Published in final edited form as:
Vision Res. 2022 November 01; 200: 108083. doi:10.1016/j.visres.2022.108083.

Human peripheral blur is optimal for object recognition

R T Pramod[1,*,#], **Harish Katti**[2,*,#], **S P Arun**[2,#]

[1]Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012

[2]Centre for Neuroscience, Indian Institute of Science, Bangalore 560012

Abstract

Our vision is sharpest at the centre of our gaze and becomes progressively blurry into the periphery. It is widely believed that this high foveal resolution evolved at the expense of peripheral acuity. But what if this sampling scheme is actually optimal for object recognition? To test this hypothesis, we trained deep neural networks on “foveated” images mimicking how our eyes sample the visual field: objects (wherever they were in the image) were sampled at high resolution, and their surroundings were sampled with decreasing resolution away from the objects. Remarkably, networks trained with the known human peripheral blur profile yielded the best performance compared to networks trained on shallower and steeper blur profiles, and compared to baseline state-of-the-art networks trained on full resolution images. This improvement, although slight, is noteworthy since the state-of-the-art networks are already trained to saturation on these datasets. When we tested human subjects on object categorization, their accuracy deteriorated only for steeper blur profiles, which is expected since they already have peripheral blur in their eyes. Taken together, our results suggest that blurry peripheral vision may have evolved to optimize object recognition rather than merely due to wiring constraints.

Introduction

Our retina contains 100 times more photoreceptors at the center compared to the periphery (Curcio and Allen, 1990; Curcio et al., 1990). It is widely believed that this sampling scheme saves on the metabolic cost of processing orders of magnitude more information that would result from full resolution scenes without affecting overall performance (Weber and Triesch, 2009; Akbas and Eckstein, 2017). But what if this sampling scheme evolved to optimize object recognition?

There are several lines of evidence suggesting that peripheral blurring can benefit object recognition. First, object detection by humans on natural scenes is slowed down by clutter as well as by partial target matches in the background (Katti, Peelen, & Arun, 2017). Thus, high spatial frequency information in the periphery interferes with recognition. Second, the

[#]Correspondence to: pramodrt9@gmail.com, harish2006@gmail.com & sparun@iisc.ac.in.
[*]Both authors contributed equally

Author Contributions
PRT, HK & SPA designed the study, interpreted the results and wrote the manuscript. PRT & HK implemented foveation, trained neural networks and collected data.

surrounding scene context can facilitate recognition (Li et al., 2002; Bar, 2004; Davenport and Potter, 2004; Larson and Loschky, 2009; Katti et al., 2017) but this information is contained in low spatial frequencies (Morrison and Schyns, 2001; Torralba, 2003; Bar, 2004; Torralba et al., 2006). Thus, sampling images densely near objects and sparsely in the surrounding context might be beneficial for recognition.

To further illustrate how such a sampling scheme benefits recognition, consider the example scene in Figure 1A. When this scene is given as input to a state-of-the-art pre-trained deep neural network (R-CNN; see Methods), it correctly identified the person but made a false alarm to a traffic cone in the background. We then “foveated” the image by resampling it at full resolution on the salient object (the person) and sampling it sparsely into the periphery according to the human blur function. The same deep network no longer showed the false alarm (Figure 1B). Thus, peripheral blurring can be beneficial in avoiding spurious target matches far away from objects of interest. Note that just fixating on the foreground object or objects (with no peripheral information at all) may not by itself yield enough information about object identity to achieve high classification accuracy (Leek et al., 2016).

We note that these arguments all pertain to visual information processing after a fixating on an object. Understanding foveal processing at the level of a single fixation is important since primates can recognize objects even with short stimulus durations at which multiple fixations are impossible (Thorpe et al., 1996; Hung et al., 2005). This core recognition behavior is also predicted by neural responses in high-level visual areas during fixation (Hung et al., 2005; Ratan Murty and Arun, 2015; Rajalingham et al., 2018). It is also supported by recent observations that high-level visual areas are organized by visual eccentricity (Arcaro and Livingstone, 2017; Gomez et al., 2019; Stewart et al., 2020).

Overview of this study

We set out to investigate whether peripheral blur leads to improvements in object recognition performance. To this end, we trained feedforward convolutional deep neural networks trained on images that were “foveated”, that is, sampled densely near objects in the scene and coarsely into the periphery. We tested two predictions. First, we predicted that, if peripheral blur is beneficial for object recognition, then neural networks trained on images with peripheral blur should show better performance. Furthermore, if the peripheral blur of our eyes are optimal for recognition, then neural networks trained on the human peripheral blur function should show the best performance. Second, we predicted that, applying peripheral blur to images should affect object categorization in humans only if the blur profile is steeper than our own peripheral vision.

Methods

Generating foveated images

Any visual stimulus can be analysed in terms of its spatial frequency content with fine details (like edges) attributed to high spatial frequencies and coarse information (like object shape) attributed to low spatial frequencies. The range of visible spatial frequencies is usually measured as the sensitivity to contrast at each spatial frequency and is summarized by the contrast sensitivity function (CSF) which varies as a function of retinal eccentricity (Campbell and Robson, 1968). Based on previous research using grating stimuli for simple detection/discrimination tasks, the contrast threshold for detecting a grating patch of spatial frequency *f* at an eccentricity *e* is given by

$$CT(f,e) = CT_0 \exp(\alpha f \frac{e+e_2}{e_2}) \tag{1}$$

where *f* is spatial frequency (cycles per degree), *e* is the retinal eccentricity (degrees), CT_0 is the minimum contrast threshold, α is the spatial frequency decay constant, and e_2 is the half-resolution eccentricity. We took the values of these variables to be $CT_0 = 0.0133$, $\alpha = 0.106$, $e_2 = 2.3$ respectively. This formula matches contrast sensitivity data measured in humans under naturalistic viewing conditions (Geisler and Perry, 1998). Although the above formula gives the contrast threshold, what is more important is the critical eccentricity e_c beyond which the spatial frequency *f* will be invisible no matter the contrast. This critical eccentricity for each such spatial frequency *f*, can be calculated by setting the left-hand side of the equation above to 1 and solving for *e*.

$$e_c = \frac{e_2}{\alpha f} \ln\left(\frac{1}{CT_0}\right) - e_2 \tag{2}$$

The above equation for critical eccentricity (in degrees) was then converted to pixel units by considering the viewing distance. Specifically, critical eccentricity in cm is calculated using the formula

$$e_{c,cm} = \mathrm{d} * \tan\frac{\pi e_c}{180} \tag{3}$$

where $e_{c,cm}$ is the critical eccentricity beyond which spatial frequency *f* (equation 2) will be unresolvable at a viewing distance *d* (in cm) (see below for choice of *d*). This was then converted into pixel units using dot-pitch of the monitor (in cm).

$$e_{c,px} = \frac{e_{c,cm}}{pitch} \tag{4}$$

The dot-pitch value of the monitor in our experiments was 0.233 cm. Then, the input image was low-pass filtered and down-sampled successively by a factor of two, to create a multi-resolution scale pyramid having up-to seven levels. Further, *f* in the above equation for ec was set to be the Nyquist frequency at each level of the multi-resolution scale pyramid and the resulting values of e_c were used to define the foveation regions at each level. That is, pixel values for the foveated image were chosen from different levels of the multi-resolution scale pyramid according to the eccentricity of the pixel from the point of fixation. In our experiments, in addition to using the default values of all the parameters, we obtained different foveation blur profiles by modulating α by a spatial decay factor γ.

$$\alpha_{\mathrm{new}} = \alpha \ \gamma \ \text{for} \ \gamma = \{0.25, 0.5, 1, 2, 4\} \tag{5}$$

where γ is the spatial decay factor with $\gamma = 1$ being the human foveation blur profile (Equation 1).

