Bayesian model of dynamic image stabilization in the visual system

Yoram Burak[a], Uri Rokni[a], Markus Meister[a,b], and Haim Sompolinsky[a,c,1]

[a]Center for Brain Science, Harvard University, Cambridge, MA 02138; [b]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138; and [c]Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel

Edited by William T. Newsome, Stanford University, Stanford, CA, and approved September 17, 2010 (received for review May 8, 2010)

Humans can resolve the fine details of visual stimuli although the image projected on the retina is constantly drifting relative to the photoreceptor array. Here we demonstrate that the brain must take this drift into account when performing high acuity visual tasks. Further, we propose a decoding strategy for interpreting the spikes emitted by the retina, which takes into account the ambiguity caused by retinal noise and the unknown trajectory of the projected image on the retina. A main difficulty, addressed in our proposal, is the exponentially large number of possible stimuli, which renders the ideal Bayesian solution to the problem computationally intractable. In contrast, the strategy that we propose suggests a realistic implementation in the visual cortex. The implementation involves two populations of cells, one that tracks the position of the image and another that represents a stabilized estimate of the image itself. Spikes from the retina are dynamically routed to the two populations and are interpreted in a probabilistic manner. We consider the architecture of neural circuitry that could implement this strategy and its performance under measured statistics of human fixational eye motion. A salient prediction is that in high acuity tasks, fixed features within the visual scene are beneficial because they provide information about the drifting position of the image. Therefore, complete elimination of peripheral features in the visual scene should degrade performance on high acuity tasks involving very small stimuli.

computation | fixational eye motion | neural network | retina | cortex

Our brain infers the structure of its surroundings from the signals of sensory neurons. When those signals are noisy, their interpretation becomes ambiguous, and multiple hypotheses about the outside world compete. Here we consider how the brain estimates a 2D image of the visual scene on the basis of the neural signals from optic nerve fibers. Ambiguity in this process derives from two primary sources: noise in the neural circuitry of the retina and random movements of the eye that lead to image jitter on the retina. An ideal Bayesian decoder in the brain would take these sources of ambiguity into account and evaluate the likelihoods of different 2D scenes leading to the spike trains from the retina. However, the full probability distribution of an image with many pixels includes an unfathomably large number of variables. Prior work on Bayesian inference focused on simplified problems in which the subject estimates only a single, typically static sensory variable (1–5). Thus there is considerable uncertainty whether Bayesian inference of full images is practicable at all. We begin by laying out the stochastic constraints on this process.

Humans with normal vision can resolve visual features spanning less than an arcminute, or approximately two receptive fields of ganglion cells in the central fovea, where each ganglion cell receives input from a single cone photoreceptor. Indeed, the letters "E" and "F" on the 20/20 line of a Snellen eye chart differ by just a few photoreceptors (Fig. 1*A*). While we perform this discrimination, the letter drifts across the retina over distances much larger than its own size. In the time between two subsequent spikes of any ganglion cell, the image shifts across several receptive fields (Fig. 1*A*), so that the cell is driven by a different part of the visual scene by the time the second spike is emitted. To properly decode the image from these spikes, it would seem that downstream visual areas require knowledge of the image trajectory. The image jitter on the retina during fixation is a combined effect of body, head, and eye movements (6, 7). Whereas the brain can often estimate the sensory effects of self-generated movement using proprioceptive or efference copy signals, such information is not available for the net eye movement at the required accuracy (8–10) (reviewed in ref. 11). Thus the image trajectory must be inferred from the incoming retinal spikes, along with the image itself. In so doing, an ideal decoder based on the Bayesian framework would keep track of the joint probability for each possible trajectory and image, updating this probability distribution in response to the incoming spikes (5, 11). However, the images encountered during natural vision are drawn from a huge ensemble. For example, there are 2^{900} possible black-and-white images with 30×30 pixels, which covers only a portion of the fovea. Clearly the brain cannot represent a distinct likelihood for each of these scenes, calling into question the practicality of a Bayesian estimator in the visual system.

Here we propose a solution to this problem, based on a factorized approximation of the probability distribution. This approximation introduces a dramatic simplification, and yet the emerging decoding scheme is useful for coping with the fixational image drift. We present a neural network that executes this dynamic algorithm and could realistically be implemented in the visual cortex. It is based on reciprocal connections between two populations of neurons, of which one encodes the content of the image and the other the retinal trajectory.

Results

To address how the visual system may deal with random drift we need, first, a model of how retinal ganglion cells (RGCs) respond to light falling on the retina, a model of the visual stimulus, and a model for how the stimulus is shifted relative to the photoreceptor array. Each one of these ingredients is probabilistic. Together, they define the likelihood of every possible stimulus given the spikes generated by the retina.

We model the fovea as a homogeneous array of retinal ganglion cells of a single type, arranged on a rectangular grid (Fig. 1*A*). The images consist of black-and-white pixels on this same grid, whose intensities are drawn independently from a binary distribution. The firing of each cell is an inhomogeneous Poisson process whose rate depends on the image pixel in the receptive field. We begin with a simple model where the cell responds instantaneously, firing at a rate λ_1 if the pixel is *on* and at a background rate λ_0 if it is *off*. Later, we consider a more realistic version where the rate depends on the past light intensity within the retina's integration time. The

Author contributions: Y.B., U.R., M.M., and H.S. designed research; Y.B., U.R., and H.S. performed research; and Y.B., M.M., and H.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: haim@fiz.huji.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1006076107/-/DCSupplemental.

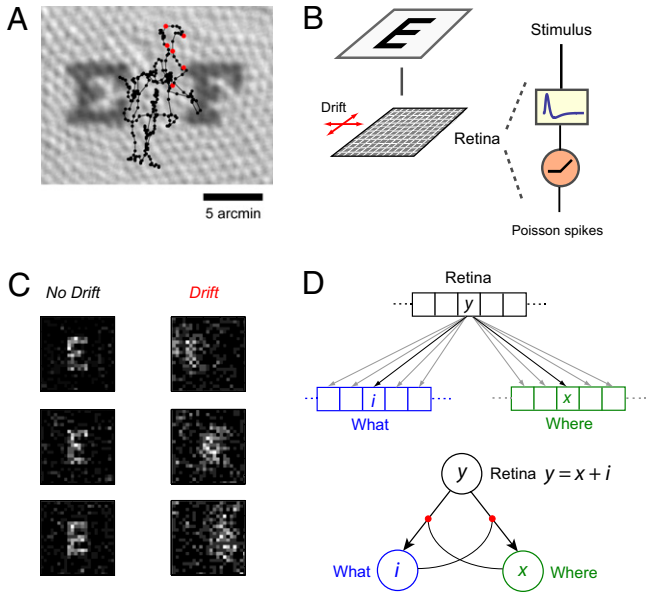

Fig. 1. (*A*) The letters E and F on the 20/20 line of the Snellen eye chart test, projected on an image of the foveal cone mosaic (photoreceptor image modified from ref. 39). The 1-arcmin features that distinguish the letters extend over only a few cones. Also shown is a sample fixational eye movement trajectory for a standing subject (courtesy of ref. 12), sampled every 2 ms for a duration of 500 ms and then smoothed with a 4-ms boxcar filter. Red dots mark the spike times from a neuron firing at 100 Hz. (*B*) Diagram of model for spike generation; see text for details. (*C*) Spikes generated by our model retina, presented with a letter E spanning 5 arcmin for 40 ms (with instantaneous RGC response), (*Left*) with no image drift and (*Right*) with image drift following statistics of human fixational eye motion. (*D*) Architecture of a neural implementation of the factorized decoder. (*Upper*) Each RGC projects to multiple *what* and *where* cells. (*Lower*) The projections are reciprocally gated between the two populations.

fixational movements of the image over the retina are modeled as a discrete random walk (12).

Spike Accumulation and the Magnitude of Fixational Motion. It is instructive to consider first what an ideal decoder would do if the image trajectory was known. An incoming spike from RGC i could then be associated uniquely with the pixel $i - x(t)$, where $x(t)$ is the known position of the image at the discharge time of the cell. After this routing of spikes to pixels, the performance would be the same as for a static image. Due to the noisy nature of ganglion cell firing, the decoder must accumulate spikes over a minimal time interval. For example, using firing rates of $\lambda_0 = 10$ Hz and $\lambda_1 = 100$ Hz, the letters on the "20/20" line of the Snellen eye chart can be estimated to reasonable accuracy within 40 ms (Fig. 1*C*, *Left*).

