High-acuity vision from retinal image motion

Alexander G. Anderson
Physics Department and Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA, USA

Kavitha Ratnam
School of Optometry, University of California, Berkeley, Berkeley, CA, USA

Austin Roorda
School of Optometry, University of California, Berkeley, Berkeley, CA, USA

Bruno A. Olshausen
School of Optometry, Helen Wills Neuroscience Institute, and Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA, USA

A mathematical model and a possible neural mechanism are proposed to account for how fixational drift motion in the retina confers a benefit for the discrimination of high-acuity targets. We show that by simultaneously estimating object shape and eye motion, neurons in visual cortex can compute a higher quality representation of an object by averaging out non-uniformities in the retinal sampling lattice. The model proposes that this is accomplished by two separate populations of cortical neurons — one providing a representation of object shape and another representing eye position or motion — which are coupled through specific multiplicative connections. Combined with recent experimental findings, our model suggests that the visual system may utilize principles not unlike those used in computational imaging for achieving "super-resolution" via camera motion.

Introduction

During visual fixation, humans have a stable, high-acuity perception of the world despite substantial drifting movements of the eyes. Recent experiments demonstrate the benefit of these movements for the discrimination of a small letter whose stroke spacing is near the sampling limit of the cone photoreceptor array (Ratnam, Domdei, Harmening, & Roorda, 2017). Subjects are shown a diffraction-limited letter E in one of four orientations (strokes pointing up, down, left, or right) during natural drift movements of the eye, and are asked to report the letter's orientation. The stimulus size is chosen to challenge the subject to the point that the orientation is discriminated correctly 40% to 60% of the time. In a second condition, the image of the E is stabilized on the retina by a real-time eye tracker with cone-level precision. Here, subjects' performance decreases. In a third condition, the stimulus moves on the retina with the same statistics as natural eye motion, but incongruent (uncorrelated) with the eye's true motion. Surprisingly, although subjects are aware of the incongruent motion of the stimulus, their task performance is the same as the natural condition in which there is no perception of motion. Taken together, these results are remarkable because the visual features defining the object span just a few photoreceptors, yet the eye's own motion spreads these features over many photoreceptors within the presumed temporal integration window of downstream cortical neurons. Thus, there must be a neural mechanism that makes use of the movement of the stimulus relative to the retina, independent of whether or not the motion is generated by the eye, for improving task performance.

Our goal here is to elucidate the neural mechanisms that could underlie these experimental results with a mathematical model capable of exhibiting the same behavior. Previous modeling efforts aimed at modeling perceptual stability in the face of fixational eye movements proposed specific neural computations to build up invariant representations of sensory signals using shifter circuits (Anderson and Van Essen, 1987) or map-seeking circuits (Arathorn, Stevenson, Yang, Tiruveedhula, & Roorda, 2013). Other investigators have approached the problem in the framework of Bayesian inference and proposed models that decode retinal ganglion cell (RGC) spikes generated from a stimulus moving owing to fixational eye movements

Citation: Anderson, A. G., Ratnam, K., Roorda, A., & Olshausen, B. A. (2020). High-acuity vision from retinal image motion. *Journal of Vision*, *20*(7):34, 1–19, https://doi.org/10.1167/jov.20.7.34.

https://doi.org/10.1167/jov.20.7.34 Received August 18, 2019; published July 31, 2020 ISSN 1534-7362 Copyright 2020, The Authors

(Pitkow, Sompolinsky, & Meister, 2007; Burak, Rokni, Meister, & Sompolinsky, 2010). Burak et al. (2010) showed that, for stimuli with binary-valued pixels, a decoder of these spikes must take into account the motion of the eye (under reasonable assumptions about the size of the eye motions and the firing rates of RGCs), otherwise the reconstructed pattern is a blur. They showed how this blur may be mitigated by simultaneously estimating form and motion in a Bayesian optimal manner. The estimated motion is used to dynamically reroute the incoming spikes onto a population of cortical neurons so as to build up an unblurred estimate of the underlying pattern on the retina. Although this model took an important step in demonstrating a computational mechanism that can account for how high acuity is preserved under fixational eye movements, it primarily aims to *mitigate* this blur, viewing the eye position drift as a hindrance.

Inspired by the results of Burak et al., we sought to show how blur could not only be mitigated, but how retinal image drift could confer a benefit because it can potentially improve visual acuity by averaging over inhomogeneities in the retinal sampling lattice. Doing so requires generalizing the model to allow for spatially continuous eye movements and gray-valued image stimuli, as opposed to the discretized eye movements and binarized stimuli assumed in Burak et al. Generalizing the model in this way is both scientifically important and technically difficult. Although the structure of our generative model is closely related to that of Burak et al., the methods for inferring the spatial pattern from the spikes are completely different because the values for the position and pixel values can no longer be discretely enumerated. Furthermore, their mean-field approximation of the image does not allow for non-trivial priors on the spatial pattern, such as in the sparse coding model of V1 (Olshausen & Field, 1997). Thus, we developed a novel, approximate Bayesian inference method based on an online approximation of the expectation maximization (EM) algorithm.

The general idea that motion is beneficial for an image sensor has been considered in a variety of disciplines. In the computational imaging community, the problem of combining a sequence of low resolution images to form a single high-resolution image has well-developed solutions (e.g., Farsiu, Robinson, Elad, & Milanfar, 2004). In the field of active perception, Rucci and colleagues (Rucci, Iovin, Poletti, & Santini, 2007; Kuang, Poletti, Victor, & Rucci, 2012; Aytekin, Victor, & Rucci, 2014; Rucci & Victor, 2015; Boi, Poletti, Victor, & Rucci, 2017; Rucci, Ahissar, & Burr, 2018; Casile, Victor, & Rucci, 2019; Intoy & Rucci, 2020) have studied the benefits that could arise from small eye motions due to the spreading of signal power from the spatial domain into the temporal domain. They show that the $1/f^2$ spatial power spectrum of natural images, when combined with the statistics of eye motion, results in a flattening of the power spectrum over the joint spatiotemporal frequency domain. They further show that, when this signal is sent through the temporal filtering properties of RGCs, high spatial frequency details get amplified and that more global spatial structures such as contours could be detected from spike synchrony. Their theory is complementary to ours in that they address limitations imposed by postreceptoral mechanisms (e.g., limited dynamic range and bandwidth of the optic nerve) and subsequent processes of feature extraction, assuming the image signal has been adequately sampled by the cones such that it can be treated as a continuous function of space and time, $I(x, y, t)$. The focus of our work, by contrast, is to understand how spatial detail at the very highest spatial frequencies (50 cycles/deg) can be perceived and discriminated despite the fact that spatial information at these scales is compromised owing to the punctate nature of cone sampling, inhomogeneities in the retinal cone mosaic, and among the cones themselves. We also take into account the punctate encoding in time by RGCs–that is, signals are conveyed to the brain not as continuous waveforms, but as a sequence of spikes. We propose a computational mechanism for decoding images that have been sampled and temporally encoded in this way, and we quantitatively evaluate its performance, corroborating the psychophysical measurements of Ratnam et al.

In what follows, we first describe our model used for estimating form and motion, with more complete details described in the Appendix. We then use our model to decode simulated spikes generated by the same letter E stimulus used in the experiments of Ratnam et al. We show that it is possible to resolve the fine spatial structure of the letter E that would otherwise be impossible to resolve in a statically viewed presentation of the stimulus on the cone lattice. We also demonstrate the ability to resolve the stimulus given a retina with holes in the cone lattice, which corroborates the fact that observers with retinal degeneration exhibit normal visual acuity. Finally, we generalize the model to the case of natural image stimuli, using a sparse latent variable model as the image prior, resulting in a model that is consistent with the known feature representations in V1 (i.e., neurons with localized, oriented, and bandpass receptive fields). We conclude by discussing neurobiological and technological implications of the model.

Methods

The simulations in this article proceed by first generating spikes from a spatial array of simplified

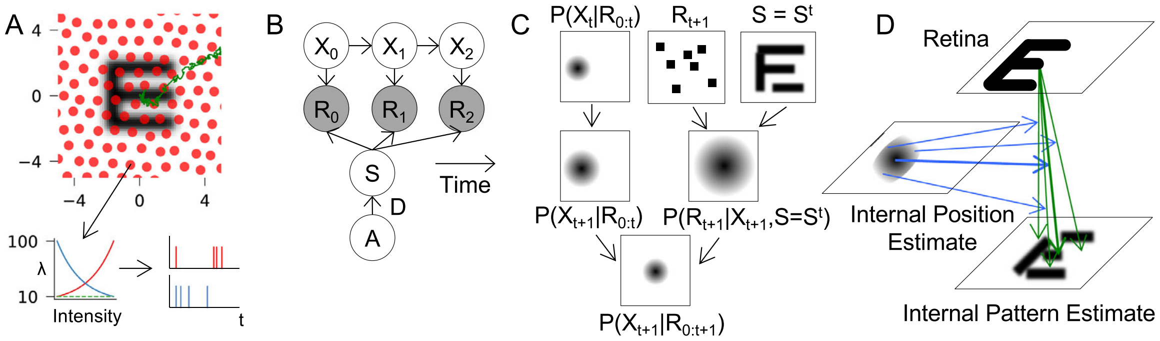

Figure 1. Model Overview: (**A**) An upright letter E (stroke width = 0.8 arcmin) projected onto a simulated cone lattice (average spacing 1.09 arcmin) with a 500 ms eye drift trajectory (Ratnam et al., 2017) superimposed (green trace). RGC cell spikes are generated using a linear-nonlinear-poisson model with ON and OFF cells. The ON and OFF RGC response functions are symmetrical, so the presence of a stimulus for an ON cell gives an equivalent response to the absence of a stimulus for an OFF cell. (**B**) Probabilistic model for inferring stimulus shape S (encoded by latent variables A) and position X from retinal spikes R. Arrows indicate causal relationships between variables. The spikes R are observed and the latent factors encoding shape A and position X must be simultaneously inferred. (**C, D**) The spike decoder repeatedly alternates between two steps: (**C**) In the first step (Equations 5), the estimate of the pattern is fixed ($S = S^t$) and new evidence coming from the next set of incoming spikes R_{t+1} is incorporated to obtain an updated posterior distribution over eye position $P(X_{t+1}|R_{0:t+1})$ (shown as a probability cloud). This update is computed by multiplying the probability distribution over the predicted position $P(X_{t+1}|R_{0:t})$ (computed from the diffusion model applied to the previous position estimate) together with the likelihood $P(R_{t+1}|X_{t+1}, S = S^t)$ (computed by cross-correlating the current estimate of the pattern with the spatial array of incoming spikes). (**D**) In the second step (Equations 8, 10), the neurons representing the internal position estimate X_t act to dynamically route incoming spikes by multiplicatively gating their connections to the internal pattern estimate, thus updating S.

RGCs in response to a spatial pattern (either an E or a natural scene patch) as it drifts over the retina, as shown in Figure 1A. These spikes are then decoded by our proposed model — an approximate inference procedure that assumes knowledge of the process by which the spikes were generated — to infer the spatial pattern and its motion, as shown in Figure 1B to 1D.