Example object detection with and without peripheral blur

To illustrate object detection with and without peripheral blur, we took a pre-trained deep neural network (Faster R-CNN) that yields state-of-the-art performance on object detection (Ren et al., 2015). This network had been pre-trained to identify instances of 20 different classes including people. To this neural network we gave as input both the full-resolution scene as well as the foveated image with human peripheral blur. The resulting object detections for the “person” class are depicted in Figure 1.

CNN training: VGG-16 architecture trained on ImageNet with foveation

To test if foveation is computationally optimal for object recognition in natural scenes, we chose ~500,000 images from the ImageNet dataset with manual object level bounding box annotations. We created 5 foveated versions of each image with the point of foveation fixed at the centre of the bounding box and trained deep neural networks for object recognition. Specifically, we used VGG-16 architecture and trained 18 separate networks (three each for the full resolution and different foveated versions of the image). Note that, all foveated images were created after scaling the image to 224x224 pixels which is the default size of input to the VGG-16 network. To create images with different levels of foveal blur, we used the equations described in the previous section. The output of those equations depends crucially on the distance between the observer and the image.

How do we find the viewing distance for the deep network? To estimate the optimal viewing distance, we trained separate networks on images foveated with a viewing distance of 30, 60, 90, 120 and 150 cm. We obtained consistent improvements in performance for all choices of viewing distance, but the best performance was obtained for a viewing distance of 120 cm. We used this value for all the reported analyses. However we confirmed that our results are qualitatively similar for other choices of viewing distance (Section S1).

We used the standard VGG-16 architecture (Simonyan and Zisserman, 2014) that has 13 convolutional layers and 3 fully connected layers. The layer specifications are as follows: two 224x224x64 convolution layers with a kernel size of 3 and a stride of 1 is followed by max-pooling layer with a window size of 2x2 and a stride of 2. This is followed by two 112x112x128 convolution layers with a 2x2 max-pool layer, which is then followed by three 56x56x256 convolutional layers and a max-pool layer. Then, the network has three 28x28x512 layers followed by a max-pool layer, three 14x14x512 convolutional layers followed by another max-pool layer, followed by a 7x7x512 convolutional layer that is flattened into 1x1x4096 and then mapped to a flattened 1x1x1000 layer corresponding to the 1000 training categories. The 1000-dimensional activation vector is then subjected to a SoftMax operation to get the belief of the network for the presence of each of the 1000 categories.

We trained each model with 3 random initializations (seeds 1-3) for a total of 18 models (3 seeds x 6 foveation levels). For each network, we started with randomly initialized weights and trained the network for 1000-way classification till accuracy on the held-out set (an unseen subset of the training set) plateaued out. We observed that this held-out set performance usually plateaued out after training the network for 70 epochs, and hence we used this as a stopping criterion. We trained each network without batch normalization with a batch size of 32 using a cross-entropy loss metric and stochastic gradient descent optimizer. All networks were defined and trained using python scripts using the PyTorch framework on NVIDIA TITAN-X/1080i GPUs. All the trained models were tested for generalization capabilities on a corresponding test set containing 50,000 images (ImageNet validation set).

Evaluation of spatial frequency content

To explore the relationship between spatial frequency content and object recognition, we selected 11 random categories from the ImageNet validation dataset - these were categories 1:100:1000 from ImageNet, which included common objects like fish, bird, animal, insect, clothing, building etc. We rescaled all images to have at least 500 pixels along both dimensions and chose 100 pixels x 100 pixels patches on concentric circles with radii 0, 50, 100, 150 and 200 pixels from the centre of the image. These patches were chosen along 8 equally spaced directions on the circle with the exception of the patch at the centre which was considered only once. We then extracted low and high spatial frequency from a bank of Gabor filters tuned for six spatial frequencies (0.06, 0.09, 0.17, 0.25, 0.33 and 0.5 cycles/pixel) and 8 orientations (uniformly sampled between 0 and 180 degrees). We then trained linear object identity decoders at both foveal as well as peripheral locations on the concatenated filter responses across all patches corresponding to high or low spatial frequencies.

Experiment 1: Animal detection task

All experiments were conducted in accordance to an experimental protocol approved by the Institutional Human Ethics Committee of the Indian Institute of Science. Subjects had normal or corrected-to-normal vision, gave written informed consent and were monetarily compensated for their participation.

Subjects—A total of 58 subjects (18-52 years, 22 females) participated in this experiment.

Procedure—Subjects were comfortably seated ~60 cm from a computer monitor with a keyboard to make responses. Image presentation and response collection was controlled by custom scripts written in MATLAB using Psychtoolbox (Brainard, 1997). Each trial began with a fixation cross at the centre of the screen shown for 500ms followed by the image. All images measured 640 x 480 pixels, and subtended 13.5° in visual angle along the longer dimension. Images were shown at the centre of the screen for 100 ms followed by a white-noise mask. The noise mask stayed for 5 s or till the subject responded, whichever was earlier. In practice, the mask stayed on for an average of 430 ms after the stimulus offset based on the participants’ responses. Subjects were instructed to maintain fixation on a fixation cross at the centre of the image and respond as quickly and as accurately as possible to indicate whether the image contained an animal or not (‘a’ for animals and ‘n’ otherwise).

Stimuli—We created three groups of foveated images with spatial decay factors of 0.25, 1 and 4. For each group, we chose 212 full resolution images of animals (for example: duck, dog, elephant, fowl, deer, rabbit, ostrich, buffalo) and an equal number of images of inanimate objects (for example: boat, bicycle, wheelbarrow, airplane, flowerpot, tower, hot air balloon, letterbox, car). All images were chosen from the ImageNet validation set. The retinal sizes of key objects in the animate and inanimate categories were comparable (average bounding box area normalized to the total image area: 0.17 for animate and 0.18 for inanimate; p = 0.1 for a ranksum test across images). In all, there were 1696 images (424 images of animals and inanimate objects x 4 levels of foveation). Subjects saw 424 images (212 each of animals and inanimate objects) such that each image was shown in only one of the foveated conditions. This was achieved by dividing the set of 212 category images into 4 mutually exclusive subsets each with 53 images and picking one of these subsets for presentation. We repeated this procedure for all versions (one full resolution and three foveated) and chose non-overlapping subsets of images across versions for the experiment. Each subject saw 424 images, and a given image was shown to 14 subjects.

Experiment 2: Person detection task

All methods were identical to Experiment 1, except for those detailed below.

Subjects—A total of 31 subjects (18-36 years, 12 female) participated in the task.

Stimuli—We chose 120 images that contained people and 120 images that had other objects (for example: dog, bird, boat, dustbin, post-box, bench, window, chair). These images were chosen from the publicly available MS-COCO (Lin et al., 2014) and ImageNet (Russakovsky et al., 2015) datasets. Like in the animal task, we generated three foveated versions of each image wherein one version had a foveal blur that matched the human contrast sensitivity function and two additional ones that were shallower or steeper (spatial decay factors 1, 0.25 or 4). Every participant saw a given scene only once across all versions. Here too, the white-noise mask stayed on for an average of 474 ms after the stimulus offset.

Results

If peripheral blur is optimal for recognition, then it follows that object classifiers trained on foveated images (with high resolution near objects and gradual blur into the periphery) should progressively improve recognition until performance peaks for the human peripheral blur profile. We tested this hypothesis by training state-of-the-art deep neural network architectures on foveated images with varying peripheral blur profiles.

To train these deep networks, we selected images from the widely used ImageNet dataset (~500,000 images annotated with object category and location across 1,000 object categories). Importantly, these images are photographs taken by humans in a variety of natural viewing conditions, making them roughly representative of our own visual experience. To obtain foveated images, we started with the well-known human contrast sensitivity function (CSF) measured at different eccentricities from the fovea (Geisler and Perry, 1998). Although the human CSF was measured in degrees of visual angle, the maximum resolvable spatial frequency is limited by the display properties (dot pitch) and resolution of the digital image. The peripheral blur function thus obtained assumes a specific distance between the observer and the scene, called the viewing distance, which is typically set by the experimenter. However, we do not have access to the viewing distance between the scene and the observer (i.e., the camera) in the ImageNet dataset. To overcome this issue, we tried different values of the viewing distance and found that value of 120 cm gave the best object recognition performance on ImageNet for the network architecture used. We obtained qualitatively similar results upon varying the viewing distance (Section S1).