Without some knowledge of the image trajectory, such a reconstruction is impossible. Human eye movements resemble a random walk with a diffusion coefficient $D \simeq 100$ arcmin^2/s (11–13). In the 40-ms interval considered above, the resulting image drift can cover some 200 different pixels. Indeed, images of a Snellen letter derived from simple spike accumulation in each pixel seem almost random (Fig. 1*C*, *Right*). Thus one is led to a decoding scheme that estimates the image trajectory and uses it to reconstruct the content of the image.

Factorized Bayesian Decoder. The ideal decoder of such spike trains would use Bayes' rule to continuously update a probabilistic estimate of the image s and the retinal position x, on the basis of all of the spikes observed up to time t. Because the number of possible images s is prohibitively large, we explored an approximate strategy that maintains the Bayesian inference scheme, but with a dramatically simplified representation of the probabilities. Specifically, the full Bayesian estimate is approximated by a factorized posterior distribution

$$p(s,x,t) = p(x,t)\prod_i p_i(s_i,t), \qquad [1]$$

where $p(x, t)$ is a probability distribution of positions and $p_i(s_i, t)$ are probability distributions for individual pixels in the stabilized coordinates of the image. This form ignores any correlations between the values of different pixels or between the image and its position. To update the posterior after a short time interval, Δt, while maintaining its factorized structure, we perform two steps. First, the factorized posterior $p(s, x, t)$ is updated according to the incoming spikes between t and $t + \Delta t$, on the basis of Bayes' rule. Subsequently, the result is recast into the factorized form. This recasting leads to update rules that are derived in the *SI Appendix* and are summarized below. We define $m_i(t)$ to be the estimated probability that $s_i = 1$: $m_i(t) = p_i(1, t) = 1 - p_i(0, t)$.

Update between spikes. Between spikes the dynamics of $p(x, t)$ are described by a diffusion equation,

$$\frac{\partial p(x,t)}{\partial t} = D\nabla^2 p\,(x,t), \qquad [2]$$

which reflects the increasing uncertainty about position due to the random walk statistics of image drift. The dynamics of $m_i(t)$ are described by the differential equation

$$\frac{\partial m_i(t)}{\partial t} = -\Delta\lambda[1 - m_i\,(t)\,]m_i(t), \qquad [3]$$

where $\Delta\lambda = \lambda_1 - \lambda_0$. Thus, $m_i(t)$ decays toward zero in the absence of spikes, with a rate proportional to $\Delta\lambda$. We note also that if m_i is either 0 or 1, such that the decoder is certain about the value of pixel i, m_i remains unchanged.

Update due to a spike. If ganglion cell k fires a spike at time t, then $p(x, t)$ changes as

$$p\,(x,t_+) \propto [\lambda_0 + \Delta\lambda m_{k-x}\,(t_-)\,]\cdot p(x,t_-), \qquad [4]$$

where t_+ designates the time right after the update, t_- represents the time right before the update, and a multiplicative prefactor keeps the probability distribution normalized. The quantity in the brackets is the estimated firing rate of ganglion cell k if the image is at position x. Thus, $p(x, t_-)$ is multiplied by the estimated likelihood that ganglion cell k has produced a spike. The update to the estimate of pixel i, following a spike in cell k, is

$$m_i\,(t_+) = m_i\,(t_-) + \phi\,[m_i\,(t_-)\,]\cdot p(k-i,t_+), \qquad [5]$$

where $m_i(t_-)$ is the value immediately before the spike, $m_i(t_+)$ is the updated value following the spike, and $\phi(m)$ is the nonlinear function $\phi(m) = \Delta\lambda m(1 - m)/(\lambda_0 + \Delta\lambda m)$. Therefore, the change in m_i is proportional to the estimated probability that the image is at position $k - i$.

Network Implementation. In contrast to the ideal Bayesian decoder, we can envision a neural implementation of the factorized decoder because the number of probabilities that must be tracked grows only linearly with the number of pixels. The update rules (Eqs. **2**–**5**) are particularly suggestive of an implementation that involves two populations of neurons: One represents the probability of image position $p(x)$ and the other the probability of pixel intensities m_i. We refer to these two populations as *where* and *what* neurons.

Within such an implementation, the update rules (Eqs. **4** and **5**) indicate how spiking of each RGC affects the activities of

multiple *where* and *what* cells (Fig. 1*D*, *Upper*). The effect of ganglion cell k on *what* cell i is gated in a multiplicative fashion by the activity of *where* cell $x = k - i$. In turn, the update to *where* cell x in response to a spike from ganglion cell k is gated by the activity of *what* cell $i = k - x$. This result suggests a network architecture with two divergent projections from retinal ganglion cells to the *what* cells and the *where* cells, along with reciprocal recurrent connections between both of these populations (Fig. 1*D*, *Lower*). The diffusion dynamics and normalization of $p(x, t)$ can be implemented by horizontal excitatory connection and divisive global inhibition within the *where* population.

For concreteness, we describe the factorized decoder in terms of the above neural implementation, although other implementations are possible.

Performance of the Factorized Decoder. The response of the factorized decoder to a sample stimulus is illustrated in Fig. 2*A*. Activity in the *where* population successfully tracks the position of the image. The estimate of the image itself, represented by activity in the *what* population, gradually improves with time. In this example almost all of the pixels are estimated correctly at 300 ms, the duration of a typical human fixation. The *what* population effectively encodes the stabilized image, from which the effects of eye motion have been removed.

Fixational image movements must be taken into account. When tested with many random images, the factorized decoder routinely reconstructed 90% of the pixels correctly in just 100 ms (Fig. 2*B*). By comparison, a *static* decoder that ignores eye movements and simply accumulates spikes performed very poorly: Shortly after stimulus onset it reached a maximum of nearly 60% correctly estimated pixels, but then the blurring from retinal motion took its toll. Clearly, the tracking of image movement is essential for successful reconstruction.

Performance improves with slower eye movements, higher firing rates, and larger image size. When D is small, the decoder easily tracks the position of the image, and performance is limited only by the stochasticity of the ganglion cell response. As D increases, the performance degrades due to uncertainty about the position (Fig. 3*A*). The convergence time increases sharply above a critical value of D. This value is proportional to the RGC firing rates, as can be deduced from dimensional analysis. With a larger image, more information is available about the trajectory, and the decoder's performance improves markedly (Fig. 3*B*). Further analysis shows that increasing the number of pixels by a factor f acts roughly like a reduction of D by a factor $\sqrt{f}$ (*SI Appendix, Section II*). This sensitivity to image size should be observable in psychophysical experiments.

Performance under conditions of human vision. With D set to 100 arcmin^2/s, corresponding to the measured statistics of human fixational drift (11–13), the factorized decoder performs well on images that cover at least 40×40 pixels (20×20 arcmin) (Fig. 3*B*). Reconstruction improves dramatically if one is satisfied with a lower resolution. For example, if the pixel size is increased from 0.5 to 1 arcmin, then the eye drift changes the pixel contents less rapidly, and four ganglion cells are available to report each pixel. Under these conditions, small 5×5 arcmin images can be decoded rapidly to high accuracy (Fig. 3*B*).