Simulating RGC responses to drifting stimuli

Each RGC is assumed to receive input from a single cone (one ON and one OFF RGC per foveal cone (Ahmad, Klug, Herr, Sterling, & Schein, 2003), and is modeled as having a Gaussian receptive field with full width at half maximum of 0.48 times the cone spacing (Macleod, Williams, & Makous, 1992). For the present purposes, we leave out the lateral inhibition and temporal filtering properties of RGCs, focusing mainly on the spatial resolution provided by the retina. The retinal cone lattice is specified by generating a hexagonal grid (random orientation) with spacing of 1.09 arcmin and then randomly jittering the position of each cone by adding noise uniformly distributed within ±25% of the spacing to the horizontal and vertical coordinates. Although jittering the centers of the cones adds more realism to the simulations and demonstrates the flexibility of the inference model, our experiments showed that it does not impact the reconstruction error as a function of time.

Eye movement trajectories are generated either as a diffusive random walk or from drift eye movement recordings from (Ratnam et al., 2017). The eye motion traces are obtained using an adaptive optics scanning laser ophthalmoscope (Roorda et al., 2002). Trials with microsaccades are thrown out. The raw data are cleaned by using interpolation to replace one timestep outliers and trials with longer sections of invalid data are thrown out. Finally, a Kalman filter with a diffusion motion prior is used to smooth the data. Because the error between the smoothed path and the true path has roughly double the standard deviation of the adaptive optics scanning laser ophthalmoscope's error (Stevenson, Roorda, & Kumar, 2010), one-half of the difference between the data and the smoothed path is added to the smoothed path to retain some of the non-smooth component of the eye motion (aka tremor).

Spiking responses of RGCs are generated using a linear-nonlinear-poisson model (Paninski, Simoncelli, & Pillow, 2004) without any spike history dependencies. The instantaneous rate parameter for each RGC is set to a baseline of 10 Hz and increases exponentially according to the inner product of the RGC's receptive field with the retinal image translated by the current eye position, and scaled so that the maximum rate is 100 Hz, as specified in Appendix Equations 11–17.

Joint inference of object shape and eye position from RGC spike trains

Our hypothesis is that the visual cortex seeks to infer the spatial stimulus pattern, S, given the incoming spikes, R, where the trajectory of the eye, X, is an unknown variable.[1] If both X and R were known, S could be easily estimated by accumulating evidence from spikes after the motion is used to correct for the translation of the eye. Likewise, if both S and R were known, X could be estimated by finding the translation of the stimulus pattern, S, that provides the best spatial alignment with the spike patterns, R, across time. In the case where only R is known, X and S must be *jointly inferred*, because one variable is needed to estimate the other.

To solve this problem from a principled perspective, we impose priors on S and X. The prior on the eye trajectory, $p(X)$, is a diffusive random walk with a diffusion constant D_C^{infer}. The prior on the stimulus pattern, $p(S)$, is constructed by constraining S to be given by $S = D\,A$, where D is a "dictionary matrix" whose columns are some elementary spatial patterns, and the vector A is a set of latent variables that specify how much of each pattern is present. The spatial structure in S can then be modeled with a simple, factorial prior over A, $p(A) = \prod_i p(A_i)$. The relationships between R, X, S, and A are described by the probabilistic graphical model shown in Figure 1B. The joint distribution of the nodes in the graph, N, is $p(N) = \prod_i p(N_i | N_{\pi(i)})$, where $\pi(i)$ denotes the parents of node i in the graph defining the model (parent-child relationships are denoted by arrows in the diagram). All quantities of interest are computed by marginalizing the joint distribution.

In an ideal Bayesian framework, one would compute the full posterior distribution over the latent variables encoding object shape $p(A|R)$, given by

$$p(A|R) \propto \sum_X p(R|X, S = DA)\, p(X)\, p(A), \qquad (1)$$

where $p(R|X, S)$ reflects the probabilistic (Poisson) model used in generating the spikes (Appendix Equation 11). The posterior $p(A|R)$ assigns a probability for every possible stimulus pattern $S = DA$ given the spikes R coming from the retina, taking into account all possible eye movement trajectories weighted by their probability. We use a series of approximations to derive a computationally tractable, causal, and online computation to estimate A (see the Appendix for details). First, only the most probable set of latent shape variables is considered, $\hat{A} = \text{argmax}_A p(A|R)$. The second is to deal with the intractable sum over all possible eye trajectories by using an online approximation of the EM algorithm. The EM algorithm maximizes $\log P(A|R)$ in an iterative manner by alternating between two steps, one for estimating X, which comes from introducing a variational distribution $q(X)$, and the other for estimating A. To make time explicit in X and R, we henceforth rewrite them as $X_{0:\,T} = (X_0, X_1, \ldots X_T)$ and $R_{0:\,T} = (R_0, R_1, \ldots R_T)$, where T is the total number of time steps in the simulation. R_t denotes the number of spikes emitted from each RGC in the time interval $[t, t + \Delta t]$. Because R_t depends only on the current eye position, X_t, and the stimulus, S, we can derive a set of EM update equations as follows:

$$q_t(X_t) \leftarrow p(X_t | R_{0:T}, S = DA') \qquad (2)$$

$$A' \leftarrow \text{argmax}_A \left[\sum_t \sum_{X_t} q_t(X_t) \log p(R_t | X_t, S = DA) + \log p(A) \right] \qquad (3)$$

A full derivation is given in Appendix Equations 31-34. Equation 2 estimates the eye position at time t, X_t, given the spikes $R_{0:\,T}$ and the current estimate of the spatial pattern A', while Equation 3 estimates A given the spikes $R_{0:\,T}$ and estimated eye positions $X_{0:\,T}$. The traditional EM algorithm repeatedly applies these equations for some number of iterations. For simplicity, A can be initialized to zero. Note that although these update equations are guaranteed to converge to a critical point of $\log P(A|R)$ by repeatedly applying them (and initializing them with $A = 0$), they are still non-causal (requiring spikes from the future to estimate quantities at the current time t), and Equation 3 is not amenable to online processing because it requires optimizing over a batch of quantities from $t = 0$: T.

To obtain a causal position estimator for Equation 2, the distribution over eye position at time t is approximated by replacing it with a filtering estimate that only takes into account spikes up to time t:

$$\begin{aligned} q_t(X_t) &\leftarrow p(X_t | R_{0:t}, S = DA) \\ &\approx \; p(X_t | R_{0:T}, S = DA) \qquad (4) \end{aligned}$$

This is then updated at each subsequent timestep via

$$\begin{aligned} q_{t+1}(X_{t+1}) \sim \; & p(R_{t+1} | X_{t+1}, \\ & S^t = D\hat{A}^t) \sum_{X_t} p(X_{t+1} | X_t) q_t(X_t) \qquad (5) \end{aligned}$$

where $\hat{A}^t$ is the current estimate of A given the spikes from 0 to t (computed via Equation 8). The steps involved in this calculation are shown graphically in Figure 1C. A particle filter with resampling (Doucet & Johansen, 2009) is used to represent and propagate $q_t(X_t)$ from one timestep to the next (see Equations 48, 49).

The optimization for A in Equation 3 is also modified to be causal and online. First, we denote the negative expected log-likelihood of A at time t as

$$E_r^t(A) \equiv -\sum_{X_t} q_t(X_t) \log p(R_t | X_t, S = DA), \qquad (6)$$

which can be thought of as an energy (to be minimized) that corresponds with how well A agrees with the position estimate and spikes at time t. A causal approximation to the update for A at time t may be obtained by considering the sum of these energies only up to time t, along with the log-prior:

$$A^t \leftarrow \mathrm{argmin}_A \left[\sum_{t'=0}^{t} E_r^{t'}(A) - \log p(A) \right], \qquad (7)$$

where we are now minimizing rather than maximizing owing to the change in sign. To make the computation online (so that the entire sum over time need not be reminimized at each time step), the sum of the energy terms up to time t is replaced by a quadratic approximation, resulting in the following update for the next time step:

$$\begin{aligned} \hat{A}^{t+1} = \mathrm{argmin}_A \Bigg[& \frac{1}{2}(A - \hat{A}^t)^T \hat{H}_t (A - \hat{A}^t) \\ & + E_r^{t+1}(A) + (A - \hat{A}^t)^T \frac{\partial \log p(A)}{\partial A}\Big|_{A=\hat{A}^t} \\ & - \log p(A) \Bigg], \end{aligned} \qquad (8)$$

where $\hat{A}^t$ is the current estimate of A given the spikes from 0 to t, and

$$\hat{H}^t = \hat{H}^{t-1} + \frac{\partial^2}{\partial A^2} E_r^t(A)|_{A=\hat{A}^t}. \qquad (9)$$

The contribution of each of the terms in this expression may be understood as follows: the first term is a running estimate of the accumulated energies $E_r^t(A)$ up to time t, the second term corresponds with the energy coming from the new set of incoming spikes at time $t + 1$, and the last two terms correspond with the log prior on A. The quadratic approximation of $E_r^t(A)$ corresponds with a Gaussian approximation in probability, and so as H grows over time, the uncertainty shrinks, meaning that this term has increasing influence in determining the optimal value of A over time. $\log p(A)$ is either the sum of absolute values of A (to encourage sparsity) or a quadratic function of A.

The minimization of Equation 8 is done using the FISTA algorithm (Beck & Teboulle, 2009), which is a version of gradient descent modified to handle the situation where the expression to be minimized contains an L_1 loss term. The basic computations required to compute the gradient are specified in Equations 50–56 of the Appendix. Figure 1D shows a graphical illustration of the computation owing to the gradient of the second term, $E_r^{t+1}(A)$, which updates A according to each new set of incoming spikes. This results in a “dynamic routing” circuit (Olshausen, Anderson, & Van Essen, 1993), in which RGC spikes R are routed into different elements of the internal shape estimate A via another set of units representing the internal position estimate X that multiplicatively gate the RGC’s.

To summarize, the full algorithm computes three equations at each timestep. First, an internal estimate of eye position at time t is updated based on the current estimate of the stimulus pattern $S^t = D\,A^t$ and the incoming spikes R_t (Equation 5). Second, the new estimate of the stimulus pattern (represented by latent factors A) is generated by minimizing Equation 8, which takes into account the new spikes and the updated estimate of eye position. Third, the estimate of the uncertainty of the latent factors, H, is updated (Equation 9).

Results

A moving retina averages out spatial inhomogeneities

Much like looking through a broken window, viewing the world through a stationary, inhomogeneous retina results in a belief about the world that is precise in some places and uncertain in others. The key idea of this work is that this detrimental, nonuniform uncertainty can be alleviated by the eye’s natural drift movements. Our main result, shown in Figure 2, is that the signal generated by a moving retina, when properly processed by downstream neural circuitry that jointly estimates the eye’s motion and the stimulus, results in a higher quality representation of the stimulus as compared the signal generated by a stationary retina. Specifically, for a stimulus duration of 700 ms, our model achieves a 50% improvement in the average signal-to-noise ratio (SNR) when the retina drifts (average SNR=5.9) as compared with when it is held stationary (average SNR=3.9). SNR is computed as the power of the ground-truth signal divided by the squared error between the ground-truth pattern and the estimated pattern (see Appendix, section SNR for details).