To vary the peripheral blur profile, we fitted this function to an exponential and modified its spatial decay by a factor of 0.5, 1, 2 or 4 (Figure 2A). We then applied this blur profile to each image, centred on the labelled object (see Methods). Example images with varying degrees of peripheral blur are shown in Figure 2. A spatial decay factor smaller than 1 indicates shallower decay than human peripheral blur, i.e. the image is in high resolution even into the periphery (Figure 2B). A value of 1 indicates images blurred according to the human peripheral blur function (Figure 2C). A value larger than 1 indicates steeper decay i.e. the image blurs out into the periphery much faster than in the human eye (Figure 2D).

Foveation leads to increased object recognition performance

Next we trained a widely used deep convolutional neural network architecture (VGG-16; Figure 3A) for 1000-way object classification on the ImageNet images with bounding box annotations. We took three random initializations of the VGG-16 architecture, and trained each of them on a total of six image sets. These six image sets corresponded to one full resolution image (no foveation) and five spatial decay factors of 0.25, 0.5, 1, 2 and 4 (Figure 3B). We then tested each network for its generalization abilities by evaluating its classification accuracy on novel images foveated with the corresponding spatial decay factor. The performance of these networks is summarized in Table 1.

Across all networks, the networks trained on images foveated according to the human peripheral blur function gave the best performance (average Top-1 accuracy = 47.4%; Top-5 accuracy = 71.4%; Figure 3B; Table 1). This performance was significantly better than the network trained on full-resolution images (Increase in top-1 accuracy: mean ± std: 0.5% ± 0.24% across 1000 categories; $p < 0.05$, signed-rank test; increase in top-5 accuracy, mean ± std: 0.53% ± 0.23%, $p < 0.05$, signed-rank test for the 1000 class-wise accuracies each for the best model trained on full-resolution and human-like foveated images). Note that the absolute accuracies of all our networks are smaller than those typically reported (Simonyan and Zisserman, 2014), even though these networks are trained to saturation. This is because we could only train our networks on the images with available bounding box annotations, which constituted only half of the ImageNet dataset (~500,000).

To investigate the underlying reasons behind the improved performance of the network trained on foveated images, we reviewed images that were correctly classified after foveation but were misclassified without foveation (Figure 3C). We observed two types of benefits. First, foveation helped to disambiguate between similar categories, such as in the “digital watch” and “freight car” images. Here, the full-resolution network incorrectly classified these images as “digital clock” and “passenger car” but the foveated network correctly classified them. Likewise the “airliner” is classified as “war plane” and “spacecraft” with higher probability than “airliner” itself by the full-resolution network but is correctly classified after foveation. Second, foveation improved the quality of top-ranked guesses as in the case of “dalmatian” where the full-resolution network determined other categories as more likely (trilobite, hook, necklace). The foveated network also made reasonable guesses for the other likely categories (Great Dane, English Foxhound, etc).

To quantify these patterns across all categories, we calculated the average rank of the correct class label in the networks trained without and with foveation. As expected, the average rank was significantly smaller for the network trained with foveation (rank of correct class label for best network, mean ± std: 19.72 ± 65.2 & 18.5 ± 61.5 for network trained without and with foveation respectively, $p < 0.05$, sign-rank test).

Next, we wondered whether the improvement due to foveation came from improved performance on specific object classes with large size, or with more eccentric position in the image, or with more clutter. To this end, we calculated the correlation between the foveation improvement for each category with the average object size, average object horizontal and vertical position, average number of interest points in the image (as a proxy for clutter; calculated using SURF feature detection -- *detectSURFFeatures* function in MATLAB). This revealed no systematic correlations (r = -0.05, 0.02, -0.04 and 0.02 for object size, object horizontal and vertical position and interest points respectively; $p > 0.1$ in all cases).

Next we wondered whether these results would generalize to other image datasets or neural network architectures. To this end we trained a ResNet-18 architecture (He et al., 2016) on ImageNet and another custom neural network architecture for person categorization over images chosen from the MSCOCO database (Lin et al., 2014). In both cases, using human-like peripheral blur yielded optimal performance (Section S2 and S3 for ResNet-18 and MSCOCO results respectively).

Evolution of the foveation advantage across neural network training

In the above results, the overall improvement of the network with human-like foveation could arise from improved detection of objects, or a decrease in the rate of false alarms. It could also arise early or late during training which may further elucidate the nature of the underlying features. To investigate this possibility, we saved the model weights every five epochs during training and calculated the overall percentages of hits and false alarms. We then calculated hits and false alarms over the course of learning for two networks: the best network (with human-like foveation) and the network trained on full resolution images (*no foveation*).We found that the improvement in accuracy for the foveated network largely came from both an increase in the hits (Figure 4A) and a reduction in false alarms (Figure 4B). This trend emerged very early during network training and remained consistent through the course of training for all three seeds. Thus, the network trained on foveated images achieves greater accuracy fairly early on during training and learns faster.

Evaluation of relevant spatial information

The above results demonstrate that human-like foveation is optimal for object recognition. This raises the intriguing possibility that foveation in the eye may have evolved to optimize object classification. Did this evolution require a complex neural network architecture, or could it arise from simpler feature detectors? To examine this possibility, we wondered whether the image features most useful for recognition vary progressively with distance from the object in a scene. Specifically, we predicted that the low spatial frequency information is more discriminative for object recognition at peripheral locations whereas high spatial frequency information is more relevant at the fovea. If this is true, then even simple classifiers based on spatial frequency features could potentially drive the evolution of foveal vision.

To verify this, we selected a subset of 11 categories from the ImageNet validation dataset. For each image, we extracted image patches at varying distances from the center and used a bank of Gabor filters to extract low and high spatial frequency filter responses from each image patch. We then trained linear classifiers on the responses of each spatial frequency filter to image patches at a particular distance from the centre. The results are summarized in Figure 5A.

Object decoding accuracy was significantly higher than the chance performance (1/11 = 9%) at all eccentricities and all spatial frequencies, indicating that there is object-relevant information at all locations and frequencies (Figure 5). However, it can be seen that classification accuracy was best for high spatial frequency features at the center, and best for low spatial frequency into the periphery. Thus, even simple detectors based on spatial frequency features show an advantage for sampling densely at the center and sparsely in the periphery.

Human categorization on foveated images

Our finding that human-like foveation is optimal for recognition is based on training neural networks. We therefore wondered how image categorization by humans would change across varying peripheral blur profiles. Since human eyes are already equipped with the typical peripheral blur profile, we predicted that foveating images with spatial decay factor of less than 1 should have no effect on recognition performance. Further, viewing images with steeper blur profiles should lead to reduced performance, due to the lack of useful low-frequency features in the periphery.

We evaluated these predictions using several behavioural experiments on humans. In Experiment 1, subjects had to indicate whether briefly presented scene contained an animal or not (see Methods). An example image is shown in Figure 6A. We used four types of images: full resolution and three levels of foveation with spatial decay factors of 0.25, 1 and 4. Critically, to avoid memory effects, subjects saw a given scene only once across all levels of foveation.

Subjects were highly accurate on this task (accuracy, mean ± std: 94% ± 1.1% across the four types of images). Importantly, accuracy was significantly lower for steeply foveated images (spatial decay factor = 4) compared to other variants (average accuracy: 93% for steeply foveated images and 94.9%, 94.6% and 94.5% for full resolution, and images with spatial decay factors of 0.25 and 1 respectively; $p < 0.005$ for ranksum test on average accuracies for foveated images with spatial decay factor of 1 vs 4; Figure 6B). Further, subjects' accuracy was comparable for full resolution images and human foveated images with spatial decay factor of 1 (p = 0.29 using ranksum test on average accuracies across images).