Dynamics of the Retinal Response. So far we assumed that RGCs modulate their firing rate instantaneously in response to the stimulus. More realistically, RGCs integrate light in their re-

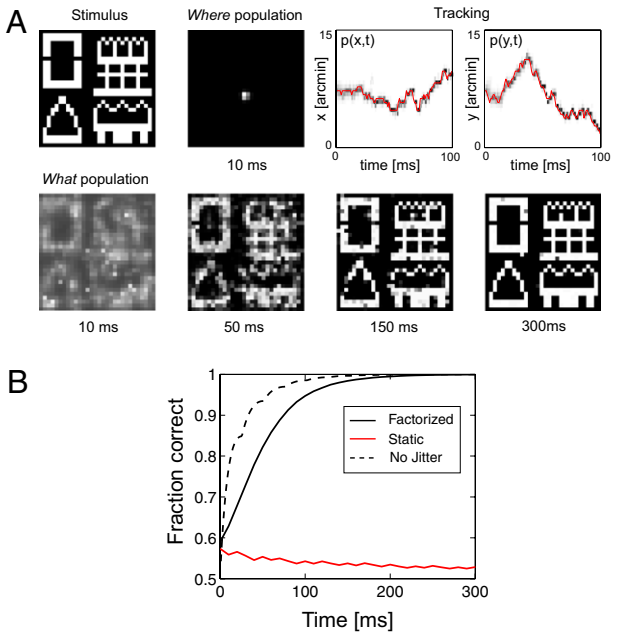

Fig. 2. (*A*) Example of image reconstruction by the factorized decoder. (*Upper*) From left to right: the stimulus; snapshot of activity in the *where* cell population at t = 10 ms; and tracking of horizontal and vertical image position over time, with probability (grayscale) compared with actual trajectory (red). Parameters: 30×30 pixels, 0.5 arcmin/pixel, $\lambda_{0,1}$ = 10/100 Hz, D = 100 arcmin^2/s. (*Lower*) Several snapshots of activity in the *what* cell population. (*B*) Fraction of correctly estimated pixels as a function of time, averaged over 100 randomly selected images each containing 50×50 pixels and spanning 25×25 arcmin. Spikes generated with image motion are presented to the factorized and static decoders (solid traces). Performance of static decoder is shown also for a static image (dashed trace).

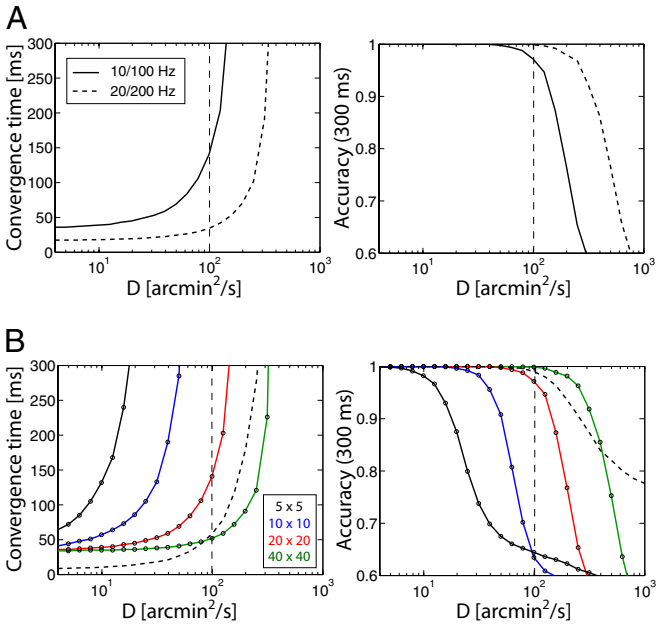

Fig. 3. (*A*) Performance as a function of D, averaged over 1,000 presentations of random images. The convergence time (at which 90% of pixels are estimated correctly) increases with D (*Left*) and the accuracy (fraction of correctly estimated pixels at t = 300 ms) decreases with D (*Right*). Results are shown for images containing 40×40 pixels (20×20 arcmin). Increasing the firing rate improves performance ($\lambda_{0,1}$ = 10/100 Hz, solid traces; $\lambda_{0,1}$ = 20/200 Hz, dashed traces). (*B*) Performance improves with image size. Solid traces show performance for several image sizes, indicated in the *Inset* in units of arcminutes. Dashed trace shows reconstruction of 5×5 arcmin images consisting of 1×1 arcmin pixels. In all other traces resolution is 0.5×0.5 arcmin. Vertical dashed lines designate the value of D that corresponds to measured statistics of human fixational eye motion (11–13).

ceptive field over a time window of ~100 ms with a biphasic impulse response (Fig. 4*A*, *Inset*) (14). Thus, a spike from a given RGC conveys partial information about all of the pixels that passed through the cell's receptive field within the integration time. Therefore eye movements affect the quality of image inference even in a hypothetical scenario where the decoder knows the image trajectory. Indeed, in this scenario, ~250 ms are required to accurately identify pixels in a drifting image at a resolution of 0.5 arcmin (Fig. 4*A*) whereas, with a small D, the required time is only 50 ms (Fig. 4*A*). These estimates for a known trajectory serve as an upper bound for any decoder that infers the image in the more realistic case of unknown trajectory (*SI Appendix*).

Because spike generation depends not only on the current image position but also on its history, a fully Bayesian decoder would need to track a probability distribution for every possible trajectory in the past ~100 ms. Given how many such trajectories exist, this approach seems unrealistic. Instead we explored performance of the above factorized decoder that ignores the dynamics of the retinal response. When presented with spike trains produced by the dynamic response model, this decoder fails to stabilize an image spanning 40 × 40 arcmin with a pixel resolution of 0.5 arcmin (Fig. 4*B*). However, if the resolution is lowered to 1 arcmin, this naive decoder performs quite well, estimating correctly 90% of the pixels in ~200 ms. Thus, the factorized decoder can successfully infer pixels at 1 arcmin resolution, over the typical time interval between saccades. As in the simpler case where RGC response is instantaneous, reducing the size of the stimulus to 5 × 5 arcmin leads to significant degradation in performance, which should be observable in psychophysical experiments (Fig. 4*B*, *Inset*).

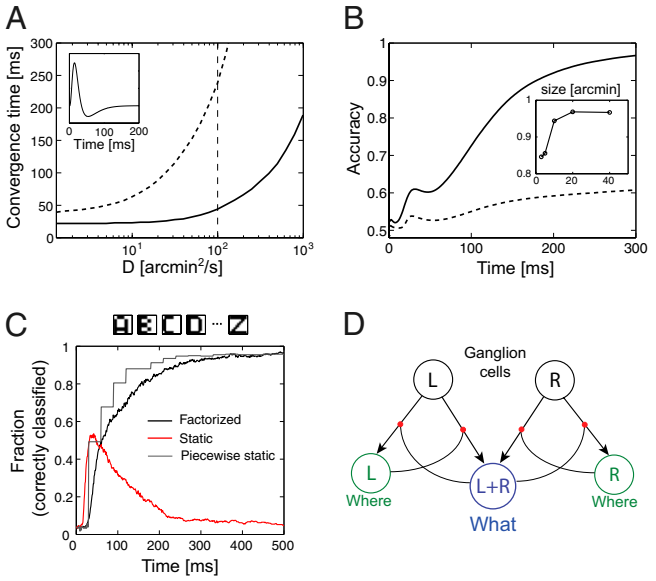

Fig. 4. Performance for spike trains generated with a temporal filter in RGC response. (*A*) Convergence time when the trajectory is known to the decoder. In contrast to the case of instantaneous response, performance depends on the diffusion statistics. Traces show the convergence time (for 90% accuracy), as a function of D for a factorized decoder that takes into account the filter (*SI Appendix, Section III*). Parameters: 20 × 20 pixel images, 0.5 arcmin/pixel (dashed trace) and 1 arcmin/pixel (solid trace). For known trajectory, image size has little effect (*SI Appendix*). Vertical dashed line: D = 100 arcmin²/s. (*Inset*) The temporal filter $f(\tau)$. (*B*) Performance of the naive factorized decoder when spikes are generated with a temporal filter (unknown trajectory). Traces show fraction of correctly estimated pixels as a function of time, averaged over 1,000 presentations of random images of sizes 40 × 40 arcmin, with D = 100 arcmin²/s. Solid and dashed traces: 1 × 1 arcmin and 0.5 × 0.5 arcmin pixels, respectively. The nonmonotonic dependence at short times is related to the structure of the temporal filter and can be eliminated using a modified version of the update rules (*SI Appendix, Section III,* and Fig. S3). (*Inset*) Accuracy at t = 300 ms measured for several image sizes, with 1 × 1 arcmin pixels (average over 1,000 presentations). (*C*) Performance on a discrimination task between 26 patterns representing the letters A–Z, averaged over 400 trials (see main text for all other parameters). Factorized decoder, black trace; static decoder, red trace; piecewise static decoder (*Discussion* and *SI Appendix*), gray trace. (*D*) Architecture of a neural implementation of the factorized decoder for binocular vision (*Discussion*).