The parameters of the stimulus, cone sampling lattice, and eye motion trajectories used in these simulations correspond directly with the experiments of (Ratnam et al., 2017). The strokes of the E have a width of 0.8 arcmin, and the cone lattice has an average spacing of 1.09 arcmin. The diffusion constant

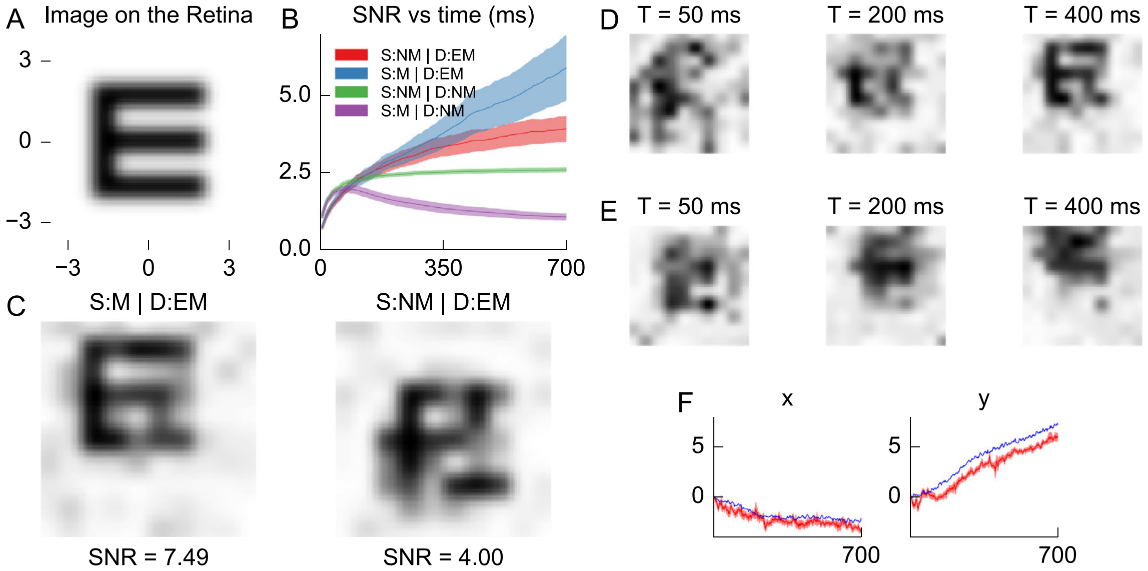

Figure 2. Benefits of motion for the discrimination of high-acuity targets: (**A**) Stimulus (S) to be recovered. The entire pattern is defined on a 20 × 20 pixel array subtending 8 arcmin. The width of each leg of the E is 2 pixels (0.8 arcmin). The cone lattice and eye trajectories are the same as in Figure 1A. (**B**) SNR of the reconstruction of the E as a function of time. The Shaded region shows 95% confidence intervals of the mean given 40 trials. Either the stimulus is moved relative to the retina (S:M = Motion), or not (S:NM = No Motion). For each of these cases, the stimulus pattern is inferred using either the approximate EM algorithm (D:EM) or an optimal decoder assuming no motion (D:NM) are used to decode the pattern. Note that D:EM > D:NM, even when there is no stimulus motion (S:NM) because the uncertainty over the position implicitly smooths the pattern. The difference between the two best methods is statistically significant (S:M | D:EM > S:NM | D:EM with $p = 0.002$ at $t = 700$ ms). (**C**) Typical reconstructions of the pattern in the case of either motion and no motion after 700 ms. (**D**) Reconstruction over time in the case of motion using the EM algorithm. (**E**) Reconstruction over time in the case of motion assuming no motion. (**F**) Estimated versus true eye position as a function of time. The red curve shows the estimated horizontal and vertical eye position using the EM algorithm (width reflects +/−1 standard deviation). The blue curve shows the true eye position. The timestep of the simulation is 1 ms.

D_C^{infer} that is used in the prior for inferring motion, $P(X)$, is set to $20\,\mathrm{arcmin}^2/s$, which matches that of the recorded eye motions. Although the subject in the experiment is asked to report which of four orientations the E is in, our task requires estimating the entire shape. The prior used to infer the shape, $P(S)$, uses a simple dictionary of non-overlapping square blocks of size 0.8 arcmin × 0.8 arcmin, with no sparsity imposed.

Because the receptive fields of the cones modeled as Gaussians have a full-width half maximum that is *half* the distance between the cones (Macleod et al., 1992), the strokes of the E can fall between the cones. In other words, even a retina with uniformly tiled cones has spatially non uniform sensitivity to the diffraction-limited stimuli in the experiments of (Ratnam et al., 2017). It is remarkable that both the mathematical model and human subjects can recover the stimulus given the gaps and irregularities in sensitivity in the retinal cone lattice (Harmening, Tuten, Roorda, & Sincich, 2014).

In additional experiments, we examined how performance changes as a function of stimulus size (Figure 5, SI). When the stimulus is very small, there is no benefit from eye motion. It cannot be well-decoded in either condition (static or moving) because the features are too small relative to the cone receptive field size. When the stimulus size is sufficiently large so that the stimulus features are large relative to gaps between the cones, both conditions accurately estimate the stimulus. Even though the SNR is higher with eye motion, there is effectively no perceptually noticeable gain because both are near perfect. There is only a nontrivial motion benefit when the strokes of the E are on the order of the spacing between the cones. Varying the magnitude of the eye motion (gain) shows that the maximum benefit from eye motion is obtained for gains between 0.5 and 1.0. The performance drops off significantly for zero motion or motion gains around 1.5 and above.

Beyond the punctate sensitivity of the cones, there are other sources of inhomogeneities in the retina

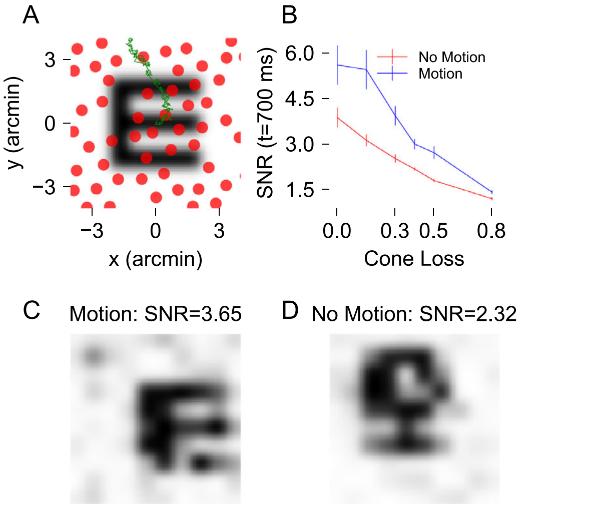

Figure 3. Motion benefit during cone loss. (**A**) Letter E stimulus sampled by a retinal cone lattice that has 30% of the cones dropped out randomly (cone loss, eye trajectories, and RGC spikes are resampled each trial). The same stimulus size, cone spacing, eye trajectories, and diffusion constant for inference were used as in Figure 2. (**B**) SNR at $t = 700$ ms as a function of cone loss for a moving and a stationary retina with $n = 21$ for each motion condition and cone loss value. The error bars correspond with plus or minus one standard error of the mean. (**C** and **D**) Examples of the reconstructed stimulus in the case of retinal drift motion and no motion for 30% cone loss.

that can compromise the accurate recovery of the luminance pattern of the retinal image, including variable cone gain factors (Li et al., 2014), different spectral sensitivities (Hofer, Carroll, Neitz, Neitz, & Williams, 2005), and disruptions in the cone mosaic caused by retinal degeneration (Duncan et al., 2007). Even in extreme cases, where retinal degeneration results in a fovea with 52% fewer cones than normal, patients still have normal visual acuity (Ratnam, Carroll, Porco, Duncan, & Roorda, 2013). Our model illustrates how these limitations can be compensated for by eye movements. Figure 3 shows the results of a simulation where a variable percentage of the cones are dropped out (besides the cone lattice, all other parameters are the same as the experiments in Figure 2). The quality of stimulus reconstruction enabled by a moving retina is dramatically improved over that with a stationary retina under conditions of cone loss.

Compounding the challenge of inferring spatial patterns defined by luminance, the visual system must additionally infer the spatial distribution of the color of objects (Sabesan, Schmidt, Tuten, & Roorda, 2016). The randomly placed cones tend to form clumps and the three cones types vary widely in their proportions (Hofer et al., 2005), which begs the question of how the joint spatiochromatic structure of small objects can be correctly inferred. Drift motion may also play a role here by averaging color appearance as an object is swept over different spectral swaths of the retina, and this merits further investigation.

Inferring natural image patterns

To infer more complex spatial patterns such as would occur in natural scenes, it is desirable to use a richer prior $p(S)$ to capture this structure. For this we turn to the sparse coding model of V1 (Olshausen & Field, 1997), which uses the generative model $S = DA$, where D is a dictionary of features learned from the statistics of natural images and A is a set of latent variables with a sparse prior $p(A)$. The goal in this case is to infer the latent factors (or image features), A, rather than a pictorial description of the pattern, S, from the incoming spikes, R. The equations for inferring A given S are usually interpreted as describing the dynamics of a neural network where the elements of A correspond with the activations of cortical neurons that have “Gabor-like” receptive fields similar to neurons in V1 (given by the dictionary, D) (Rozell, Johnson, Baraniuk, & Olshausen, 2008). In this case, we infer A given only the spikes R, which change as patterns drift over the retina. The resulting Equations (5, 8) can be interpreted as describing the interactions between two separate populations of neurons that work together to jointly infer the eye position X and the latent factors A. The neurons representing the latent factors A will appear to have dynamic, Gabor-like receptive fields that track features as they drift across the retina rather than remaining locked in retinotopic coordinates. Our experiments simulating this model on whitened natural scene patches (whitening the stimulus serves as an approximation to the center surround receptive field structure of RGCs) demonstrate that the sparse prior improves the inference of spatial patterns drawn from natural images (Figure 4).

Discussion

The drift motions that occur during fixation create a problem, but also an opportunity, for neural circuits downstream tasked with inferring the structure of high acuity targets. The prior work of Burak et al. showed how the problem may be solved by a Bayesian decoder that factorizes the time-varying spikes arriving from the retina into separate representations of form and motion. Our contribution here is to take this work a step further to realize the opportunity provided by retinal drift to obtain a higher quality visual representation than would otherwise be available given the inhomogeneities of the retinal sampling lattice.