We found similar but stronger effects of foveation on reaction times. Reaction times were slowed down only for the highest spatial decay factor (reaction times, mean ± std: 529 ± 102 ms, 523 ± 93 ms, 527 ± 95 ms and 545 ± 98 ms for full resolution images, and foveated images with spatial decay factors of 0.5, 1 and 4 respectively; p < 0.0005 for ranksum test on reaction times for human foveated and steep foveated images, p > 0.05 for all other pairwise comparisons; Figure 6C).

Next, we asked whether these results would generalize to other categories. To this end, in Experiment 2, subjects had to detect the presence of people in an image (example scene in Figure 6D). Subjects were highly accurate in detecting the target object (accuracy, mean ± std across subjects: 75.3% ±1.5% across the four types of images). As with the animal categorization task, accuracy was lowest for steeply foveated images (average accuracy: 76% for full resolution; 76.8% for shallow foveation; 75.4% for human foveation; 73.3% for steep foveation; Figure 6E). Moreover, the decrease in categorization accuracy for steeper levels of foveation was not systematically related to the size of the objects in both experiments (correlation between performance difference and object size: r = -0.07, p = 0.47 for animals, r = - 0.08, p = 0.39 for people).

We found similar but stronger effects in reaction times. Reaction times were the slowest for steeply foveated images (reaction times, mean ± std: 566 ± 182 ms, 547 ± 199 ms, 558 ± 374 ms and 577 ± 215 ms for full resolution images, and foveated images with spatial decay factors of 0.5, 1 and 4 respectively; p = 0.009 for ranksum test on reaction times for human foveated and steep foveated images; p = 0.02 for ranksum test on reaction times for full-resolution and human foveation; p > 0.05 for all other pairwise comparisons; Figure 6F).

To verify whether this effect is specific to animate objects, we performed an additional experiment in which subjects performed car detection. Here too, we observed similar results (Section S4).

To summarize, categorization performance in humans remained similar for both full resolution and foveated images, and worsened only for steeper levels of foveation. This is expected because humans already have peripheral blur in their eyes, as a result of which only steep foveation has any impact on performance.

Discussion

Our vision is sharpest at the center of gaze and blurs out into the periphery. The coarse sampling of the periphery is widely thought to save on wiring and metabolic cost without impacting performance. Here, we challenge this belief by showing that the human peripheral blur profile is actually optimal for object recognition on natural images. This in turn implies that the evolution of a fovea might have been driven by the demands of visual recognition rather than to simply satisfy wiring constraints.

Our specific findings in support of this conclusion are: (1) Deep networks trained on natural images show optimal performance for human-like foveation; (2) The relevant features for object recognition require high spatial frequencies near the image center and low spatial frequencies in the periphery; and (3) Humans performing categorization on natural scenes show a decline in categorization only when scenes are blurred beyond the normal level of peripheral blur. Below we discuss these findings in the context of the relevant literature.

Our main finding is that deep networks trained on foveated images achieve optimal performance for human-like peripheral blur (Figure 3). This raises several important concerns that merit careful consideration. First, we have observed only modest improvements in performance (Table 1) due to foveation, and it could be argued that this marginal improvement may not provide sufficient evolutionary drive. While it is unclear how much improvement “is enough”, we note that these improvements might be modest only because state-of-the-art networks are already saturated in their performance on standard image datasets, and that similar improvements have been reported in the literature (Felzenszwalb et al., 2010; Zhu et al., 2016). Second, could this improvement come from the foreground object becoming more salient with peripheral blurring? We consider it unlikely because this would predict a monotonic increase in accuracy with steeper blur profiles, which is opposite to what we observed. Third, if full-resolution images contain more information than foveated images, then why do deep networks achieve lower accuracy on full-resolution images? This could be because full-resolution images contain target-like features in the periphery that result in false alarms or slow detection (Katti et al., 2017). It could also be that deep networks trained on full-resolution images fail to pick up important scene context features (Zhu et al., 2016; Katti et al., 2019). Fourth, if foveation is optimal for recognition, then how does the visual system know where to foveate before initiating recognition? There is a large body of evidence showing that the primate oculomotor system uses a saliency map to guide saccades, and that low-level features can be used to guide eye movements towards potential objects of interest (Itti and Koch, 2001; Akbas and Eckstein, 2017). Whether and how the ventral stream visual regions influence the saliency map can be elucidated through paired recordings in both regions.

The finding that human-like peripheral blur yields optimal recognition in deep networks alone does not constitute proof that human peripheral blur evolved to optimize recognition. However, it is a remarkable coincidence that the exact human peripheral blur profile is what ends up being optimal for recognition. It could be argued that feature detectors in our brains are qualitatively different from deep networks, but there is growing evidence that this is not the case: object representations in deep networks have strong parallels to the ventral visual stream neural representations (Yamins et al., 2014; Ponce et al., 2019). Given this, any comparison of brains (both neural and behavioral data) and neural networks will benefit from training the computational models on appropriately foveated inputs.

Our conclusion that foveation might have evolved for optimal recognition stands in stark contrast to the widely held belief in the literature. Previous studies have used foveation as a pre-processing step to achieve image compression (Geisler and Perry, 1998) or to create saliency maps to guide eye movements (Itti and Koch, 2001). However no previous study has systematically varied peripheral blur profiles to examine the impact on recognition. A recent study has shown that foveation yields equivalent object detection performance to full-resolution images but with significant computational cost savings (Akbas and Eckstein, 2017). If foveation is so beneficial for object recognition, then why has this not been noticed previously? In our experiments, we observed consistently better performance for foveated images, but this benefit varied with the viewing distance used in the foveation calculations. We speculate that these studies may have used sub-optimal values of viewing distance, resulting in only marginal improvements.

We have shown that low-spatial frequency features are most informative for object detection in the image periphery, whereas high-spatial frequency features are most informative at the image center. Our results are in agreement with past reports that coarse representations play an important role in the categorisation of high eccentricity stimuli (Boucart et al., 2016) and that coarse global shape can be extracted rapidly and used for visual categorisation (Sripati and Olson, 2009; Leek et al., 2016). Interestingly, representation of peripheral vision has been reported to be stable even when central vision is impaired or lost (Boucart et al., 2010). Our results are also concordant with the recent observation that a fovea-like sampling lattice evolves after training a deep network for handwritten digit recognition (Cheung et al., 2017). These findings suggest that the evolution of a fovea can be driven by object detectors based on simple Gabor-like features as have been observed in the primary visual cortex.

The foveal bias towards high spatial frequency and peripheral bias towards coarse representations could play an important role in the temporal interactions that have been reported between respective retinotopic regions using transcranial stimulation (Chambers et al., 2013) and deficits in categorisation of peripherally presented stimuli due to temporally delayed presentation of noise at the fovea (Ramezani et al., 2019). We want to emphasize that our results relate to information processing when fixation has already been attained on a target, although coarse sampling of peripheral information has been shown to be a beneficial contextual signal that can guide exploratory eye movements as well (Torralba et al., 2006).

Our work is also consistent with accounts of spatial attention gating where task relevant information is selectively processed according to some spatial constraints (Leek et al., 2003; Reppa et al., 2012). However, since we are using the ImageNet dataset which is likely to have objects in the center of the image, we cannot test the interplay between object-based and location-based attentional mechanisms as previously studied in human behavior (Leek et al., 2003). Just the fact that foveation improves object recognition performance points towards limitations in off-the-shelf CNNs which can be potentially overcome using biological constraints (like we have shown here using foveation). Moreover, in future work, networks with these constrains can be probed for attentional gating mechanisms. More generally, we note that the organization of the fovea varies widely across animals (Land and Nilsson, 2012). We speculate that the fovea and peripheral blur profile in each species may be optimized for its high-level visual demands, just as our eyes are optimized to ours.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by Intermediate and Senior Fellowships from the DBT/Wellcome Trust India Alliance to SPA (Grant #: 500027/Z/09/Z and IA/S/17/1/503081) and by the DBT-IISc partnership programme. HK was supported by a DST-CSRI postdoctoral fellowship from the Department of Science and Technology (DST).