Discrimination Task. It is useful also to assess the performance of the factorized decoder on a task for which there are clear performance measures from human psychophysics. We thus considered a discrimination task similar to the 20/20 row of the Snellen eye chart (Fig. 4*C*). The 26 possible images represent the letters A–Z; each letter subtends 5 arcmin and occupies 10 × 10 pixels on a 30 × 30 background of *off* pixels. Spikes are generated by a model retina with a biphasic temporal filter and diffusion coefficient D = 100 arcmin²/s and fed into the decoder. We evaluated the posterior probability for each letter and performed a maximum-likelihood decision. The decoder achieves a 90% success rate after ~300 ms, about the length of a human fixation, and is thus compatible with human vision on this task. To test whether trajectory tracking is required on this task, we also considered the simple static decoder that ignores eye movements altogether. The static decoder reaches peak performance ~40 ms after stimulus onset, when it correctly identifies the letter in ~50% of the trials, far short of human performance on this task.

Discussion

We proposed a computation by which the brain might interpret the spikes obtained from the fovea of the retina, while taking into account the statistics of image drift and the noisy nature of retinal responses. First, our analysis confirmed the intuition that the visual system must indeed take fixational movements into account to achieve high acuity vision. Simply integrating the retinal spikes with downstream neurons, while ignoring the eye movements, results in poor performance inconsistent with human abilities (Figs. 2*B* and 4*C*). Our proposed strategy therefore simultaneously estimates the image and its trajectory on the retina (Fig. 2). The method relies on Bayesian inference and thus needs to grapple with the "curse of dimensionality" from the combinatorially large ensemble of random images. To circumvent this challenge, the factorized decoder keeps track of separate probability distributions for each pixel in the image and for the image position. We hypothesize that this strategy is implemented in the brain by a neural network architecture that involves two cell populations, one that tracks the position of the image and another that accumulates evidence about the image content in a stabilized representation devoid of any image drifts (Fig. 1*D*).

Dependence on Image Size. The performance of the decoder is sensitive to the size of the presented image, because it rests largely on the estimate of the image trajectory. In our model this estimate was based only on spikes from the foveal region of the retina, which also encode the image itself. However, the ocular drift trajectory is common to all parts of the visual field. Thus the brain might use signals from more peripheral areas for estimating the trajectory, their sheer number possibly outweighing the sharp decrease in spatial resolution compared with the fovea. Additionally, direction-selective ganglion cells specialized to encode fine image motion might be recruited for the task. We therefore suggest that careful control of peripheral cues may be instructive in psychophysical measurements of visual acuity. For small stimuli a few arcminutes in size, embedded in a featureless background, we expect to see a significant degradation of fine spatial vision, compared with conditions where a larger area is stimulated or fixed features are added in the peripheral visual field (Figs. 3*B* and 4*B*).

Alternative Approaches. The detailed architecture and nonlinearity of the circuit model, Fig. 1*D*, shares notable similarities with the previously proposed *shifter circuits* for invariant object recognition (15, 16): Information from the retina is dynamically routed to form a stabilized representation of the image, based on multiplicative control signals representing the eye's position. Here we show that for retinal image stabilization, the control signal can be derived from the retinal inputs, as was previously suggested in the context of visual attention and invariant object recognition (17) (see also ref. 18), and we propose a specific algorithm to achieve this. Furthermore, our approach treats in a probabilistic framework the signal-to-noise levels of retinal responses and the statistics of rapid eye movements. Hence the nature of the computations and their neuronal implementation are more complex than the deterministic shifter circuit model.

By stabilizing the retinal image, as proposed here, fixational image motion is dealt with once and for all by dedicated neural circuitry that performs the same computation regardless of the image content. Subsequent stages of the visual system can then probe the content of this stabilized image to perform any number of visual tasks without needing to deal with image jitter. This division of labor is functionally attractive, but one can imagine an alternative scenario in which the visual system deals with fixational motion separately whenever it analyzes the foveal image for a specific visual task. We tested this scenario for the letter discrimination task (Fig. 4*C*) and found that, in principle, such an approach may be successful: Whereas the spikes from a single 30-ms time window were not sufficient to discriminate between letters, a procedure that accumulates evidence from many subsequent windows performed quite well (Fig. 4*C*). This strategy, which we call the *piecewise static* decoder (*SI Appendix*), involves two steps: First, in each short time window, generate a position-invariant likelihood that each of the possible letters is in the image, using the static decoder. Second, summate these log-likelihoods across windows to accumulate evidence over time, while ignoring the continuity of the trajectory across adjoining windows.

The piecewise static decoder does not involve an intermediate stage where the image is represented in stabilized coordinates. Compared with the factorized decoder, the piecewise static decoder seems complicated, because intricate neural circuitry must be set up for each possible pattern and every kind of visual task. Additionally, position-invariant pattern recognition apparently takes place late in the visual cortex, long after inputs from the two eyes have converged. Therefore, it would be difficult to eliminate the relative jitter of the two eyes, compared with a solution based on neural circuitry at an early stage of the visual process.

When the temporal response properties of RGCs are taken into account, eye motion has two competing effects within our model. On one hand, it introduces ambiguity in the interpretation of retinal spikes. On the other hand, it helps drive the RGCs, whose response to completely static stimuli is weak. Previous analysis of ideal discrimination between two small stimuli at the limit of visual acuity suggested that a small drift would be beneficial, but the actual eye movements of human subjects are much larger and on balance deleterious (11). This was confirmed in the present analysis for larger images at the resolution limit (Fig. 4*A*). For other visual tasks involving coarser features, the smearing effect of eye movements will be less severe, and the beneficial effect, coming from more robust activation of the RGCs, will be more prominent. Indeed, recent eye-tracking experiments demonstrated that fixational drift can be beneficial under those conditions (19).

The global image shifts introduced by eye movements are such a prominent aspect of the retinal input that one imagines multiple strategies may have evolved to deal with them. Indeed, certain types of retinal ganglion cells appear designed to ignore global image motion entirely and respond only when an object moves relative to the background scene (20). Clearly these RGCs cannot contribute to a reconstruction of static scenes. Their version of image processing—implemented already within retinal circuits—can be seen as complementary to the image stabilization discussed here.

We considered here only the smooth fixational drifts between saccades or microsaccades (6). A broader question is how the brain forms a stable scene representation across saccades (21). The computational principles presented here may be useful also for treatment of these larger motions. However, the size and speed of saccades are much larger than those of fixational drift, and it seems unlikely that the brain deals with both extremes of eye motion using the same neural circuitry.

Implementation in the Brain. We considered image pixels as the fundamental units that are reconstructed by the factorized decoder. More realistically, if the computation is performed in the visual cortex (see below), the decoder may represent probabilities for presence of more complex features, such as oriented edges.

Our neural implementation of the factorized decoding strategy has several salient features. First, the computation requires a divergence of afferents from ganglion cells to the populations of *what* and *where* units (Fig. 1*D*). The required span of divergence to the *what* population is determined by the typical range of fixational drifts, ~10 min of arc in each direction, whereas the number of *what* cells should correspond at least to the size of the fovea. The *where* cells need represent only the possible range of drift, and because this range is smaller than the size of the fovea, we expect far fewer *where* cells than *what* cells. Thus, every ganglion cell in the foveal region is expected to synapse into a subset of the *what* cells and into all *where* cells. Second, the dynamic routing of information from the retina to the *what* and *where* populations requires a multiplicative gating controlled in a reciprocal fashion by the signals in those populations (Fig. 1*D*). Multiplicative gain is prevalent in sensory cortical areas (22, 23), and many mechanisms for achieving it have been proposed (24–27). Third, in the *where* population, local excitatory connections (28) are required to implement the diffusive update between spikes, and a global divisive mechanism (24, 25, 29, 30) is needed to maintain normalization of the total activity. Finally, the rate dynamics in both populations involve local nonlinearities as described by Eqs. **4** and **5**.

Neural activity. What are the distinctive predictive features of activity in the *what* and *where* populations? The *what* cells represent a stabilized version of the image. Their receptive fields should shift on the retina according to the eye movements, but remain locked in the external visual space. Further, ramping firing rates after the onset of fixation should reflect the gradual accumulation of evidence about the image content. The *where* cells are expected to have large receptive fields, comparable at least to the size to the fovea. During conditions conducive to image tracking their activity should reflect the eye movement.

Location. Where might one find these circuits in the visual system? Fixational eye drifts are largely independent in the two eyes (31), so their compensation must occur within the monocular part of the visual pathway, including the lateral geniculate nucleus (LGN) and parts of V1. The LGN does not provide the required convergence of afferents from the retina, over an area ~20 arcmin in diameter. Thus the recipient circuits in V1 are the first stage at which fixational eye movements could be compensated.