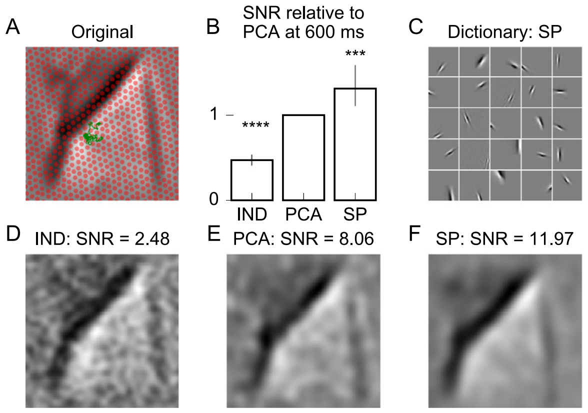

Figure 4. Neurons with structured receptive fields improve inference. (**A**) A whitened 32 × 32 pixel natural scene patch scaled to subtend a square with side length 24 arcmin is projected onto a simulated cone lattice with an average spacing of 1 arcmin. The retinal drift motion in this case is generated by a random walk with $D_c = 20\,\mathrm{arcmin}^2/s$. (**B**) SNR of the decoded image at $t = 600$ ms. RGC spikes are decoded using three pattern priors. The SNR is plotted relative to PCA averaged over 15 trials (different natural scene patches and eye trajectories). Error bars show 95% confidence intervals. The p-values are calculated between the uniform prior and PCA, and between the sparse coding prior and PCA (**** $p < 0.0001$; *** $p < 0.001$). (**C**) A random set of 25 elements from the learned sparse coding dictionary, D. Sparse coding seeks to describe any given image pattern as a sparse linear combination of these features. (**D** – **F**) Example reconstructed image patterns for each method after 600 ms. IND, independent pixel prior; PCA, Gaussian prior; SP, dictionary trained with sparse coding with both a L1 and L2 prior.

The model proposed here should be seen as a first step to establish the basic neural computations that would need to occur for a causal, online system to perform approximate Bayesian inference that could account for the improvement in acuity observed in the experiments of Ratnam et al. There are obviously many important neurobiological elements missing from our model — the temporal filtering known to occur in RGCs, Magno versus Parvo streams, wavelength selectivity of cones and color-opponency of RGCs, and so on. For this first step, we sought to include the most important biophysical factors that make the recovery of fine spatial detail a challenge — that is, cone sampling properties, and the spiking nature of neural activity, which requires temporal integration by neurons downstream. Further work is needed to realize a more accurate neurobiological implementation of the model to demonstrate its true feasibility, and thus in the meantime any conclusions from our results should be tempered accordingly.

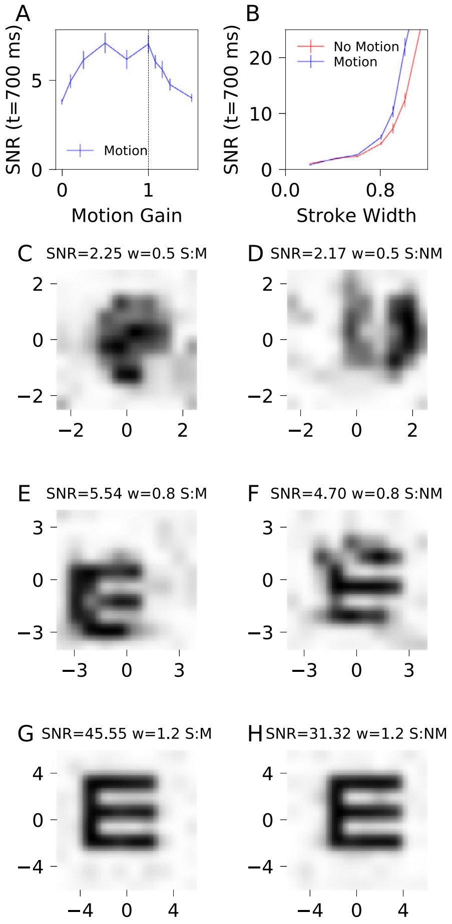

Figure 5. Extended Tuning Plots: (**A**) The SNR as a function of motion gain ($n = 40$ for each value of the motion gain). The experimentally measured eye trajectories are used, except that the overall position is multiplied by the gain factor. (**B**) The SNR as a function of stimulus size ($n = 20$). Both plots use the same parameters as in Figure 2. The error bars in both plots show the standard error. (**E–H**) Example reconstructions for the stimulus size experiments with stroke width (w), and motion (S:M) or no motion (S:NM). The horizontal and vertical axes are in arcmin. (**C, D**) For small stimuli, the orientation of the stimulus is unrecognizable in both cases. (**E, F**) For stimuli with a stroke width on the order of the spacing of the cones, the orientation of the stimulus is barely recognizable. (**G, H**) For larger stimuli, although the SNRs are different, the orientation of the stimulus is unambiguous, despite a large difference in the SNR.

In this section, we discuss some of the considerations that arise in mapping different elements of our inference model onto neural circuits in the brain, as well as further modeling and experimental efforts suggested by this work.

Neural implementation

The update Equations (5 and 8) can be interpreted as describing the interactions between two separate populations of neurons — one representing hypotheses about eye position, X, and another representing the stimulus pattern, A. We hypothesize these two populations to reside in area V1. The incoming spikes R would be carried by the LGN afferents innervating layer 4 of V1 (assuming LGN to be a simple relay of RGCs). The neurons representing A would likely be those in layer 4, or possibly layers 2 and 3. The hypotheses about eye position X would be represented by a population of neurons corresponding to the particles supporting $q(X)$. Such a scheme for neurally representing and updating probability distributions was proposed previously by (Lee & Mumford, 2003).

Importantly, the neural representations of A and X are not computed independently from the input, but rather jointly by multiplicative interactions between the two populations. The neurons representing X essentially compute a cross-correlation between the spatial pattern of incoming spikes R and the current estimate of the pattern represented by A (Figure 1C). Conversely, the neurons representing A are computed (in part) by dynamically routing the incoming spikes R via multiplicative gating by neurons representing X (Figure 1D).

The idea of dynamic routing (i.e., shifter circuits) was proposed more than 30 years ago as a model for stabilizing the cortical image representation in the face of retinal drift (Anderson & Van Essen, 1987). Here, rather than proposing a routing circuit a priori, the routing dynamics emerge from the principled objective of doing optimal (Bayesian) estimation of a moving spatial pattern using a log-linear Poisson observation model. To see this, consider the gradient of the second term in the cost function of Equation 8, $E_r^{t+1}(A)$, which is minimized using Fast Iterative Shrinkage-Thresholding Algorithm (FISTA). Ignoring the slight modifications from using FISTA instead of gradient descent, the update equation for the k^{th} element of A is:

$$\frac{dA_k}{dt} \propto \sum_j R_{t,j} \langle g_{j,k}(x) \rangle_t - \langle \lambda_{t,j} g_{j,k}(x) \rangle_t \Delta t, \quad (10)$$

where $R_{t,j}$ is the number of spikes arriving from RGC j in the time interval $[t, t + \Delta t]$ and $\lambda_{t,j}$ is the corresponding rate parameter of its Poisson distribution $p(R_t|X_t, S = DA)$ (a full derivation is the Appendix Equations 50–56). $g_{j,k}(x)$ corresponds to a dynamically controlled connection strength between RGC j and latent factor k that is determined by the eye position estimate X. $\langle \cdot \rangle_t$ denotes averaging with respect to $q_t(X_t)$. The first term of Equation 10 corresponds with a multiplicative gating of the incoming spikes by the internal position estimate. The second term is a homeostatic correction that corresponds with the expected number of incoming spikes given the internal estimate of the spatial pattern. The precise mathematical form of $g_{j,k}(x)$ is determined by the parameters of the spiking model and the receptive fields of the latent factors (Appendix Equation 56).

An interesting future direction will be to reformulate the model to directly estimate motion rather than position, and to use this to update the pattern estimate. The model currently assumes RGCs with receptive fields that are static in time and the inference model effectively updates its position estimate via spatial cross-correlation between the current pattern estimate and the image features. Alternatively, a shift signal relative to the current position could be estimated via spatiotemporal correlation and then used to dynamically route spikes into the latent representation of shape. Reformulating the model in this way could allow for a more direct correspondence with the temporal filtering and direction selective properties of RGCs and V1 neurons, respectively.

Neurobiological implications and questions

A key question that arises from this model is whether neurons in the foveal region of V1 form a locally stabilized, object-centered representation or a dynamically changing representation that moves in the presence of fixational drift. This question has been investigated previously with conflicting results from different laboratories (Motter & Poggio, 1990; Motter, 1995; Gur & Snodderly, 1997). In the case of microsaccades (Meirovithz, Ayzenshtat, Werner-Reiss, Shamir, & Slovin, 2011), observe that a local population response evoked by a small stimulus is shifted over the V1 retinotopic map after each microsaccade. Recent experimental work on mapping the receptive fields of V1 neurons while compensating for eye motion is a promising approach to resolve this question (McFarland, Bondy, Cumming, & Butts, 2014). Another promising direction is to use an adaptive optics scanning laser ophthalmoscope with targeted stimulus delivery combined with V1 electrophysiological measurements (or two photon imaging) to study V1 activity in response to motion-controlled stimuli presented to the fovea (Sincich, Zhang, Tiruveedhula, Horton, & Roorda, 2009). It should be noted that, although our model recovers an explicit stabilized representation of the object, it is also possible that these computations could be done in a nonstabilized representation that still integrates information efficiently (Appendix: Alternative Representations). Another possibility that is important to consider is that it may be the case that the visual cortex has instances of both types of cells.

In addition to neurons sensitive to the shape of the stimulus, we also predict that there is a collection of neurons that track the position of the eye to high precision in visual cortex. Although the computations to integrate information in our simulations are handled by a particle filter, the same computations could be executed by an integrator circuit that tracks the position of the left and right eyes. In line with this, Snodderly, Kagan, & Gur (2001) find that some V1 cells have varying activation in response to drift and microsaccades (e.g., tuned to one or the other, or a combination).

More generally, there is good reason to believe that the neural computations associated with the fovea are fundamentally different than the periphery. Fixational drift is large relative to the receptive field sizes of RGCs in the fovea (but not in the periphery) and there is an additional factor of cortical amplification in the fovea. There are four times more LGN cells per RGC and 10 times more striate cells per LGN projection (Connolly & Van Essen, 1984) in the fovea versus the periphery. How exactly this over-representation is being used for high-acuity vision merits further attention.

Future directions

Beyond understanding the neural computations associated with the fovea, there are many important ways to extend the model and the associated experiments. First, more work needs to be done to understand the way in which spike history dependencies in RGC firing contribute to the perception of high-acuity stimuli. On one hand, the spatial pattern is moving so fast that there may be an effect akin to motion blur, where there is less spatial information content available in the RGC spikes owing to the duration of the temporal integration of light on a particular RGC. On the other hand, the temporal filtering may serve as a preprocessing step that whitens the stimulus and decreases the impact of RGC noise on the final estimate of the stimulus. Regardless, new methods of approximate Bayesian inference need to be developed to extend the inference model to the case where the RGC cells have spike history dependencies. Second, there are many unanswered questions about our ability to infer the spatial color profile of small objects. For instance, how is our ability to infer color impacted by the nonuniform placement of the different cone types? Furthermore, to what extent does the natural motion of the retina help to alleviate these nonuniformities? Finally, although our psychophysical experiments and mathematical model probe the inner workings of our retinal circuitry, more work toward understanding simultaneous estimation of form and motion given high-acuity stimuli presented without adaptive optics is warranted.