Data Availability

All codes and analyses required to replicate the results are available on OSF at https://osf.io/dcngp/

References

Akbas E, Eckstein MP. Object detection through search with a foveated visual system. Einhäuser W. PLOS Comput Biol. 2017; 13 e1005743 [PubMed: 28991906]

Arcaro MJ, Livingstone MS. A hierarchical, retinotopic proto-organization of the primate visual system at birth. Elife. 2017; 6

Bar M. Visual objects in context. Nat Rev Neurosci. 2004; 5: 617–629. [PubMed: 15263892]

Boucart M, Naili F, Despretz P, Defoort-Dhellemmes S, Fabre-Thorpe M. Implicit and explicit object recognition at very large visual eccentricities: No improvementafter loss of central vision. Vis cogn. 2010; 18: 839–858.

Boucart MV, Lenoble Q, Quettelart J, Szaffarczyk S, Despretz P, Thorpe SJ. Finding faces, animals, and vehicles in far peripheral vision. J Vis. 2016; 16: 1–13.

Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997; 10: 433–436. [PubMed: 9176952]

Campbell FW, Robson JG. Application of fourier analysis to the visibility of gratings. J Physiol. 1968; 197: 551–566. [PubMed: 5666169]

Chambers CD, Allen CPG, Maizey L, Williams MA. Is delayed foveal feedback critical for extra-foveal perception? Cortex. 2013; 49: 327–335. [PubMed: 22503283]

Cheung, B; Weiss, E; Olshausen, B. Emergence of foveal image sampling from learning to attend in visual scenes; 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings; 2017.

Curcio CA, Allen KA. Topography of ganglion cells in human retina. J Comp Neurol. 1990; 300: 5–25. [PubMed: 2229487]

Curcio CA, Sloan KR, Kalina RE, Hendrickson AE. Human photoreceptor topography. J Comp Neurol. 1990; 292: 497–523. [PubMed: 2324310]

Davenport JL, Potter MC. Scene consistency in object and background perception. Psychol Sci. 2004; 15: 559–564. [PubMed: 15271002]

Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object Detection with Discriminative Trained Part Based Models. IEEE Trans Pattern Anal Mach Intell. 2010; 32: 1627–1645. [PubMed: 20634557]

Geisler WS, Perry JS. Real-time foveated multiresolution system for low-bandwidth video communication. Proc SPIE. 1998; 3299: 294–305.

Gomez J, Barnett M, Grill-Spector K. Extensive childhood experience with Pokémon suggests eccentricity drives organization of visual cortex. Nat Hum Behav. 2019; 3: 611–624. [PubMed: 31061489]

He, K; Zhang, X; Ren, S; Sun, J. Deep residual learning for image recognition; Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit; 2016. 770–778.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. Science (80-). 2005; 310: 863–866.

Itti L, Koch C. Computational modelling of visual attention. Nat Rev Neurosci. 2001; 2: 194–203. [PubMed: 11256080]

Katti H, Peelen MV, Arun SP. How do targets, nontargets, and scene context influence real-world object detection? Atten Percept Psychophys. 2017; 79: 2021–2036. [PubMed: 28660468]

Katti H, Peelen MV, Arun SP. Machine vision benefits from human contextual expectations. Sci Rep. 2019; 9 2112 [PubMed: 30765753]

Land, MF, Nilsson, D-E. Animal eyes. 2nd Editio. New York, NY: Oxford University Press; 2012.

Larson AM, Loschky LC. The contributions of central versus peripheral vision to scene gist recognition. J Vis. 2009; 9: 6.1–16. [PubMed: 19761321]

Leek EC, Reppa I, Tipper SP. Inhibition of return for objects and locations in static displays. Percept Psychophys. 2003; 65: 388–395. [PubMed: 12785069]

Leek EC, Roberts M, Oliver ZJ, Cristino F, Pegna AJ. Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. Neuropsychologia. 2016; 89: 495–509. [PubMed: 27396674]

Li FF, VanRullen R, Koch C, Perona P. Rapid natural scene categorization in the near absence of attention. Proc Natl Acad Sci U S A. 2002; 99: 9596–9601. [PubMed: 12077298]

Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P. Microsoft COCO: Common Objects in Context. arXiv. 2014.

Morrison DJ, Schyns PG. Usage of spatial scales for the categorization of faces, objects, and scenes. Psychon Bull Rev. 2001; 8: 454–469. [PubMed: 11700896]

Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principlesand Neuronal Preferences. Cell. 2019; 177: 999–1009. e10 [PubMed: 31051108]

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. J Neurosci. 2018; 38: 7255–7269. [PubMed: 30006365]

Ramezani F, Kheradpisheh SR, Thorpe SJ, Ghodrati M. Object categorization in visual periphery is modulated by delayed foveal noise. J Vis. 2019; 19: 1–12.

Ratan Murty NA, Arun SP. Dynamics of 3D view invariance in monkey inferotemporal cortex. J Neurophysiol. 2015; 113: 2180–2194. [PubMed: 25609108]

Reppa I, Schmidt WC, Leek EC. Successes and failures in producing attentional object-based cueing effects. Attention, Perception, Psychophys. 2012; 74: 43–69.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015; 115: 211–252.

Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv. 2014. 1–14.

Sripati AP, Olson CR. Representing the forest before the trees: a global advantage effect in monkey inferotemporalcortex. J Neurosci. 2009; 29: 7788–7796. [PubMed: 19535590]

Stewart EEM, Valsecchi M, Schütz AC. A review of interactions between peripheral and foveal vision. J Vis. 2020; 20: 2.

Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature. 1996; 381: 520–522. [PubMed: 8632824]

Torralba A. Contextual priming for object detection. Int J Comput Vis. 2003; 53: 169–191.

Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev. 2006; 113: 766–786. [PubMed: 17014302]

Weber C, Triesch J. Implementations and Implications of Foveated Vision. Recent Patents Comput Sci. 2009; 2: 75–85.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A. 2014; 111: 8619–8624. [PubMed: 24812127]

Zhu Z, Xie L, Yuille AL. Object Recognition with and without Objects. 2016. arXiv1611.06596

