:10.1167/9.12.13. [PubMed] [Article]

Bulakowski, P. F., Post, R. B., & Whitney, D. (2011). Reexamining the possible benefits of visual crowding: Dissociating crowding from ensemble percepts. *Attention, Perception, & Psychophysics*, *73*(4), 1003–1009.

Chaney, W., Fischer, J., & Whitney, D. (2014). The hierarchical sparse selection model of visual crowding. *Frontiers in Integrative Neuroscience*, *8*, 1–11.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*, pp. 886–893.

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, *159*(2), 203–221.

Doersch, C., Singh, S., & Gupta, A. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics*, *31*(4).

Eberhardt, S., & Zetzsche, C. (2013). Low-level global features for vision-based localization. In M. Ragni, M. Raschke, & F. Stolzenburg (Eds.), *KI 2013 Workshop on Visual and Spatial Cognition*, *1055*, pp. 5–12, Koblenz, Germany: CEUR Workshop Proceedings.

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, *106*(1), 59–70.

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Attention, Perception, & Psychophysics*, *74*(2), 379–396.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics, 5*, 826–834.

Green, D. M., & Dai, H. (1991). Probability of being correct with 1 of M orthogonal signals. *Attention, Perception, & Psychophysics*, *49*(1), 100–101.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision*, *15*(6):5, 1–18, doi:10.1167/15.6.5. [PubMed] [Article]

Isik, L., Leibo, J. Z., Mutch, J., Lee, S. W., & Poggio, T. (2011). A hierarchical model of peripheral vision. *Technical Report MIT-CSAIL-TR-2011-031, CBCL-300.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural Networks. In *Advances in neural information processing systems* (pp. 1097–1105). Washington, DC: IEEE Computer Society.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, *9*(10):6, 1–16, doi:10.1167/9.10.6. [PubMed] [Article]

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 2 (pp. 2169–2178). Washington, DC: IEEE Computer Society.

Leung, T., & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, *43*(1), 29–44.

Levi, D. M. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654.

Loschky, L., McConkie, G., Yang, J., & Miller, M. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition*, *12*(6), 1057–1092.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision, vol. 2, 1999* (pp. 1150–1157). Washington, DC: IEEE Computer Society.

Marko, H., & Giebel, H. (1970). Recognition of handwritten characters with a system of homogeneous layers. *Nachrichtentechnische Zeitschrift*, *23*(9), 455.

Mutch, J., Knoblich, U., & Poggio, T. (2010). *CNS: A GPU-based framework for simulating cortically-organized networks.* Technical Report MIT-CSAIL-TR-2010-013 / CBCL-286, Cambridge, MA: Massachusetts Institute of Technology.

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, *696*(64), 251–258.

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.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744.

Pelli, D. G., Palomares, M., & Majaj, N. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12):12, 1136–1169, doi:10.1167/4.12.12. [PubMed] [Article]

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135.

Pinto, N., Barhomi, Y., Cox, D. D., & DiCarlo, J. J. (2011). Comparing state-of-the-art visual features on invariant object recognition tasks. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)* (pp. 463–470). Washington, DC: IEEE Computer Society.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), e27.

Polyak, S. L. (1941). *The retina.* University of Chicago Press: Chicago.

Ponce, J., Berg, T. L., Everingham, M., Forsyth, D. A., Hebert, M., Lazebnik, S., . . . Zisserman, A. (2006). Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, & A. Zisserman (Eds.), *Toward category-level object recognition* (pp. 29–48). Berlin/Heidelberg, Germany: Springer.

Portilla, J., & 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.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rodieck, R. W. (1998). *The first steps in seeing, vol. 15.* Sunderland, Massachusetts: Sinauer Associates.

Rosenholtz, R. (2011). What your visual system sees where you are not looking. In *Proceedings of SPIE, 7865*, Human Vision and Electronic Imaging XVI, 786510.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012a). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology*, 3.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012b). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, doi:10.1167/12.4.14. [PubMed] [Article]

Schwartz, E. L. (1994). Computational studies of the spatial architecture of primate visual cortex. In *Primary visual cortex in primates* (pp. 359–411). Berlin, Germany: Springer.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, USA, 104*(15), 6424–6429.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, *11*(5):13, 1–82, doi:10.1167/11.5.13. [PubMed] [Article]

Tacchetti, A., Mallapragada, P. K., Santoro, M., & Rosasco, L. (2011). GURLS: A toolbox for large scale multiclass learning. Presented at Workshop: "Big Learning: Algorithms, Systems, and Tools for Learning at Scale" at NIPS 2011, December 16–17, 2011, Sierra Nevada, Spain.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience*, *14*(5), 869–876.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766.

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLoS Computational Biology*, *6*(1), e1000646.

Virsu, V., Näsänen, R., & Osmoviita, K. (1987). Cortical magnification and peripheral vision. *Journal of the Optical Society of America A*, *4*(8), 1568–1578.

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, *37*(3), 475–494.

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision*, *16*(2):4, 1–30, doi:10.1167/16.2.4. [PubMed] [Article]

Weymouth, F. W. (1958). Visual sensory units and the minimal angle of resolution. *American Journal of Ophthalmology*, *46*(1), 102–113.

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168.

Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *Journal of the Optical Society of America A*, *14*(9), 2057–2068.

Wilkinson, M. O., Anderson, R. S., Bradley, A., & Thibos, L. N. (2016). Neural bandwidth of veridical perception across the visual field. *Journal of Vision*, *16*(2):1, 1–17, doi:10.1167/16.2.1. [PubMed] [Article]

Wilson, H. R., Levi, D., Maffei, L., Rovamo, J., & DeValois, R. (1990). The perception of form: Retina to striate cortex. In L. Spillmann, (Ed.), *Visual perception: The neurophysiological foundations* (pp. 231–272). San Diego, CA: Academic Press.

Wolter, J., Reineking, T., Zetzsche, C., & Schill, K. (2009). From visual perception to place. *Cognitive Processing*, *10*, 351–354.