Conclusions

The role of eye movements in visual perception is an important and long-studied problem. We use psychophysical experiments and mathematical modeling to identify a novel principle by which one can understand the benefits of drift eye movements for the perception of high-acuity targets: eye movements carry the stimulus across the retina to acquire a higher acuity representation of the spatial structure in the world than would otherwise be possible owing to inhomogeneities in retinal sampling. This principle has far-reaching consequences, both for understanding biological sensory systems and the design of novel sensors. From the biological side, this principle informs future experiments on the high-acuity perception of color and active perception for vision and other sensory modalities. From the technological side, the novel algorithms of this work motivate the design of imaging systems that exploit (rather than avoid) image motion in order to infer high-quality images from cheap non-uniform or noisy sensors.

Keywords: spatial vision, visual acuity, eye movements, Bayesian inference

Acknowledgments

The authors thank the members of the Redwood Center for Theoretical Neuroscience and the Roorda lab for useful feedback. This material is based on work supported by the National Science Foundation under Grant No. DGE 1106400 (A.G.A.) and IIS-1111765 (A.G.A., B.A.O.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The experimental work was supported by NIH grant EY023591 (A.R.), Foundation Fighting Blindness (A.R.), and the Irvin M. Borish/Essilor Ezell Fellowship (K.R.).

Commercial relationships: none.
Corresponding author: Alexander G. Anderson.
Email: aga@berkeley.edu.
Address: University of California 575A Evans Hall, MC# 3198, Berkeley, CA 94720, USA.

Footnote

[1] Although the brain can, in principle, obtain motion estimates on the scale of visual drift from proprioceptive or efference copy signals, a number of lines of evidence suggest that this is not the case

(Guthrie, et al., 1983; Donaldson, 2000; Murakami & Cavanagh, 2001). Furthermore, the incongruent motion experiments of (Ratnam et al., 2017) demonstrate that such an efference copy is not necessary for a high-acuity vision task.

References

Ahmad, K. M., Klug, K., Herr, S., Sterling, P., & Schein, S. (2003). Cell density ratios in a foveal patch in macaque retina. *Visual Neuroscience, 20*(02), 189–209.

Anderson, C. H., & Van Essen, D. C. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences of the United States of America, 84*(17), 6297–6301.

Arathorn, D. W., Stevenson, S. B., Yang, Q., Tiruveedhula, P., & Roorda, A. (2013). How the unstable eye sees a stable and moving world. *Journal of Vision, 13*(10), 22.

Aytekin, M., Victor, J. D., & Rucci, M. (2014). The visual input to the retina during natural head-free fixation. *Journal of Neuroscience, 34*(38), 12701–12715.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences, 2*(1), 183–202.

Boi, M., Poletti, M., Victor, J. D., & Rucci, M. (2017). Consequences of the oculomotor cycle for the dynamics of perception. *Current Biology, 27*(9), 1268–1277.

Burak, Y., Rokni, U., Meister, M., & Sompolinsky, H. (2010). Bayesian model of dynamic image stabilization in the visual system. *Proceedings of the National Academy of Sciences of the United States of America, 107*(45), 19525–19530.

Casile, A., Victor, J. D., & Rucci, M. (2019). Contrast sensitivity reveals an oculomotor strategy for temporally encoding space. *eLife, 8*, e40924.

Connolly, M., & Van Essen, D. (1984). The representation of the visual field in parvicellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey. *Journal of Comparative Neurology, 226*(4), 544–564.

Donaldson, I. M. L. (2000). The functions of the proprioceptors of the eye muscles. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 355*(1404), 1685–1754.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering, 12*, 656–704.

Duncan, J. L., Zhang, Y., Gandhi, J., Nakanishi, C., Othman, M., & Branham, K. E. H. (2007). High-resolution imaging with adaptive optics in patients with inherited retinal degeneration. *Investigative Ophthalmology and Visual Science, 48*(7), 3283–3291.

Farsiu, S., Robinson, M. D., Elad, M., & Milanfar, P. (2004). Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing, 13*(10), 1327–1344.

Gur, M., & Snodderly, D. M. (1997). Visual receptive fields of neurons in primary visual cortex (v1) move in space with the eye movements of fixation. *Vision Research, 37*(3), 257–265.

Guthrie, B. L., Porter, J. D., & Sparks, D. L. (1983). Corollary discharge provides accurate eye position information to the oculomotor system. *Science, 221*(4616), 1193–1195.

Harmening, W. M., Tuten, W. S., Roorda, A., & Sincich, L. C. (2014). Mapping the perceptual grain of the human retina. *Journal of Neuroscience, 34*(16), 5667–5677.

Hateren, J. H. v., & Schaaf, A. v. d. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences, 265*(1394), 359–366.

Hofer, H., Carroll, J., Neitz, J., Neitz, M., & Williams, D. R. (2005). Organization of the human trichromatic cone mosaic. *Journal of Neuroscience, 25*(42), 9669–9679.

Intoy, J., & Rucci, M. (2020). Finely tuned eye movements enhance visual acuity. *Nature Communications, 11*(1), 1–11.

Kuang, X., Poletti, M., Victor, J. D., & Rucci, M. (2012). Temporal encoding of spatial information during active visual fixation. *Current Biology, 22*(6), 510–514.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A, 20*(7), 1434–1448.

Li, P. H., Field, G. D., Greschner, M., Ahn, D., Gunning, D. E., & Mathieson, K. et al. (2014). Retinal representation of the elementary visual signal. *Neuron, 81*(1), 130–139.

Macleod, D. I. A., Williams, D. R., & Makous, W. (1992). A visual nonlinearity fed by single cones. *Vision Research, 32*(2), 347–363.

McFarland, J. M., Bondy, A. G., Cumming, B. G., & Butts, D. A. (2014). High-resolution eye tracking using v1 neuron activity. *Nature Communications, 5*(1), 4605, https://doi.org/10.1038/ncomms5605.

Meirovithz, E., Ayzenshtat, I., Werner-Reiss, U., Shamir, I., & Slovin, H. (2011). Spatiotemporal effects of microsaccades on population activity in the visual cortex of monkeys during fixation. *Cerebral Cortex, 22*(2), 294–307.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine, 13*(6), 47–60.

Motter, B. C. (1995). Receptive-field border stabilization during visual fixation. *Investigative Ophthalmology and Visual Science, 36*(4), S691–S691.

Motter, B. C., & Poggio, G. F. (1990). Dynamic stabilization of receptive fields of cortical neurons (vi) during fixation of gaze in the macaque. *Experimental Brain Research, 83*(1), 37–43.

Murakami, I., & Cavanagh, P. (2001). Visual jitter: Evidence for visual-motion-based compensation of retinal slip due to small eye movements. *Vision Research, 41*(2), 173–186.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13*(11), 4700–4719.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research, 37*(23), 3311–3325.

Paninski, L., Simoncelli, E. P., & Pillow, J. W. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural model. *Advances in Neural Information Processing Systems,* 1311–1318.

Pitkow, X., Sompolinsky, H., & Meister, M. (2007). A neural computation for visual acuity in the presence of eye movements. *PLoS Biology, 5*(12), e331.

Ratnam, K., Carroll, J., Porco, T. C., Duncan, J. L., & Roorda, A. (2013). Relationship between foveal cone structure and clinical measures of visual function in patients with inherited retinal degenerations. *Investigative Ophthalmology & Visual Science, 54*(8), 5836–5847.

Ratnam, K., Domdei, N., Harmening, W. M., & Roorda, A. (2017). Benefits of retinal image motion at the limits of spatial vision. *Journal of Vision, 17*(1), 30–30.

Roorda, A., Romero-Borja, F., Donnelly, W. J., III, Queener, H., Hebert, T. J., & Campbell, M. C. (2002). Adaptive optics scanning laser ophthalmoscopy. *Optics Express, 10*(9), 405–412.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Computation, 20*(10), 2526–2563.

Rucci, M., Ahissar, E., & Burr, D. (2018). Temporal coding of visual space. *Trends in Cognitive Sciences, 22*(10), 883–895.

Rucci, M., Iovin, R., Poletti, M., & Santini, F. (2007). Miniature eye movements enhance fine spatial detail. *Nature, 447*(7146), 852–855.

Rucci, M., & Victor, J. D. (2015). The unsteady eye: An information-processing stage, not a bug. *Trends in Neurosciences, 38*(4), 195–206.

Sabesan, R., Schmidt, B. P., Tuten, W. S., & Roorda, A. (2016). The elementary representation of spatial and color vision in the human retina. *Science Advances, 2*(9), e1600797.

Sincich, L. C., Zhang, Y., Tiruveedhula, P., Horton, J. C., & Roorda, A. (2009). Resolving single cone inputs to visual receptive fields. *Nature Neuroscience, 12*(8), 967.

Snodderly, D. M., Kagan, I., & Gur, M. (2001). Selective activation of visual cortex neurons by fixational eye movements: Implications for neural coding. *Visual Neuroscience, 18*(2), 259–277.

Stevenson, S. B., Roorda, A., & Kumar, G. (2010). Eye tracking with the adaptive optics scanning laser ophthalmoscope. In: *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 195–198).

Troy, J. B., & Lee, B. B. (1994). Steady discharges of macaque retinal ganglion cells. *Visual Neuroscience, 11*(01), 111–118.

Appendix

The mathematical model is developed in full detail. The appendix begins with the model notation. Next, the RGC model, the priors for inference and the steps of the algorithm are discussed. Next, we give a full derivation of the inference algorithm and the computations involved in the optimization. Finally, we include a Supplementary Movie that shows reconstruction emerge as the information from the spikes are integrated over time.

Model notation

Subscripts: t: time step, i: pixel index, j: RGC index, k: latent factor, p: particle number.

(1) S is the spatial pattern to be inferred, in a pixel representation. S_i denotes a particular pixel of the pattern. S_i is constrained to be between s_{min} and s_{max} (for natural scenes, $(-0.5, 0.5)$, $(0, 1)$ for the tumbling E). X_i^S denotes the center of pixel i. The pixels are placed in a grid with spacing ds. The simulated projected image of the pattern, $I(x)$, is smoothed using Gaussian interpolation, with $\sigma^S =$

$0.5{*}ds$. Thus, $I(x) = \sum_i S_i N(x; \mu = X_i^S, \sigma = \sigma^S)$, where N denotes a Gaussian.

(2) A is the vector of latent factors that generate S through a dictionary, D (e.g., $S = DA$, where S, A are column vectors and D is a matrix). A_k denotes the kth latent factor.

(3) D is the dictionary, where D_k is the kth dictionary element.

(4) X_t^R (abbreviated as X_t) denotes the position of the retina relative the cone lattice. X is used as an abbreviation of X_t for all t.

(5) $R_{t,j}$ denotes the number of spikes of RGC j in the time window $[t, t + \Delta t]$. Δt is the timestep of the simulation (taken to be 1 ms). R is used as an abbreviation for $R_{t,j}$ for all t and j.