It was suggested previously that primary visual cortex generates a stabilized representation of the visual image (32), but more recent work (33, 34) concluded that receptive fields of V1 neurons are fixed in retinal coordinates. In the present context, it is relevant that these recordings were from V1 cells in the parafoveal region with relatively large receptive fields 20–40 arcmin in diameter. For these neurons the receptive field diameter exceeds the total drift during a fixation, which obviates a strong need for stabilization. By the same token, these receptive fields, if they are indeed fixed on the retina, are too coarse to support visual acuity corresponding to 20/20 vision or the equivalent in macaques (35). Thus, the available

evidence does not exclude a network for fixational image stabilization within the foveal region of V1.

If, in fact, each of the two monocular pathways decodes the image independently, one needs to ask how their image estimates are combined. The simplest solution would be for both monocular decoders to feed the same image estimate. In the context of our factorized representation, this solution involves two monocular populations of *where* neurons that control the inputs to a single population of *what* neurons (Fig. 4*D*). Such a binocular representation of the stabilized image may appear in disparity-selective neurons in V1 or downstream of V1, for example in a binocular population in V2 that receives monocular inputs. To test these predictions it would be very instructive to record from cortical neurons that represent the primate fovea, whose receptive field structure is fine enough to resolve patterns close to the animal's acuity.

Methods

Stimulus and Simulated Spike Trains. We assume that the size *a* of each pixel matches the receptive field of a single RGC, and because there is little overlap between receptive fields in the fovea (36), each ganglion cell reports on the value of a single pixel (for 0.5 arcmin reconstruction; for 1 arcmin reconstruction, we assume that each pixels covers four receptive fields). For each presentation of the stimulus, we first generate a random walk trajectory for the image. Image shifts occur randomly with a rate $4D/a^2$ and Poisson statistics. Jump size is *a* and the direction is selected randomly with equal probabilities for up, down, left, and right shifts. We then evaluate the time-dependent firing rate of each RGC, determined either from the instantaneous pixel intensity at position or by the recent history as

$$\lambda_i(t) = \phi\left[\lambda_0 + \Delta\lambda \int d\tau f(\tau) s_{i-x(t-\tau)}\right], \quad [6]$$

where $x(t)$ is the position of the image at time t. The temporal kernel $f(\tau)$ is biphasic and is chosen as described (11) (see also ref. 14 and *SI Appendix*). We chose a background firing rate $\lambda_0 = 20$ Hz on the basis of measurements in macaque retina (37) and chose $\Delta\lambda$ such that the maximal possible firing rate of the neuron is 200 Hz. Firing rates are then almost always within the range 0–100 Hz (*SI Appendix*, Fig. S3*A*), chosen to match maximal firing rates observed in macaque retina (14, 38). The linear rectification function $\phi;(x) = \min(x, \lambda_C)$, where we chose the cutoff $\lambda_C = 1$ Hz. On the basis of the rates $\lambda_i(t)$, we generate a spike train for each RGC using inhomogeneous Poisson statistics. To simplify the numerical simulation, we use periodic boundary conditions and discretize time in steps $dt = 0.1$ ms.

Factorized Decoder. In Eq. **2** the Laplacian operator stands for a discrete operator, $\sum_{x' \in NN(x)} p(x', t) - 4p(x, t)$, where NN($x$) are the four nearest-neighbor locations near x. To speed up the numerical calculation we used a version of the update rules as described in *SI Appendix, Section I.E*, with a time step $dt = 0.1$ ms. In all simulations where the naive factorized decoder is applied to spikes generated with a temporal filter, the decoder assumes $\lambda_0 = 20$ Hz and $\lambda_1 = 100$ Hz. Measurements of accuracy were performed as described in *SI Appendix, Section V*.

ACKNOWLEDGMENTS. We thank Dan Lee, Ofer Mazor, and Xaq Pitkow for helpful discussions and Eran Mukamel for comments on the manuscript. We acknowledge support from the Swartz Foundation (Y.B. and U.R.), the National Eye Institute (M.M.), the Israeli Science Foundation (H.S.), and the Israeli Ministry of Defense (H.S.).

1. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J Neurosci* 12: 4745–4765.
2. Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J Neurophysiol* 86:1916–1936.
3. Rao RPN (2005) Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* 16:1843–1848.
4. Deneve S, Latham PE, Pouget A (2001) Efficient computation and cue integration with noisy population codes. *Nat Neurosci* 4:826–831.
5. Huys QJM, Zemel RS, Natarajan R, Dayan P (2007) Fast population coding. *Neural Comput* 19:404–441.
6. Martinez-Conde S, Macknik SL, Hubel DH (2004) The role of fixational eye movements in visual perception. *Nat Rev Neurosci* 5:229–240.
7. Skavenski AA, Hansen RM, Steinman RM, Winterson BJ (1979) Quality of retinal image stabilization during small natural and artificial body rotations in man. *Vision Res* 19: 675–683.
8. Guthrie BL, Porter JD, Sparks DL (1983) Corollary discharge provides accurate eye position information to the oculomotor system. *Science* 221:1193–1195.
9. Donaldson IM (2000) The functions of the proprioceptors of the eye muscles. *Philos Trans R Soc Lond B Biol Sci* 355:1685–1754.
10. Murakami I, Cavanagh P (2001) Visual jitter: Evidence for visual-motion-based compensation of retinal slip due to small eye movements. *Vision Res* 41:173–186.
11. Pitkow X, Sompolinsky H, Meister M (2007) A neural computation for visual acuity in the presence of eye movements. *PLoS Biol* 5:e331.
12. Engbert R, Kliegl R (2004) Microsaccades keep the eyes' balance during fixation. *Psychol Sci* 15:431–436.
13. Eizenman M, Hallett PE, Frecker RC (1985) Power spectra for ocular drift and tremor. *Vision Res* 25:1635–1640.
14. Chichilnisky EJ, Kalmar RS (2002) Functional asymmetries in on and off ganglion cells of primate retina. *J Neurosci* 22:2737–2747.
15. Anderson CH, Van Essen DC (1987) Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proc Natl Acad Sci USA* 84:6297–6301.
16. Olshausen BA, Anderson CH, Van Essen DC (1995) A multiscale dynamic routing circuit for forming size- and position-invariant object representations. *J Comput Neurosci* 2: 45–62.
17. Olshausen BA, Anderson CH, Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J Neurosci* 13:4700–4719.
18. Arathorn DW (2002) *Map-Seeking Circuits in Visual Cognition: A Computational Mechanism for Biological and Machine Vision* (Stanford Univ Press, Palo Alto, CA).
19. Rucci M, Iovin R, Poletti M, Santini F (2007) Miniature eye movements enhance fine spatial detail. *Nature* 447:851–854.
20. Olveczky BP, Baccus SA, Meister M (2003) Segregation of object and background motion in the retina. *Nature* 423:401–408.
21. Melcher D, Colby CL (2008) Trans-saccadic perception. *Trends Cogn Sci* 12:466–473.
22. Salinas E, Thier P (2000) Gain modulation: A major computational principle of the central nervous system. *Neuron* 27:15–21.
23. Peña JL, Konishi M (2001) Auditory spatial receptive fields created by multiplication. *Science* 292:249–252.
24. Murphy BK, Miller KD (2003) Multiplicative gain changes are induced by excitation or inhibition alone. *J Neurosci* 23:10040–10051.
25. Mel BW (1993) Synaptic integration in an excitable dendritic tree. *J Neurophysiol* 70: 1086–1101.
26. Mehaffey WH, Doiron B, Maler L, Turner RW (2005) Deterministic multiplicative gain control with active dendrites. *J Neurosci* 25:9968–9977.
27. Chance FS, Abbott LF, Reyes AD (2002) Gain modulation from background synaptic input. *Neuron* 35:773–782.
28. Gilbert CD, Wiesel TN (1989) Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J Neurosci* 9:2432–2442.
29. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. *Vis Neurosci* 9: 181–197.
30. Carandini M, Heeger DJ, Movshon JA (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci* 17:8621–8644.
31. Steinman RM, Collewijn H (1980) Binocular retinal image motion during active head rotation. *Vision Res* 20:415–429.
32. Motter BC, Poggio GF (1990) Dynamic stabilization of receptive fields of cortical neurons (vi) during fixation of gaze in the macaque. *Exp Brain Res* 83:37–43.
33. Gur M, Snodderly DM (1997) Visual receptive fields of neurons in primary visual cortex (V1) move in space with the eye movements of fixation. *Vision Res* 37:257–265.
34. Tang Y, et al. (2007) Eye position compensation improves estimates of response magnitude and receptive field geometry in alert monkeys. *J Neurophysiol* 97: 3439–3448.
35. Merigan WH, Katz LM (1990) Spatial resolution across the macaque retina. *Vision Res* 30:985–991.
36. Schein SJ (1988) Anatomy of macaque fovea and spatial densities of neurons in foveal representation. *J Comp Neurol* 269:479–505.
37. Troy JB, Lee BB (1994) Steady discharges of macaque retinal ganglion cells. *Vis Neurosci* 11:111–118.
38. Shapley RM, Victor JD (1978) The effect of contrast on the transfer properties of cat retinal ganglion cells. *J Physiol* 285:275–298.
39. Roorda A, Williams DR (1999) The arrangement of the three cone classes in the living human eye. *Nature* 397:520–522.