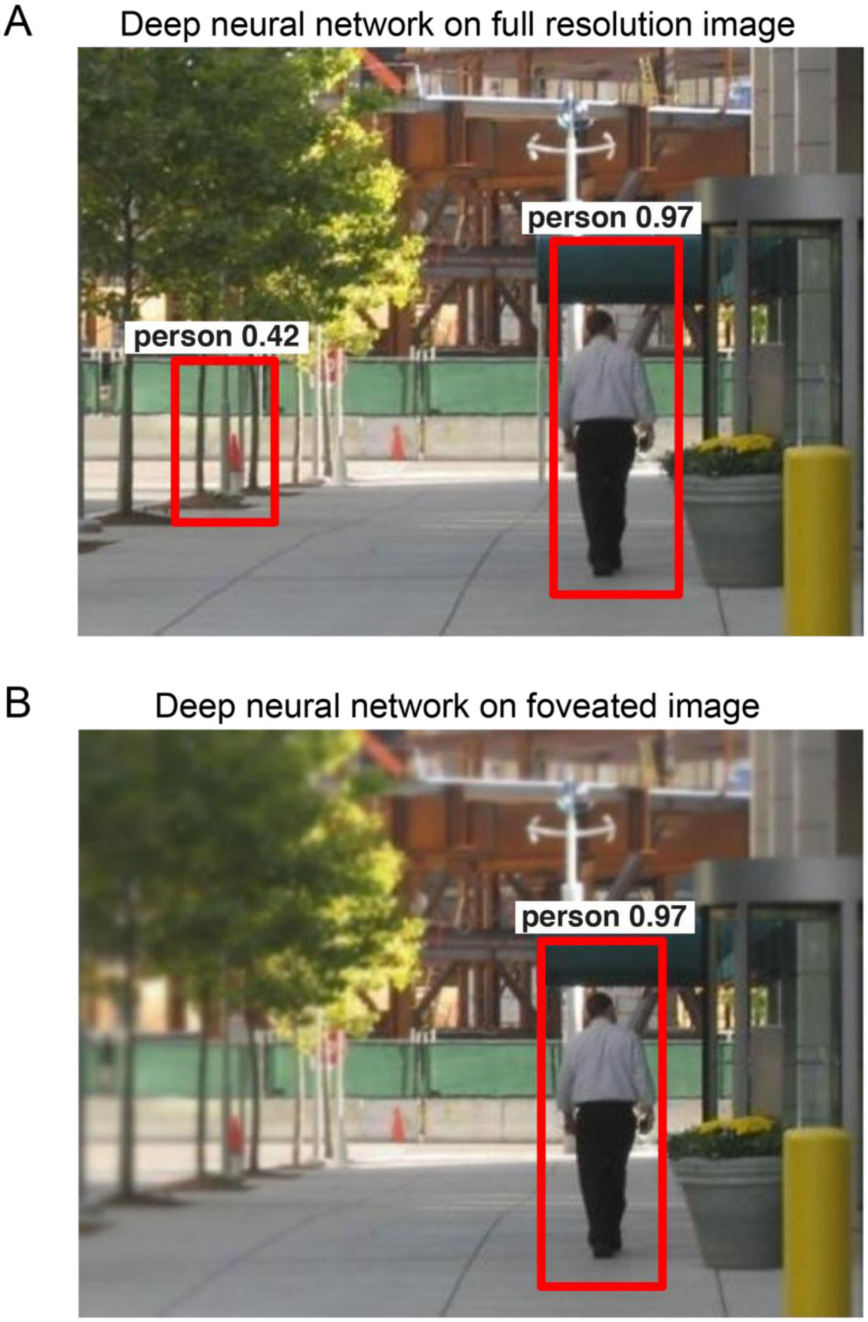
A
Deep neural network on full resolution image
person 0.97
person 0.42
B
Deep neural network on foveated image
person 0.97

Figure 1. Example object detection with and without peripheral blur.

(A) Example object detections from a state-of-the-art deep neural network (R-CNN), showing a correctly identified person and a false alarm in which a traffic cone is mistaken for a person.

(B) Example object detections on a foveated version of the image using the same network, showing the correctly identified person but without the false alarm.

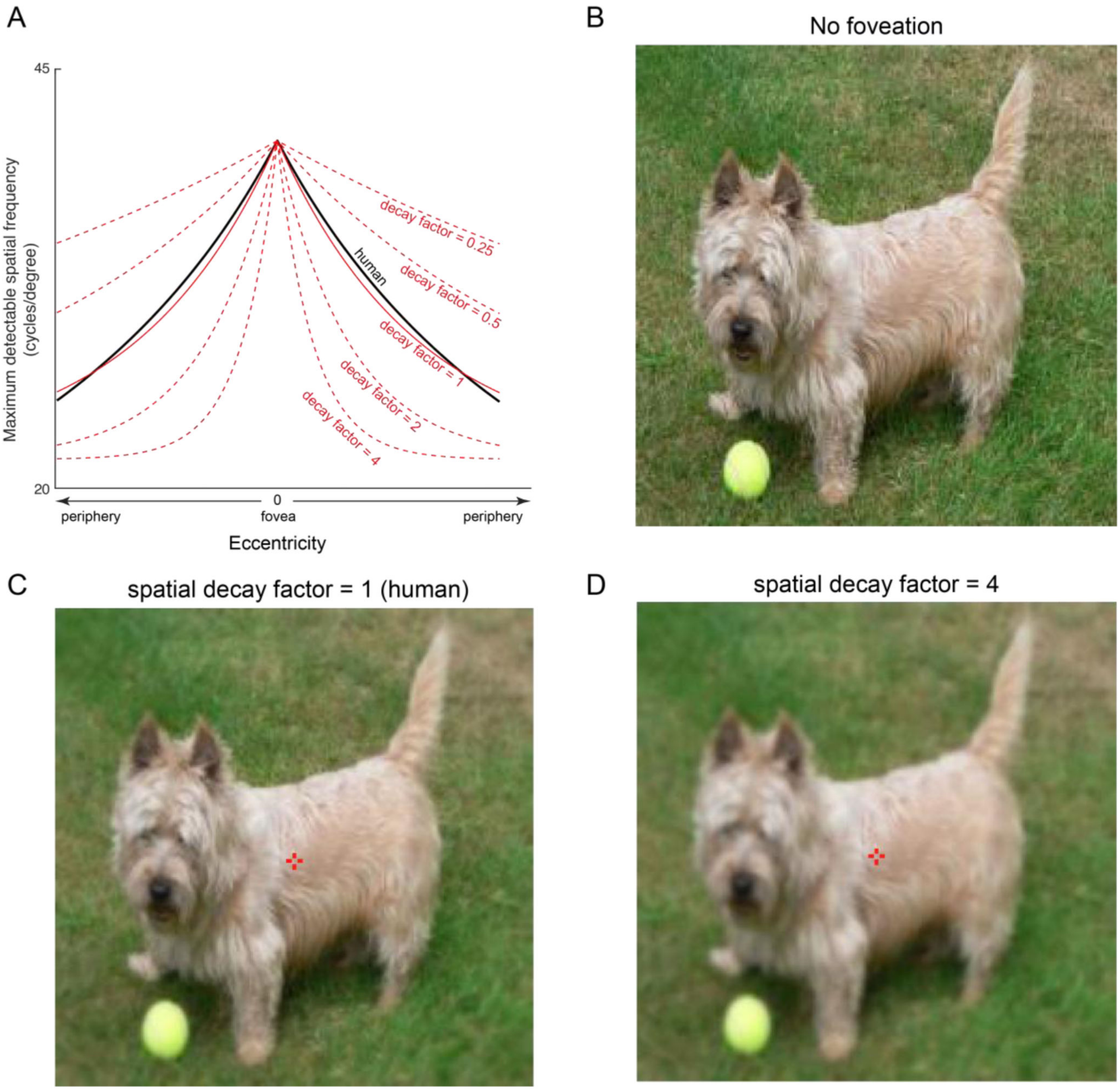

Figure 2. Example foveated images with varying peripheral blur profiles.

(A) Human contrast sensitivity function (*solid black line*) and the corresponding exponential fit (*solid red line*). The spatial decay of the exponential was varied by scaling the human exponential fit to obtain shallower or deeper blur profiles (*dashed red lines*).

(B) Example full resolution image

(C) Same as panel B but foveated on the object center (*red cross*) with a spatial decay of 1, which corresponds to the human peripheral blur function. At the appropriate viewing distance, fixating on the red cross will make this image look identical to the full resolution image in panel B.

(D) Same as panel B but foveated with a more extreme peripheral blur (spatial decay factor = 4).