(6) A jittered cone lattice with spacing de is constructed. Each cone is connected to one ON cell and one OFF cell. The jth RGC has a Gaussian receptive field $N(x; \mu = X_j^E, \sigma = \sigma_j^E)$ with a full width half maximum of 0.48 times the cone spacing (Macleod et al., 1992) (thus, $\sigma^E = 0.203 \cdot de$, where de is the spacing of the cones).

(7) D_C is the diffusion coefficient of the eye movements, $\lambda_0 = 10$ Hz, $\lambda_1 = 100$ Hz are the baseline and maximum firing rates of the neurons (Troy & Lee, 1994).

RGC model

The spiking of the RGC's are modeled using an linear-nonlinear-poisson model with no spike history dependencies.

$$\log p(R_{t,j}|S, X_t) = R_{t,j} \log[\lambda_j(S, X_t)dt] - \lambda_j(S, X_t)dt \qquad (11)$$

$$\lambda_j(S, X_t) = \exp\left(\log \lambda_0 + \log(\lambda_1/\lambda_0) \cdot c''_{j,t}\right) \qquad (12)$$

$$c''_{j,t} = c'_{j,t} \text{ if } j \in \text{ON} \text{ or } 1 - c'_{j,t} \text{ if } j \in \text{OFF} \qquad (13)$$

$$c'_{j,t} = (c_{j,t} - s_{min})/(s_{max} - s_{min}) \qquad (14)$$

$$c_{j,t} = g \cdot \sum_i S_i T(X_t^R)_{i,j} \qquad (15)$$

$$T(X^R)_{i,j} = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{||X_i^S - X_j^E - X^R||^2}{2\sigma^2}\right] \qquad (16)$$

$$\sigma^2 = (\sigma^S)^2 + (\sigma^E)^2 \quad (17)$$

g is a gain factor that sets the maximum size of $c_{j,t}$ to be 1 when S is the vector of 1s. The scaling ensures that $c''_{j,t} \in [0, 1]$ when $S_i \in [s_{min}, s_{max}]$. The inner product of the pattern projected onto the retina with the Gaussian receptive field for each cone for an arbitrary displacement of the retina is calculated as follows: let T denote the translation operator and E_j denote the receptive field of the jth neuron. Then $\int d\vec{x}[S(\vec{x})T_{X^R}E_j(\vec{x})] = \sum_i S_i \int d\vec{x}[N(\vec{x}; \vec{X}_i^S, \sigma_S^2)N(\vec{x}; \vec{X}_j^E + \vec{X}^R, \sigma_E^2)] = \sum_i S_i \cdot N(\vec{X}_i^S - \vec{X}_j^E - \vec{X}^R; 0, \sigma_S^2 + \sigma_E^2)$. Thus, the inner product can be written as $\Sigma_i S_i T(x^R)_{i,j}$

Spike generation model

In order To generate a spike train for our decoder, a spatial pattern and an eye motion trajectory are fed into the spiking model. The eye path is either generated by a diffusive random walk with a diffusion constant $D_C = D_C^{gen}$ or a drift eye motion trajectories (from Ratnam et al., 2017). The stimulus is either an E or a natural scene patch.

Motion and pattern priors

There are many possible pairs of eye motion paths and spatial patterns that are consistent with the incoming retinal spikes. To deal with this ambiguity, we impose priors on the eye trajectory and the pattern. The spiking model $p(R|X, S)$ (defined above) and the priors define the relationships between the random variables in the probabilistic graphical model shown in Figure 1.

Motion prior

The fixational eye movements are modeled as a diffusion process with a diffusion constant $D_C \equiv D_C^{infer}$. Note that the optimal D_C^{infer} may not be equal to D_C^{gen}, so this parameter is optimized using cross-validation.

$$p(X_0) = \delta(X_0) \quad (18)$$

$$-\log p(X_t|X_{t-1}) = \frac{1}{2(D_C/2)\Delta t}(X_t - X_{t-1})^2 + C \quad (19)$$

Note that X_t is a 2D vector, so for the overall vector to have a diffusion constant of $D_C \Delta t$, then each individual component has a diffusion constant of $D_C/2\Delta t$. Higher-order priors that seek to both infer the eye position and the velocity of the eye were investigated, but did not provide significant benefits for inference. However, such a direction could help to reformulate the model in terms of tracking the velocity of the eye instead of the absolute position.

Pattern prior

Priors of the form

$$-\log p(S|A) = \delta(S - DA) \quad (20)$$

are considered where $\delta(x)$ is a delta function. As a result of this pattern prior, the latent generative factors of the pattern are estimated instead of directly estimating pixels.

The independent pixel prior as in Burak et al. (Burak et al., 2010) may be seen as a special case where D is the identity matrix and there is a uniform prior on A.

In the natural scenes experiments, the prior on A is chosen to be a combination of a $L1$ and a $L2$ prior. The $L1$ part of the prior is $-\log p(A) = \beta \Sigma_k |A_k|$. The $L2$ part of the prior is of the form $0.5 \cdot (A - \mu)^T \Sigma^{-1} (A - \mu)$, where μ and Σ are precomputed mean and covariances of the latent factors. This prior is implemented by setting $\hat{A}_0 = \mu$ and $\hat{H}_0 = \Sigma^{-1}$ in the initialization of the algorithm.

An additional term is added to the cost function to force the pixels to be in the range $[S_{min}, S_{max}]$: $-\log p(S_i) = \gamma * (\Theta(S_i - S_{max}) + \Theta(-(S_i - S_{min})))$, where $\Theta(x)$ is x if $x > 0$ and zero otherwise and γ is a parameter (chosen to be 10).

Sparse coding prior

Natural scenes

Sparse coding was used to train the dictionary. A set of 32×32 image patches were chosen randomly from a set of natural scenes images from the Van Hateren natural images database (Hateren and Schaaf, 1998). The images are whitened by convolving with a whitening filter based on natural scene statistics: $f * e^{-(f/0.35N)^2}$, where N is the length in pixels of the image patches. Sparse coding minimizes the objective function $(S - DA)^2 + k|A|$, where S is the pattern in a pixel representation, D is a dictionary, and A is a set of sparse generative factors. k is a constant that trades off reconstruction error and sparsity (Olshausen & Field, 1997). The value of k is chosen so that the reconstruction error and the sparsity penalty are the same order of magnitude. This step ensures that the minimization procedure attempts to seek a trade off between sparsity and reconstruction error. In jointly optimizing for A and D, for fixed D, FISTA is used to find the best A. A gradient step of size ϵ for D is taken and the dictionary elements are normalized to have L2 norm of one. The value of ϵ is annealed during learning. The dimensionality of the sparse code is three times the effective dimensionality of the data (computed by using PCA to find the number of components that account for 90% of the variance). Because convergence is usually poor for a fullyconnected dictionary, D, the dictionary elements are divided into 10 groups. The first group has dictionary elements whose values are constrained to be zero except the top left 16 × 16 part of the pattern. The next set of dictionary elements have nonzero values with this 16 × 16 patch is shifted by 8 pixels to the right. Doing all such shifts gives nine groups. The last group is fullyconnected.

For the $L1$ part of the sparse coding natural scenes prior, β is chosen through cross validation in the full simulation. For the $L2$ part of the sparse coding natural scenes prior, the mean and covariance of the sparse codes of 10^4 held-out patterns were computed in the sparse coding simulations (e.g., minimizing $(S - DA)^2 + k|A|$).

For the natural scenes data, a prior based on second order statistics (PCA) was also considered. The covariance matrix of the input patterns in a pixel representation, C, can be written as $C = PVP^T$, where P is an orthonormal matrix and V is a diagonal matrix with non-negative entries. P is used as the dictionary, D. If μ is the mean of the data after converting to the PCA basis, then $-\log p(A) = 0.5 * \sum_k (A_k - \mu_k)^2 V_{k,k}^{-1}$. This prior is implemented by setting $\hat{A}_0 = \mu$ and $\hat{H}_0 = V$ (see the definitions of $\hat{A}$ and $\hat{H}$ below). For the PCA basis, note that the whitening filter does not fully whiten the data because of the high frequency cut-off in our whitening filter. The 20% of the principal components that contributed the smallest amount of variance were removed to improve numerical stability of the inference algorithm (the prior takes the inverse of variance associated with each of these component, which are very small numbers).

Algorithm for inferring shape and position from the retinal spikes

The following algorithm is a causal and online method for decoding a stimulus from spikes when the eye position is unknown (SI, Section 1). The algorithm requires storing three quantities in memory:

(1) $q_t(X_t)$ is the algorithm's current estimate of the position of the eye at time t. This distribution is estimated as $q_t(X_t) = \Sigma_p W_{t,p} \cdot \delta(X_t, X_{t,p})$ where $X_{t,p}$ is a collection of positions, and $W_{t,p}$ are the corresponding weights.

(2) $\hat{A}_t$, is a vector of size N_l (the number of latent factors) that represents the algorithm's estimate of the underlying spatial pattern, represented in a latent space, after looking at spikes in the time interval [0, t].

(3) $\hat{H}_t$ is a matrix of size N_l by N_l that represents the inverse of the covariance associated with the estimate of A_t after looking at spikes in the time interval $[0, t]$.

The algorithm consists of the following steps:
1. Initialize $\hat{A}_0$ and $\hat{H}_0$ (set to 0 unless otherwise noted).
2. Update q:

$$q_{t+1}(X_{t+1}) \sim p(R_{t+1}|X_{t+1}, S = D\hat{A}^t) \times \sum_{X_t} p(X_{t+1}|X_t) q_t(X_t) \quad (21)$$

A particle filter with resampling (Doucet & Johansen, 2009) is used to estimate this equation with $N_p = 20$ particles (performance saturates at this N_p). $p(X_{t+1}|X_t)$ is used for the proposal distribution (see Equations 48 and 49)
3. Update the estimate of the latent factors:

$$\hat{A}^{t+1} = \text{argmin}_A \left[E_g(A) + E_r^{t+1}(A) + E_p(A)\right] \quad (22)$$

$$E_g(A) = \frac{1}{2}(A - \hat{A}^t)^T H_t (A - \hat{A}^t) \quad (23)$$

$$-E_r^t(A) = \langle \log p(R_t|X_t, S = DA)\rangle_{q_t(X_t)} \quad (24)$$

$$-E_p(A) = [\log p(A)] - (A - \hat{A}^t)\frac{\partial \log p(A)}{\partial A}|_{A=\hat{A}^t} \quad (25)$$

$$+\gamma * \sum_i \Theta(S_i - s_{max}) + \Theta(-(S_i - s_{min})) \quad (26)$$

where $\Theta(x) = x$ for $x \geq 0$, and zero otherwise. The minimization is executed using the FISTA algorithm (Beck & Teboulle, 2009) with 320 gradient steps per timestep (often not necessary, but ensures minimization). The Lipschitz constant is chosen using cross validation. A neural interpretation emerges when writing out the FISTA equations for this minimization (SI, Section 2).
4. Update the value for the Hessian (SI, Section 3):

$$\hat{H}^{t+1} = \hat{H}^t + \frac{\partial^2}{\partial A^2} E_r^{t+1}(A)|_{A=\hat{A}^{t+1}} \quad (27)$$

Indeed, this is an online algorithm because the previous state, $(q, \hat{A}, \hat{H})_t$, is combined with new data, R_{t+1}, to calculate the new state $(q, \hat{A}, \hat{H})_{t+1}$.