Supporting Text: Bayesian dynamics of image stabilization in the brain

Yoram Burak,[1] Uri Rokni,[1] Markus Meister,[1,2] and Haim Sompolinsky[3,4]

[1] *Center for Brain Science, Harvard University, Cambridge, MA 02138, USA*

[2] *Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA*

[3] *Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem, 91904, Israel*

[4] *Center for Brain Science, Harvard University, Cambridge, MA 02138*

Contents

I. THE FACTORIZED BAYESIAN DECODER

In this section we consider the case where RGC response is instantaneous. We first define the problem mathematically (part A). We then derive the ideal Bayesian decoder (part B) and the factorized Bayesian decoder (part C). The implementation of the factorized decoder to the case of binary pixels yields equations (2)–(5) of the main text (part D, Eqs. (S21)–(S23) and (S25)). In simulations we used a version of the decoder in which time is discretized, as described in part E. Finally, we present a comparison between the factorized decoder and the ideal decoder for small 1-dimensional images (part F).

A. The problem

We assume a 2D image of n x n pixels with fixed intensities $\{s_i\}$. The image shifts with time with a trajectory $x\left(t\right)$. For simplicity, we assume that changes in $x(t)$ occur only at discrete times, separated by fixed intervals of duration Δt. Eventually we will take the limit of $\Delta t \to 0$. The trajectory is drawn by a Markov process with a transition matrix T:

$$P\left[x\left(t+\Delta t\right) \mid x\left(t\right)\right] = \delta_{x(t+\Delta t),x(t)} + T\left[x\left(t+\Delta t\right) \mid x\left(t\right)\right]\Delta t \tag{S1}$$

where

$$\sum_x T\left(x \mid x'\right) = 0 \tag{S2}$$

and the summation is over all N possible values of x. The derivations below are valid for any choice of the transition matrix. To describe two-dimensional diffusion we choose

$$T(x \mid x') = \begin{cases} D & , \quad |x-x'| = 1 \\ -4D & , \quad x = x' \\ 0 & , \quad |x-x'| > 1 \end{cases} \tag{S3}$$

We define the initial shift to be $x\left(0\right) = 0$. There is a set of n x n neurons, each observing a single pixel. The response $r_i\left(t\right)$ of neuron i is an inhomogeneous Poisson process, whose instantaneous

rate depends on the incident image pixel $s_{i-x(t)}$

$$P\left[r_i\left(t\right) \mid s_{i-x}\right] = \delta_{r_i(t),0}\left[1-\lambda\left(s_{i-x(t)}\right)\Delta t\right] + \delta_{r_i(t),1}\lambda\left(s_{i-x(t)}\right)\Delta t \tag{S4}$$

where we assumed that Δt is sufficiently small that in each interval the neuron emits at most one spike. The function λ describes the relation between instantaneous firing rate and incident pixel intensity. We assume that given the stimulus, the neurons are uncorrelated. In the following, we denote the vectors of pixel intensities and neural responses by omitting the index i, i.e. by s and $r\left(t\right)$, respectively. The problem is to infer, at any time t, the fixed vector of pixel intensities s, given the past spike trains $r\left(0\right),...,r\left(t\right)$. Specifically, we are interested in the continuous time limit, i.e. $\Delta t \to 0$.

B. The Bayesian filter

The optimal Bayesian estimation requires a computation of the posterior distribution $P\left[s, x\left(t\right) \mid r\left(0\right),...,r\left(t\right)\right]$, which we denote more shortly by $P\left(s,x;t\right)$. This distribution can be marginalized over $x\left(t\right)$ to obtain the the posterior of s. $P\left(s,x;t\right)$ can be computed iteratively by

$$P\left(s,x;t\right) = \frac{1}{Z}\prod_{i=1}^{N} P\left[r_i\left(t\right) \mid s_{i-x}\right] \sum_{x'=1}^{N} P\left(x \mid x'\right) P\left(s,x';t-\Delta t\right) \tag{S5}$$

where Z is a normalization constant. We substitute Eqs. (S4) and (S1) in Eq. (S5), and take $\Delta t \to 0$. At times when there are no spikes $P\left(s,x;t\right)$ evolves smoothly according to

$$\frac{\partial P\left(s,x;t\right)}{\partial t} = \left[R_{\text{tot}}\left(t\right) - \sum_{i=1}^{N}\lambda\left(s_i\right)\right] P\left(s,x;t\right) + \sum_{x'=1}^{N} T\left(x \mid x'\right) P\left(s,x';t\right) \tag{S6}$$

where $R_{\text{tot}}\left(t\right)$ is the total expected firing rate,

$$R_{\text{tot}}\left(t\right) = \sum_{s,x} P\left(s,x;t\right) \sum_{i=1}^{N}\lambda\left(s_i\right) \tag{S7}$$

When there is a spike at neuron i at time t_i, $P\left(s,x;t\right)$ evolves discontinuously according to

$$P\left(s,x;t_{i+}\right) = \frac{1}{R_i\left(t_{i-}\right)}\lambda\left(s_{i-x}\right) P\left(s,x;t_{i-}\right) \tag{S8}$$

where $R_i\left(t\right)$ is the firing rate expected at neuron i

$$R_i\left(t\right) = \sum_{s,x}\lambda\left(s_{i-x}\right) P\left(s,x;t\right) \tag{S9}$$

C. The factorized approximation

In the factorized approximation we approximate the posterior as a product of probabilities for position and for each pixel,

$$P\left(s,x;t\right)\cong P\left(x;t\right)\prod_{i=1}^{N}P_{i}\left(s_{i};t\right) \tag{S10}$$

To update $P(x;t)$ and $P_i(s_i;t)$ we first use Eq. (S5). We then recast the updated $p(s,x;t)$ into the factorized form by marginalizing it on x and s_i to obtain the updated $P(x;t)$ and $P_i(s_i;t)$, respectively. This procedure minimizes the Kullback-Leibler divergence between $p(s,x;t)$ and the updated factorized approximation. We insert Eq. (S10) into Eq. (S6), and obtain the dynamics of the factorized posterior when there are no spikes

$$\frac{\partial P\left(x;t\right)}{\partial t}=\sum_{x'=1}^{N}T\left(x\mid x'\right)P\left(x';t\right) \tag{S11}$$

$$\frac{\partial P_{i}\left(s_{i},t\right)}{\partial t}=\left[\rho_{i}\left(t\right)-\lambda\left(s_{i}\right)\right]P_{i}\left(s_{i},t\right) \tag{S12}$$

where $\rho_i(t)$ is the firing rate expected by the neuron which at time t observes s_i

$$\rho_{i}\left(t\right)=\sum_{s_{i}}\lambda\left(s_{i}\right)P_{i}\left(s_{i},t\right) \tag{S13}$$