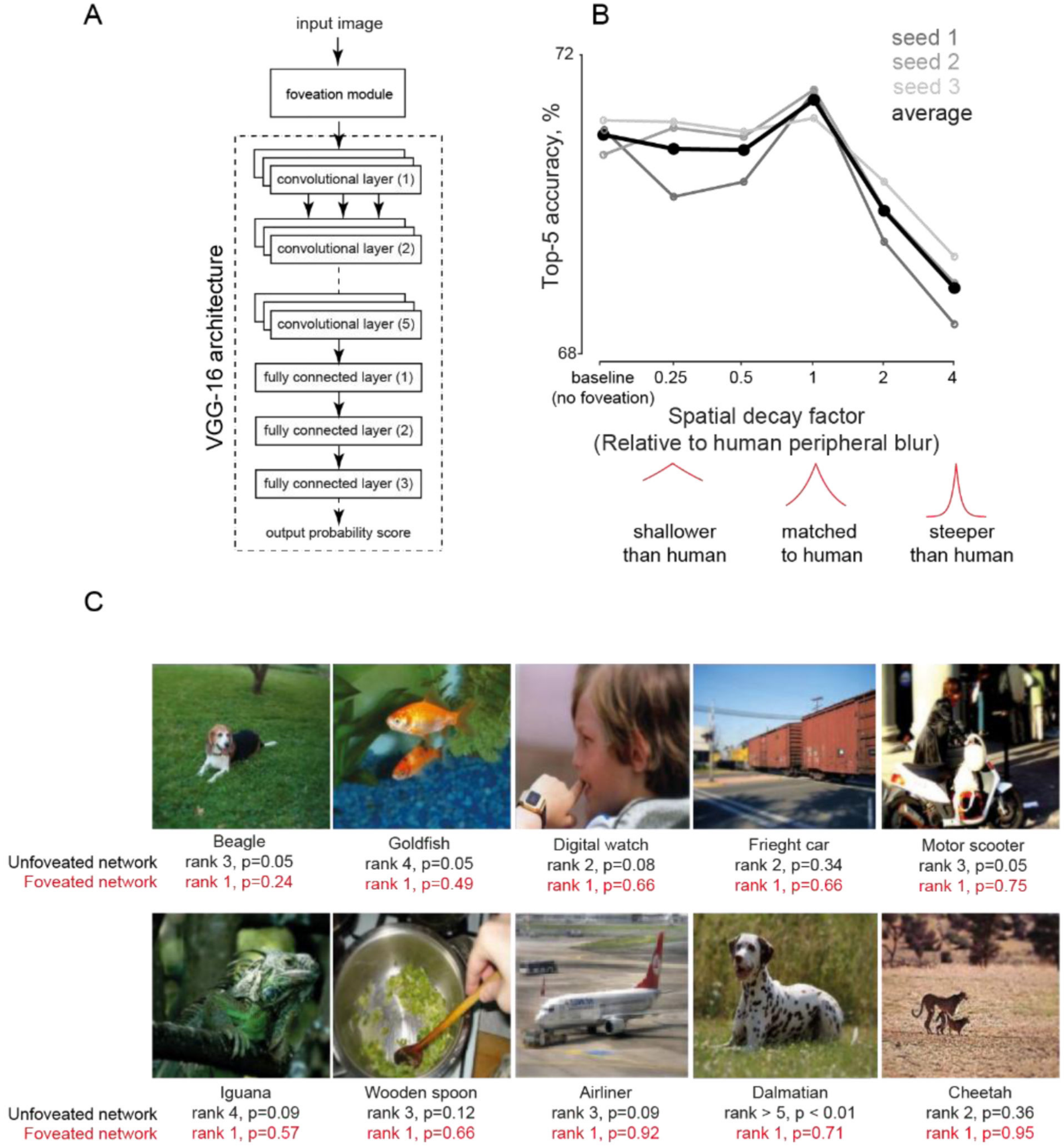

Figure 3. Human-like peripheral blur is optimal for object recognition.

(A) Schematic of the VGG-16 neural network architecture used to train images

(B) Top-5 accuracy of neural networks with varying peripheral blur. The accuracy of each network was calculated on test images after training it on foveated images with the corresponding blur profile. Each grey line corresponds to Top-5 test accuracies for individual instances of VGG-16 architecture trained from scratch with the random initialization, and the black line represents the average performance over the three instances.

(C) Example images for which the correct category was identified only by the foveated network (with human-like peripheral blur) but not the full-resolution (unfoveated) network. Below each image, the correct object label is shown (*top*), followed by its rank and posterior probability returned by the unfoveated network (*black, second row*) and by the foveated network (*red, third row*).

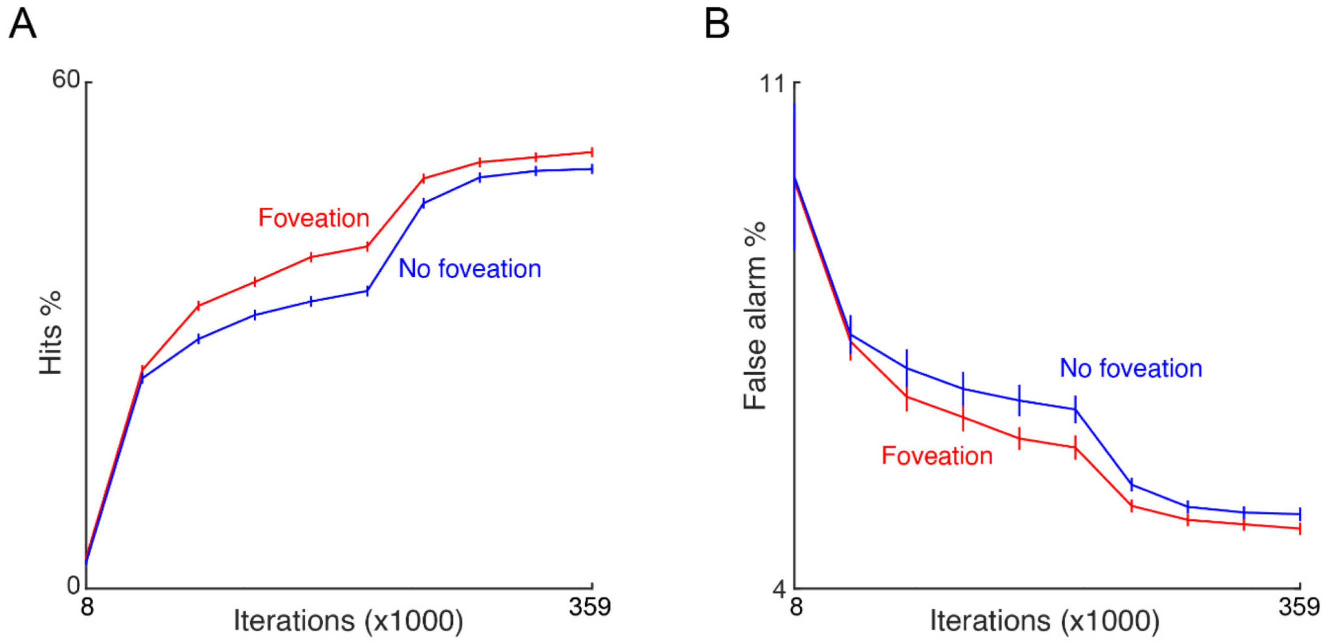

Figure 4. Object recognition performance over the course of training.

(A) Plot of percentage hits as a function of learning for networks trained on foveated images (*red*) and full resolution images (*blue*). (B) Same as in (A) but for false alarms. In both plots the x-axis indicates the number of iterations (or batches of data) in multiples of 1000. Error bars indicate s.e.m. across 1000 categories.