Full derivation

Although there are many possible decoders, the goal of this work is to create a neurally plausible decoder with the following properties. First, the algorithm should be causal (e.g., not using information from the future to infer the current state). Second, the algorithm should be an online algorithm. That is to say that the algorithm has a finite memory buffer that it updates using observations. A Kalman filter is a good example of an algorithm of satisfying these two requirements. Third, the algorithm should work with a pattern representation that does not necessarily consist of pixels, but where each neuron could have a structured (e.g., oriented) receptive field as in V1 (as compared with pixel receptive fields that result from an independent pixel prior). Fourth, the algorithm should be computable in a nonexponential amount of time as is typical in the full Bayesian treatment of many problems.

Our model is specified using a probabilistic graphical model. If N are all the random variables of the graphical model, then the joint distribution is $p(N) = \prod_i p(N_i|N_{\pi(i)})$, where $\pi(i)$ denotes the parents of node i in the graph defining the model. All other quantities of interest are computed by marginalizing the resulting distribution. Our approach is based on using the EM algorithm to approximate $\text{argmax}_A p(A|R)$ and then expanding the resulting equations using a Gaussian approximation to the data terms. In an ideal Bayesian world, one would compute $p(A|R)$. This assigns a probability for every possible set of latent factors given the observed spikes. Because this is intractable, the argmax is taken:

$$\hat{A} = \text{argmax}_A p(A|R)$$
$$= \text{argmax}_A \sum_{S,X} p(A, S, X|R) \quad (28)$$

$$= \text{argmax}_A \log \sum_{S,X} p(A, S, X, R) \quad (29)$$

$$= \text{argmax}_A \log \sum_X p(R|X, S = DA)p(X) + \log p(A) \quad (30)$$

The evaluation of the sum over X (for fixed A) involves numerically integrating over all timesteps of the simulation using a particle filter (which may not behave well numerically). Because the expression must be evaluated many times for each value of A during optimization, EM is used (Moon, 1996). Following the traditional EM recipe (which comes from introducing a variational distribution q), first initialize $A \rightarrow A'$ and then alternate between two steps:

$$q(X) \leftarrow p(X|R, S = DA') \quad (31)$$

$$A' \leftarrow \text{argmax}_A \sum_X q(X) \log p(R|X, S = DA) + \log p(A) \quad (32)$$

The temporal structure of the problem allows for a number simplifications. First, the logarithm splits up the evidence coming in at different times. Expanding out the logarithm shows us that only the marginals of $q(X)$ at each time point are needed to evaluate this expression.

$$\text{argmax}_A \sum_X \sum_t q(X) \log p(R_t|X_t, S = DA) \quad (33)$$

$$= \text{argmax}_A \sum_t \sum_{X_t} q_t(X_t) \log p(R_t|X_t, S = DA) \quad (34)$$

$$= \text{argmax}_A \sum_t \langle \log p(R_t|X_t, S = DA) \rangle_t \quad (35)$$

$$= \text{argmin}_A \sum_t E_r^t(A), \quad (36)$$

where $q_t(X_t)$ is the X_t marginal of $q(X)$ (recall $X = (X_0, \ldots X_T)$, where T is the total number of timesteps in the simulation) and the sum over X_t weighted by $q_t(X_t)$ is abbreviated as $\langle \cdot \rangle_t$. According to the EM equations, the marginals of $q(X)$ are

$$q_t(X_t) = \sum_{X_{-t}} p(X|R, S = DA') = p(X_t|R_{0:T}, S = DA') \quad (37)$$

$$\approx p(X_t|R_{0:t}, S = DA'), \quad (38)$$

where X_{-t} denotes the set $\{X_{t'}|t' \neq t\}$. Conditioned on a fixed A, the model is a hidden Markov model and the desired marginals are the smoothing estimate of the position. The smoothing estimate is replaced with the filtering estimate in order to get a causal position estimator (i.e., no spikes from the future are used to estimate the current position of the eye). A particle filter is used to estimate these marginals. Next, the optimization for A is modified to be causal. First, the summation over all time is broken up to only use the data up to a time t.

$$\hat{A}^t = \text{argmin}_A \sum_{t'=0}^{t} E_r^{t'}(A) - \log p(A), \quad (39)$$

where q_t is an estimate of X_t at time t and $\hat{A}_t$ is the current estimate of the latent factors. This computation requires memory that grows linearly with t. We seek an online algorithm. The sum of data terms from $t' \in [0, t]$ is modified such that all of the terms in the past are replaced by a Gaussian approximation expanded about the estimate of the position at that time point:

$$\sum_{t'=0}^{t} E_r^{t'}(A) \approx \sum_{t'=0}^{t} \frac{1}{2}(A - \hat{A}^{t'})^T \bar{H}_t (A - \hat{A}^{t'}) + \bar{B}_t^T (A - \hat{A}^{t'}) + \bar{C}_t \quad (40)$$

$$= \frac{1}{2}(A - \hat{A}^t)^T \left[\sum_{t'=0}^{t} \bar{H}_t \right] (A - \hat{A}^t) + B_t^T (A - \hat{A}^t) + C_t \quad (41)$$

$$H_t = \left[\sum_{t'=0}^{t} \bar{H}_t \right] = \sum_{t'=0}^{t} \frac{\partial^2}{\partial A^2} E_r^{t'}(A)|_{A=\hat{A}^{t'}} \quad (42)$$

$$H_t = H_{t-1} + \frac{\partial^2}{\partial A^2} E_r^t(A)|_{A=\hat{A}^t}. \quad (43)$$

The second line follows from collecting the terms that are quadratic in A, and noting the leftover polynomial is linear in A. H, B and C do not depend on A. The polynomial expansion would be more accurate if the evidence terms were expanded about the most accurate value of A possible, but for the algorithm to be online, the best available estimate of A is used. C_t does not need to be computed because we are optimizing relative to A. B_t could be computed recursively, but it is simpler to note that $\hat{A}_t$ is a local minima:

$$B - \frac{\partial \log(A)}{\partial A}|_{A=\hat{A}^t} = 0 \quad (44)$$

The prior is kept separate because the sparsity prior is not twice differentiable. Putting this approximation of the past data into the equation to be optimized, we get:

$$\hat{A}^{t+1} = \text{argmin}_A \frac{1}{2}(A - \hat{A}^t)^T H_t (A - \hat{A}^t) + E_r^{t+1}(A) \quad (45)$$

$$+ (A - \hat{A}^t)^T \frac{\partial \log p(A)}{\partial A}|_{A=\hat{A}^t} - \log p(A) \quad (46)$$

To prevent problems in the inference of A from the pixels going out of bounds in the log-linear firing rate equation, a term is added to the cost function that penalizes the entries of the vector $S = DA$ from going out of bounds. The quadratic prior is incorporated by initializing H_0 and $\hat{A}_0$ to be nonzero.

Particle filtering

The equation for the E step of the EM algorithm after the causal approximation requires us to calculate $q_t(X_t) = p(X_t|R_{0:\,t}, S = DA)$ where A is a fixed estimate of the latent variables. Given a fixed A, the model is a hidden markov model, whose probabilities can be estimated using sequential importance sampling. Following the tutorial (Doucet & Johansen, 2009), suppose that one has a sequence of distributions ($t = 0, \ldots T$): $\pi(X_{0:t}) = \frac{\gamma(X_{0:t})}{Z_t}$, where π is normalized but γ is not normalized. Sequential Monte Carlo (SMC) methods help one to sample from these distributions iteratively. SMC methods estimate this distribution using a collection of samples with weights: $\pi(X_{0:t}) \approx \sum_p W_t^p \delta(X_{0:t}, X_{0:t}^p)$. The weights are defined by adding an auxillary distribution that is easier to sample from called an importance density, or a proposal distribution, $r_t(X_{0:\,t})$ chosen by the user. The weights are then calculated by the following equation: $w_t(X_{0:t}) = \frac{\gamma_t(X_{0:t})}{r_t(X_{0:t})}$. To have the computation per sampling step not increase linearly in time, one typically uses a factorized importance density: $r_t(X_{0:\,t}) = r_{t\,-\,1}(X_{0:\,t\,-\,1})r_t(X_t|X_{t\,-\,1})$. In this article, the unnormalized distribution is:

$$\gamma(X_{0:t}) = p(X_{0:t}, R_{0:t}|A) = \prod_{t'=0}^{t} p(R_{t'}|S = DA, X_{t'}) \times \prod_{t'=1}^{t} p(X_{t'}|X_{t'-1})p(X_0) \quad (47)$$

The corresponding normalized distribution is $\pi(X_{0:\,t}) = p(X_{0:\,t}|R_{0:\,t}, A)$ and the normalizer is $Z_t = p(R_{0:\,t}|A)$. For the EM algorithm requires the estimation of the X_t marginal of $\pi(X_{0:\,t})$. In this article, we use the proposal distribution: $r_t = p(X_t|X_{t\,-\,1})$, which is a Gaussian. Given the SMC framework, there are a number of sampling techniques that can be used to estimate the sequence of distributions. This article uses sequential importance resampling, which achieves good performance and is easy to implement. Each E step is achieved by executing the following steps:

(1) Sample according to the proposal distribution:

$$X_t^p \sim r_t(X_t|X_{t-1}^p) = p(X_t|X_{t-1}^p) \quad (48)$$

(2) Compute the weights

$$W_t(X_{0:t}^p) \sim W_{t-1}(X_{0:t-1}^p) \times \frac{p(R_t|S = DA, X_t^p)p(X_t^p|X_{t-1}^p)}{r_t(X_t^p|X_{t-1}^p)} = W_{t-1}(X_{0:t-1}^p)p(R_t|S = DA, X_t^p) \quad (49)$$

(3) Resample if the effective sample size goes below threshold (e.g., one-half of the number of particles). Resampling is done using systematic resampling, which takes $O(|p|)$ steps.

The ESS is defined as the reciprocal of the sum of the squares of the weights. If the weights are all equal, then the ESS is the number of particles. One subtlety in this sampling process is that at step t, W_t is the weight associated with $X_{0:\,t}$. That is to say, at each step, each particle represents a full trajectory of the eye from 0 to t. Thus, we get an approximation of $\pi(X_{0:\,t})$. Because the proposal distribution only looks at the most recent position, we need to only store the current position associated with each particle instead of the entire trajectory.

It is worth noting that, in the case of a spiking model with spike history dependencies, the SMC framework would allow one to sample the appropriate distributions at the expense of needing to store not only the current position associated with each particle, but also the portion of the trajectory that summarizes the history relevant to the spiking model.