Similarly, we insert Eq. (S10) into Eq. (S8) to obtain the discontinuous change in the factorized posterior when neuron i spikes

$$P\left(x;t_{i+}\right)=\frac{1}{R_{i}\left(t_{i-}\right)}\rho_{i-x}\left(t_{i-}\right)P\left(x;t_{i-}\right) \tag{S14}$$

where

$$R_{i}\left(t\right)=\sum_{x}\rho_{i-x}\left(t\right)P\left(x;t\right) \tag{S15}$$

and

$$\begin{aligned}P_{k}\left(s_{k},t_{i+}\right) &= P_{k}\left(s_{k},t_{i-}\right) \\ &\times \frac{1}{R_{i}\left(t_{i-}\right)}\left[\lambda\left(s_{k}\right)P\left(x=i-k;t_{i-}\right)+\sum_{x\neq i-k}\rho_{i-x}\left(t_{i-}\right)P\left(x;t_{i-}\right)\right]\end{aligned} \tag{S16}$$

which can be rewritten as

$$P_{k}\left(s_{k},t_{i+}\right)=\left\{1+\frac{\left[\lambda\left(s_{k}\right)-\rho_{k}\left(t\right)\right]P\left(x=i-k;t_{i-}\right)}{R_{i}\left(t_{i-}\right)}\right\}P_{k}\left(s_{k},t_{i-}\right) \tag{S17}$$

or, using Eq. (S14), as

$$P_{k}\left(s_{k},t_{i+}\right)=\left\{1+\frac{\left[\lambda\left(s_{k}\right)-\rho_{k}\left(t\right)\right]P\left(x=i-k;t_{i+}\right)}{\rho_{k}(t_{i-})}\right\}P_{k}\left(s_{k},t_{i-}\right) \tag{S18}$$

D. The binary s_i case

When s_i are binary variables, we can describe $P_i(s_i,t)$ by

$$m_i(t) = P_i(s_i = 1, t) \tag{S19}$$

In this case, we can write the firing rate function $\lambda(s_i)$ as

$$\lambda(s_i) = \lambda_0 + \Delta\lambda s_i \tag{S20}$$

where λ_0 is the firing rate for $s_i = 0$, λ_1 is the firing rate for $s_i = 1$, and $\Delta\lambda = \lambda_1 - \lambda_0$. The evolution of $P(x,t)$ in the absence of spikes is unchanged from Eq. (S11)

$$\frac{\partial P(x;t)}{\partial t} = \sum_{x'=1}^{N} T\left(x \mid x'\right) P\left(x';t\right). \tag{S21}$$

This is equation (2) in the main text, where the operator $D\nabla^2$ in discrete space represents the transition probabilities $T(x|x')$ of Eq. (S3). The evolution of $m_i(t)$ in the absence of spikes is derived from Eq. (S12) and yields Eq. (3),

$$\frac{\partial m_i(t)}{\partial t} = -\Delta\lambda\left[1 - m_i(t)\right] m_i(t) \tag{S22}$$

When neuron i fires a spike, $P(x,t)$ is updated by [see Eq. (S14)]

$$P(x;t_{i+}) = \frac{1}{R_i(t_{i-})}\lambda\left(m_{i-x}(t_{i-})\right) P(x;t_{i-}) \tag{S23}$$

where

$$R_i(t) = \lambda_0 + \Delta\lambda \sum_x m_{i-x} P(x;t) \tag{S24}$$

[Eq. (4)] and from Eq. (S18) we find that $m_k(t)$ is updated by

$$m_k(t_{i+}) = \left\{1 + \frac{\Delta\lambda P(x = i-k; t_{i+})\left[1 - m_k(t_{i-})\right]}{\lambda\left(m_k(t_{i-})\right)}\right\} m_k(t_{i-}) \tag{S25}$$

This can be written in the form of Eq. (5) by defining the nonlinear transfer function $\phi(m) = \Delta\lambda m(1-m)/(\lambda_0 + \Delta\lambda m)$.

E. Large Δt

In the above derivation we took the limit $\Delta t \to 0$. This limit is relevant for biological implementation, and it simplifies the equations. However, for the purpose of computer implementation

it requires taking a very small time step, such that at any given time step the probability that any neuron fires a spike is small. This constraint becomes more restrictive as the number of neurons in the model increase. To allow faster computer implementation, here we derive the algorithm without assuming a small time step. When $\lambda_i \Delta t \ll 1$ the implementation of this algorithm should be identical to the original algorithm up to small truncation and roundoff errors.

Our starting point is the equation of the full Bayesian filter (Eq. S5) which applies for non-vanishing Δt. Next, we apply the mean field approximation (see section on mean field approximation above) and derive the update rules:

$$P\left(x;t\right)=\frac{1}{Z}P'\left(x;t\right)\prod_{i}\sum_{s_i}P\left[r_{i+x}\left(t\right)\mid s_i\right]P\left(s_i;t-\Delta t\right) \tag{S26}$$

where

$$P'\left(x;t\right)=\sum_{x'}P\left(x\mid x'\right)P\left(x';t-\Delta t\right) \tag{S27}$$

and

$$P\left(s_i;t\right)=\sum_{x}P\left(x;t\right)\frac{P\left[r_{i+x}\left(t\right)\mid s_i\right]P\left(s_i;t-\Delta t\right)}{\sum_{s_i'}P\left[r_{i+x}\left(t\right)\mid s_i'\right]P\left(s_i';t-\Delta t\right)} \tag{S28}$$

Inserting (S4) into (S26) we obtain the following update rule for $P\left(x;t\right)$

$$P\left(x;t\right)=\frac{1}{Z}P'\left(x;t\right)\exp\left[\sum_{i}r_{i+x}\left(t\right)\log\frac{\rho_i\left(t-\Delta t\right)\Delta t}{1-\rho_i\left(t-\Delta t\right)\Delta t}\right] \tag{S29}$$

where $\rho_i\left(t\right)$ is defined in (S13). Similarly, by inserting (S4) into (S28) we obtain the update rule for $P\left(s_i;t\right)$

$$\begin{aligned}P\left(s_i;t\right) &= P\left(s_i;t-\Delta t\right) \\ &\times \sum_{x}P\left(x;t\right)\left[r_{i+x}\left(t\right)\frac{\lambda\left(s_i\right)}{\rho_i\left(t-\Delta t\right)}+\left(1-r_{i+x}\left(t\right)\right)\frac{1-\lambda\left(s_i\right)\Delta t}{1-\rho_i\left(t-\Delta t\right)\Delta t}\right]\end{aligned} \tag{S30}$$

Keeping terms up to first order in Δt yields

$$\begin{aligned}P\left(s_i;t\right) &= P\left(s_i;t-\Delta t\right)\times\left\{1-\left[\lambda(s_i)-\rho_i(t-\Delta t)\right]\Delta t\right. \\ &+ \left.\sum_{x}P\left(x;t\right)r_{i+x}\left(t\right)\frac{\lambda\left(s_i\right)-\rho_i(t-\Delta t)}{\rho_i\left(t-\Delta t\right)}\right\}\end{aligned}$$

from which the continuous time limit of Eqs. (S12) and (S18) follows by taking the limit $\Delta t\to 0$.

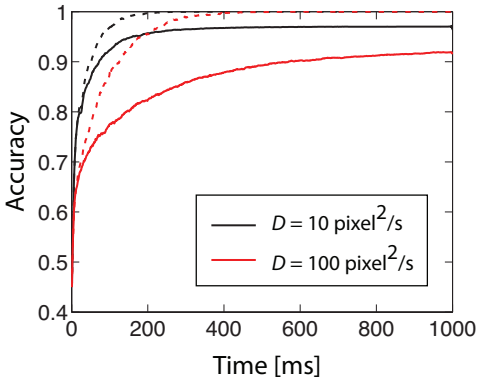

FIG. S1: Comparison of the factorized decoder (full traces) with the ideal Bayesian decoder (dashed traces) for small, one-dimensional images containing 10 pixels. Accuracy (fraction of correctly estimated pixels), averaged over 1000 presentations of random images, is shown as a function of time for two values of the diffusion coefficient (legend).

F. Comparison with the ideal filter

For small one-dimensional images, we can compare the performance of the factorized decoder with that of the ideal Bayesian decoder. Such a comparison is shown for images containing 10 pixels in Fig. S1. As expected, the factorized decoder infers the pixels more slowly than the ideal decoder. Further, the plots demonstrate that after a long presentation of the image, the factorized decoder may converge on an imperfect estimate of the image. In contrast, the ideal decoder infers all the pixels correctly if given enough time.