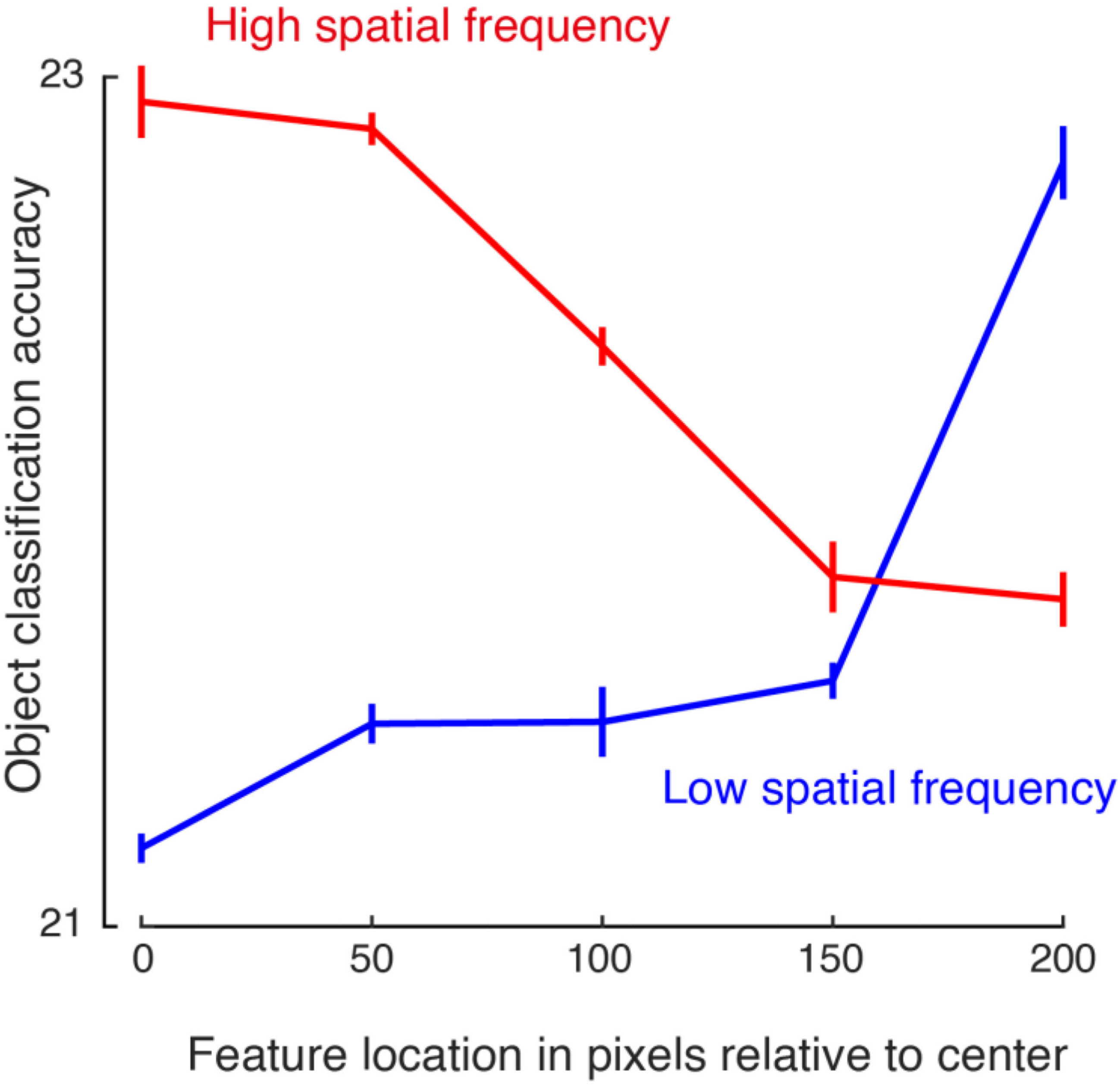

Figure 5. Relative importance of spatial frequency features as a function of image eccentricity.
Accuracy of a 11-way object decoder is plotted as a function of eccentricity i.e. feature location in pixels relative to the image center, for high spatial frequencies (*red*) and low spatial frequencies (*blue*).

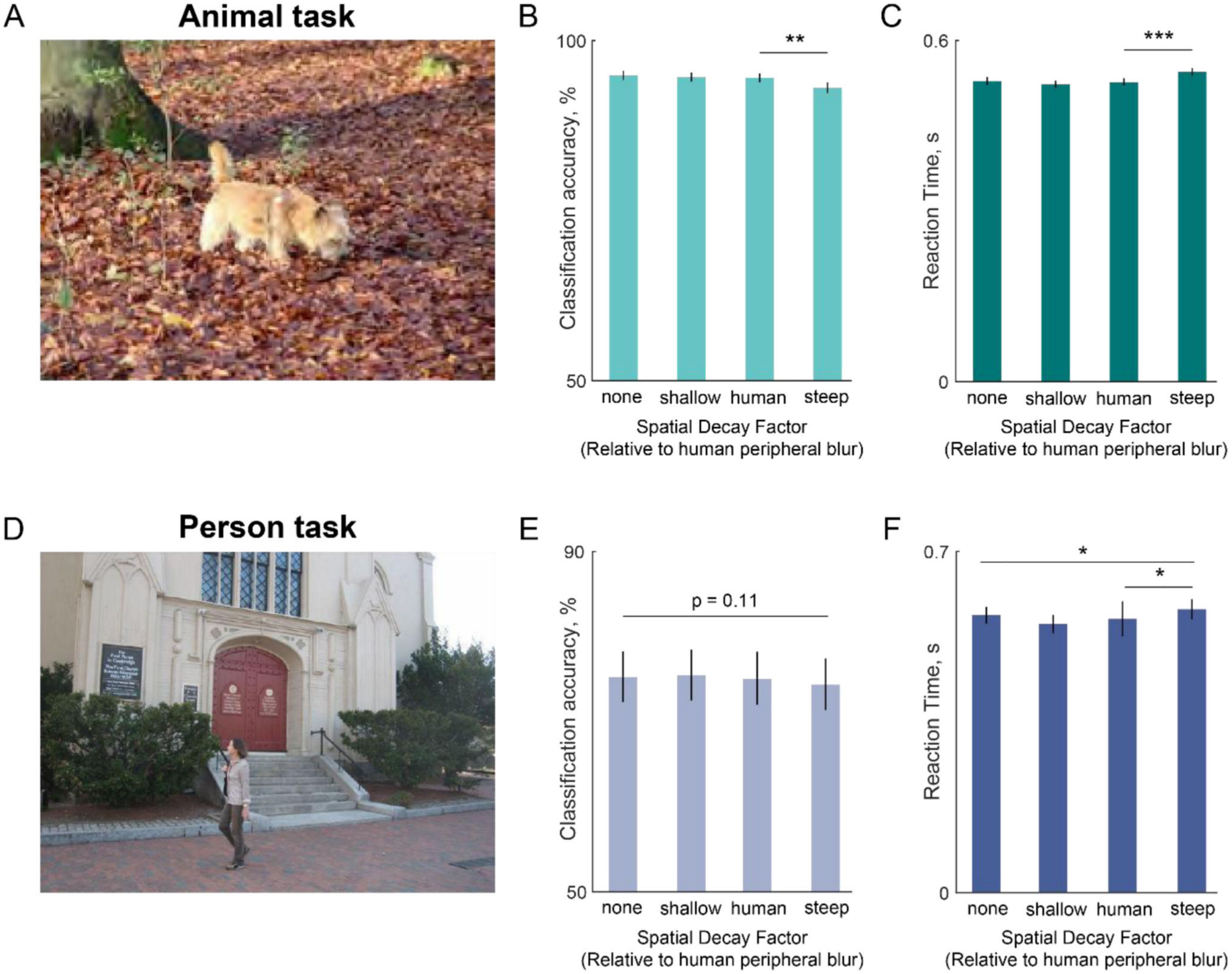

Figure 6. Human categorization declines only for steep foveation.

(A) Example full resolution image from the animal categorization task.

(B) Accuracy for different levels of foveation. Error bars indicate s.e.m. calculated across all images used in the task. Asterisks indicate statistical significance using a Wilcoxon signed rank-sum test across images (* is p < 0.05, ** is p < 0.005, *** is p < 0.0005). All other comparisons are not significant.

(C) Same as (B) but for reaction times on correct trials, with error bars indicating s.e.m across images. Conventions are as in (B).

(D) Example full resolution image from the person categorization task.

(E-F) Same as (B) and (C), but for the person categorization task.

Table 1
Classification performance of VGG-16 networks on foveated and full resolution images.

We report both Top-1 and Top-5 accuracies on test sets averaged across three networks, each trained from a specific random seed. The Top-1 accuracy refers to the accuracy with which the best guess of the network matched the object label. The Top-5 accuracy is the accuracy with which the correct label was present in the top 5 guesses of the network. The network trained on images with human-like peripheral blur (spatial decay factor = 1) is highlighted in *red*.

	Top-1 test accuracy	Top-5 test accuracy
4	44.8	68.9
2	46.1	69.9
1 (human)	**47.5**	**71.4**
0.5	46.9	70.7
0.25	46.9	70.7
0 (No foveation)	47.1	70.9