No motion decoder

As a control for the motion benefit owing to motion, a decoder that assumes that there is no motion is considered. In this case, the model collapses into a simple model $\hat{A} = \text{argmax}_A p(A|R) = \text{argmax}_A \log p(R|X = 0, S = DA) \sim \sum_j \bar{R}_j \log \lambda_{t,j} - \lambda_{t,j}$, where $\bar{R}_j$ is the average firing rate of neuron j. This loss function is optimized using AdaDelta (other methods are also possible).

SNR

Given a ground truth pattern, S, and an estimated pattern, $S^{'} = DA^{'}$, the SNR is computed as follows. First, the average difference between the estimated path and the true path is computed. This is used to shift the estimated pattern, (call that $S^{''}$). Then $S \cdot S/(S - S^{''}) \cdot (S - S^{''})$ is computed where each pattern is represented by a sum of Gaussians. For example, if $\{U, V\} = \sum_i \{U, V\}_i N(x; x_i^{\{u,v\}}, \sigma)$, then $U \cdot V = \int dx \sum_{i,i'} U_i V_{i'} N(x; x_i^u, \sigma) N(x; x_{i'}^v, \sigma) = \sum_{i,i'} U_i V_{i'} N(0; x_i^u - x_{i'}^v, \sqrt{2}\sigma)$.

Creating a dynamical system by following the gradient

Because a gradient-based optimization scheme is used, the latent factors evolve according to the dynamical system: $\frac{dA}{dt} \sim -\frac{dE}{dA}$. For simplicity, consider an ON cell where $s_{min} = 0$ and $s_{max} = 1$.

$$-\frac{\partial E_r^t(A)}{\partial A} = \frac{\partial}{\partial A} \sum_j \langle \log p(R_{t,j}|X_t, S = DA) \rangle_t \quad (50)$$

$$= \sum_j \frac{\partial}{\partial A} \langle R_{t,j} \log \lambda_{t,j} - \lambda_{t,j}\, dt \rangle_t \quad (51)$$

$$= \sum_j \langle [R_{t,j} - \lambda_{t,j}\, dt] \frac{\partial \log \lambda_{t,j}}{\partial A} \rangle_t \quad (52)$$

$$\Delta A_k \sim -\frac{\partial E_t}{\partial A_k} + [H_t(A - \hat{A}^t)]_k + [\delta(A)]_k \quad (53)$$

$$\delta(A) = \frac{\partial \log p(A)}{\partial A}|_{A=\hat{A}^t} - \frac{\partial \log p(A)}{\partial A} \quad (54)$$

$$-\frac{\partial E_t}{\partial A_k} \sim \sum_j R_{t,j} \langle g_{j,k}(x) \rangle_t - \langle \lambda_{t,j} g_{j,k}(x) \rangle_t dt \quad (55)$$

$$g_{j,k}(x) \equiv g \sum_i D_{i,k} T_{i,j}(x) \quad (56)$$

using the fact that $\frac{\partial \log \lambda_{t,j}}{\partial A_k} = g \log \frac{\lambda_1}{\lambda_0} * \sum_i D_{i,k} T(x_t)_{i,j} \sim g \sum_i D_{i,k} T(x_t)_{i,j}$. The sign of this derivative is flipped for the OFF cells. $[\,\cdot\,]_k$ denotes extracting the kth entry of a vector.

The equation for the derivative of the data term admits a neural interpretation. The position-dependent gain factor, $g_{j,\,k}(x)$, is the product of the connection strength between pixel i and neuron j, $T_{i,\,j}(x)$, and the dictionary. The result is the connection strength between RGC j and latent factor k. The equation for the derivative of the data term, E_t, has two parts. In the first part, the average position-dependent gain factor is computed by averaging over the internal position estimate. This value modulates the impact of the spike R_j on latent factor k. The second term is a decay term that looks at the expected number of spikes that are coming in given the current estimate of the latent factors. In particular, $\lambda_{t,\,j}dt$ is the number of spikes that the circuit expects to come in the interval $[t, t + \Delta t]$.

Second derivative of spike log-likelihood

Taking the derivative of Equation 50 gives:

$$\frac{\partial^2}{\partial A_{k'} \partial A_k} E_r^t(A)$$

$$= -dt \sum_j \left\langle \lambda_{t,j} \frac{\partial \log \lambda_{t,j}}{\partial A_{k'}} \frac{\partial \log \lambda_{t,j}}{\partial A_k} \right\rangle_t \quad (57)$$

noting that the second derivative of $\log \lambda$ with respect to A is zero because it is a log-linear model. As before, the derivative of $\log \lambda$ is replaced to get

$$\frac{\partial^2}{\partial A_{k'} \partial A_k} E_r^t(A) \sim \sum_{j,p} W_{p,t} \lambda_{t,j,p} g_{j,k'}$$

$$\times (X_{t,p}) g_{j,k}(X_{t,p}), \quad (58)$$

where the proportionality constant is $dt \left(\log \frac{\lambda_1}{\lambda_0}\right)^2$ and the particle filter has weights $W_{p,\,t}$ associated with positions $X_{t,\,p}$.

Although this equation is ostensibly complex, a direction for future work is to run simulations with a simpler uncertainty estimate. Whereas λ is the internal estimate of the number of incoming spikes per unit time, it can be approximated by a constant. Furthermore, replacing g with its definition gives:

$$\sim \sum_{j,p,i,i'} W_{p,t} D_{i',k'} T_{i',j}(X_{t,p}) D_{i,k'} T_{i,j}(X_{t,p}) \quad (59)$$

$$= D^T \left(\sum_j \left\langle T_{i',j}(X) T_{i,j}(X) \right\rangle_t \right) D, \quad (60)$$

where the object in the parenthesis is a matrix with indices i', i. Recalling the definition $T_{i,\,j}(x)$ is a Gaussian with a constant standard deviation across i and j and a mean $X_j^E - X_i^S$ (j indexes the position of the cones and i indexes the position of the pixels of the pattern). This uncertainty is independent of the incoming measurements given the estimated position.

Alternative representations

It should be noted that, although our model recovers an explicit stabilized representation of the object, it is also possible that these computations could be done in a nonstabilized representation. From the perspective of Bayesian inference, there are observations, R, and a hidden state, X, A. To map the problem into a hidden Markov model, consider the random variables A_t for $t \in \{0, \ldots T\}$, where $A_{t+1} = A_t$ and $A_0 = A$. These quantities are related by an observation model $p(R|X_t, A_t)$ and a state transition model $p(A_t, X_t|A_{t-1}, X_{t-1})$. In principle, it is possible to do a change of variables from X_t, A_t, to another set of hidden variables, which would result in a different neural representation (e.g., an unstabilized representation). In particular, suppose that the representation of the stimulus is in retinotopic coordinates. Define $\bar{A}_t$ to be the latent factors representing the stimulus as it lands on the retina at time t. For example, $\bar{A}_t = \bar{T}_{X_t} A$ where $\bar{T}$ is the translation operator that acts on the latent factors.

Although this simplifies the observation model, it complicates the state transition model. For example, we would need to update $\bar{A}_{t+1}$ from $\bar{A}_t$ and $X_{t+1} - X_t$. This would require the circuit to know how to compute a translation in an arbitrary direction in the current encoding of the pattern (e.g., if T_X is the translation operator in pixel space, then the circuit would need to implement $\bar{T}_X \approx D^{-1} T_X D$, which is a translation operator in the latent factor space). In experiments with such a model, we found it difficult to model that translation operator. More theoretical work on a translation operator that acts on a sparse code of a pattern could enable such a model.

This idea can be explored further in equations. Define $\Delta X_t = X_{t+1} - X_t$. Then write out the Bayesian equations and use conditional independences in the model:

$$p(A_{t+1}|R_{0:t+1}) \sim \quad (61)$$

$$\sum_{\Delta X_t, A_t} p(A_{t+1}, \Delta X_t, A_t, R_{t+1}|R_{0:t}) \quad (62)$$

$$= \sum_{\Delta X_t, A_t} p(R_{t+1}|A_{t+1}) p(A_{t+1}|A_t, \Delta X_t) \times p(\Delta X_t) p(A_t|R_{0:t}) \quad (63)$$

$$= p(R_{t+1}|A_{t+1}) * \sum_{\Delta X_t} p(\Delta X_t) * p(A_t$$

$$= \bar{T}_{-\Delta X_t} A_{t+1}|R_{0:t}) \quad (64)$$

using the additional fact that the motion prior used in this work is translation invariant (e.g., $p(\Delta X_t|X_t) = p(\Delta X_t)$) and $A_{t+1} = \bar{T}_{\Delta X_t} A_t$. Compared with before, the observation model is simpler: $p(R_{t+1}|A_{t+1}, X_{t+1}) = p(R_{t+1}|A_{t+1})$ because the location of the object in the world in retinotopic coordinates determines the spikes. However, there is a more complex hidden state update equation for A_t that does not have a simple analytical form.

Diffusion constant comparison

To facilitate comparisons between different models, a method to match the diffusion constants for different models of eye movements is presented. Regardless of the model, it should be the case that the quantity $\frac{E[(X(t+\Delta t) - X(t))^2]}{\Delta t}$ is the same constant.

(1) Discrete time diffusion on a lattice (as in (Burak et al., 2010)): Diffusion happens on a rectangular lattice with lattice spacing a. Each of the four possible steps happens with a probability $D_C \Delta t$. Thus, the ratio is $\frac{4 D_C \Delta t a^2}{\Delta t} = 4 D_C a^2$. For the majority of their paper, $a = \frac{1}{2}$ is used.

(2) Continuous time diffusion in continuous space (as in (Kuang et al., 2012)): Here, diffusion is modeled using the diffusion equation $u_t = \frac{1}{2} D_C \nabla^2 u$. In this equation, the variance as a function of time is $2 D_C \Delta t$, so the ratio is $2 D_C$.

(3) Discrete time in a continuous space (this article): Position is updated as $X_{t+1} = X_t + (D_C \Delta t/2)^* \epsilon$ where ϵ is drawn from a 2-D standard normal distribution. Thus the expected difference is $D_C \Delta t/2$ for each component of the eye position, so the total expectation divided by Δt is D_C.

Thus, $4 D_C^{burak} a^2 = 2 D_C^{kuang} = D_C^{us}$.

Supplementary Material

Supplementary Movie S1. Reconstruction as a function of time for the EM decoder. Each frame of the video shows the simulation results after a certain amount of time (same parameters as in Figure 2). (Top left) stimulus moving relative to the retina over time. (Bottom left) the reconstructed pattern visualized in pixel space. (Middle) exponential moving average of the spikes from the ON and OFF cells as a function of time. Although the OFF cells fire in the absence of a stimulus, such spiking does not convey additional information — only the spatial variations in the spiking conveys useful information for the decoder. The spikes are the input to our decoding algorithm. Right: decoded eye position (blue) relative to the true eye position (green) as a function of time. The shaded regions show plus or minus 1 standard deviation of the estimate. Although the position is quickly known to a relatively high certainty, the pattern needs a longer integration time before becoming sharp. This is because the edges of the stimulus are relatively sharp and many measurements (i.e., spikes) contribute to the estimate of position (a 2D quantity). In contrast, the stimulus is a higher dimensional quantity that needs to be inferred with the same number of observations.