II. PERFORMANCE AS A FUNCTION OF IMAGE SIZE

A. Accuracy of tracking for a known image

We assume that the image is known and evaluate how well the decoder estimates the position of the image at steady state. The pixels are assumed to be uncorrelated with an equal distribution of *on* and *off* values. This part of the calculation applies to the optimal Bayesian decoder as well as to the factorized one, since both of them have the same dynamics when the image is known.

Specifically, we consider the following quantity

$$\log P(\Delta\mathbf{x}) \equiv \langle \log p(\mathbf{x}(t) + \Delta\mathbf{x}) \rangle_{\mathbf{x}(t),r} \tag{S31}$$

where $\mathbf{x}(t)$ is the true position of the drifting image at time t. The averaging is performed over all possible trajectories $\mathbf{x}(t)$ and over the ganglion cell firing patterns.

For convenience we work with a one-dimensional image containing n pixels and assume that image drift occurs in discrete steps: At each Δt the image moves one step to the left, with probability $D\Delta t$, or to the right, with the same probability. Ultimately we will be interested in the limit of small $\Delta t \to 0$. At this stage we only require that Δt is sufficiently small such that any individual neuron is unlikely to produce two spikes in a single time interval,

$$\lambda_0 \Delta t \ll 1 \quad , \quad \lambda_1 \Delta t \ll 1. \tag{S32}$$

In addition we assume that $D\Delta t \ll 1$.

The update of $p(x,t)$ is:

$$p(x, t+\Delta t) = \frac{1}{Z} p(r \mid x) \left\{ (1 - 2D\Delta t) p(x,t) + D\Delta t \left[p(x+1,t) + p(x-1,t) \right] \right\} \tag{S33}$$

To estimate the steady state behavior of $p\,(x,t)$ we make use of the following approximations. First, we approximate $\log p(r \mid x)$ by replacing it with its average over r. In the limit of large n this quantity is the same for all values of x except for the true position of the image:

$$\langle \log p(r \mid x) \rangle = \begin{cases} c_0 & x = x(t) \\ c_0 - n d_{\mathrm{KL}} \Delta t & x \neq x(t) \end{cases} \tag{S34}$$

where $d_{\mathrm{KL}} \Delta t$ is the Kullback-Leibler distance per pixel between the distribution of firing patterns given the correct image, and the distribution of firing patterns given a shifted version of the image,

$$d_{\mathrm{KL}} = \frac{1}{4} (\lambda_1 - \lambda_0) \log \frac{\lambda_1}{\lambda_0}. \tag{S35}$$

and where c_0 is independent of x. In deriving this expression we made use of the assumption that the pixels are drawn independently from a binary distribution. We thus replace the dynamics of Eq. (S33) by:

$$\begin{aligned} \log p(x, t+\Delta t) \;&=\; -\log Z + n d_{\mathrm{KL}} \Delta t \delta_{\mathbf{x},\mathbf{x}(t)} \\ &+\; \log \left\{ (1 - 2D\Delta t) p(x,t) + D\Delta t \left[p(x+1,t) + p(x-1,t) \right] \right\} \end{aligned} \tag{S36}$$

where Z is determined from the normalization requirement on p. Here the spiking is no longer considered as a stochastic process: The combined influence of all spikes on the Bayesian estimate is encapsulated deterministically in the second term on right hand side of Eq. (S36). In the limit $\Delta t \to 0$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} p(x) \;&=\; n d_{\mathrm{KL}} \delta \left[x - x(t) \right] p(x) + D \left[p(x+1) + p(x-1) - 2p(x) \right] \\ &-\; n d_{\mathrm{KL}} p(x^*) p(x) \end{aligned} \tag{S37}$$

As a further simplification we consider a particular trajectory, where the image remains at $x(t) = 0$ at all times. In other words, the image is static, but the estimator is tuned to a randomly drifting image with statistics characterized by the diffusion coefficient D. At steady state we then have,

$$D\left[p(x+1) + p(x-1) - 2p(x)\right] + nd_{\mathrm{KL}}p(0)\left[\delta_{x,0} - p(x)\right] = 0 \tag{S38}$$

This equation describes the steady state distribution of particles which are created by a point source at $x = 0$ and undergo one-dimensional random diffusion. The particles are created at a rate $nd_K Lp(0)$ and are randomly removed, independent of their position, at the same rate such that their total number remains constant in time. The parameter $p(0)$ must be obtained self consistently from the requirement that

$$\sum_{-\infty}^{\infty} p(x) = 1. \tag{S39}$$

By dividing this equation by D we see that the solution depends only on the ratio nd_{KL}/D. Therefore doubling the diffusion coefficient has the same effect as reducing the number of pixels by a factor of two. For two-dimensional images, however, n is replaced everywhere by n^2. Hence performance depends on the diffusion coefficient, scaled by the number of pixels. This is the main result of this section. We proceed to analyze the form of the solution to Eq. (S38) in the 1-d case. The solution to Eq. (S38) is found by assuming the ansatz

$$P(x) \propto \exp(-\alpha|x|). \tag{S40}$$

Inserting this expression in Eq (S38), we get

$$2\left[\cosh(\alpha) - 1\right] = \frac{nd_{\mathrm{KL}}p(0)}{D} \tag{S41}$$

From the normalization requirement (S39), $p(0) = \mathrm{tgh}(\alpha/2)$. Therefore

$$\alpha = \sinh^{-1}\left(\frac{nd_{\mathrm{KL}}}{2D}\right) \tag{S42}$$

Fig. S2 A shows a measurement of $\log P(\Delta x)$ (Eq. S31) from a long simulation of the factorized decoder tracking a known one-dimensional image containing 1000 pixels. These results compare well with the approximate analytical expression [Eqs. (S40),(S42)].

B. Accuracy of tracking for an unknown image

Here we consider the opposite limit where the image is completely unknown to the decoder. This is the situation when the image is first presented to the decoder. We consider an approximate

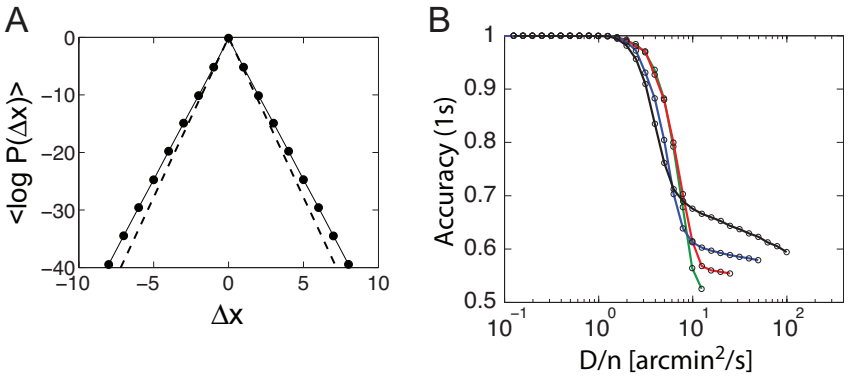

FIG. S2: **A** For large images the probability of position inferred by the decoder is sharply distributed around the true position. Symbols and solid trace show $\log P(\Delta x)$ [Eq (S31)] as a function of Δx, for a one-dimensional image containing 1000 pixels. These results are compared with the analytical estimate, Eqs. (S40) and (S42) (dashed trace). Parameters: $\lambda_{0,1} = 10/100\,\mathrm{Hz}$, $D = 200$ pixels^2/s. **B** Accuracy of decoded pixels after a long presentation (1 s) becomes roughly independent of the number of pixels n^2 when plotted as a function of D/n. Data is shown for $n = 10$ (black), 20 (blue), 40 (red), and 80 (green), which corresponds to 5 x 5, 10 x 10, 20 x 20, and 40 x 40 arcmin. All parameters are as in Fig. 3B.

decoder which assumes that the image moves only at discrete times, separated by regular intervals of duration Δt. We choose

$$\Delta t = \frac{1}{4D} \tag{S43}$$

because over this time scale it is reasonable to assume that the image is static. In order to track the position of the image accurately, the decoder must be able to infer the relative position of the image in the second interval, compared to its position during the first interval, which we denote by Δx^*. The inference is based on the RGC spike counts observed during the first and second intervals, which we denote by r and r'. In order for the decoder to successfully distinguish between a shift Δx and the true shift Δx^*, the following quantity must be large compared to unity for any $\Delta x \neq \Delta x^*$:

$$d_{\mathrm{KL}}\left[p(r, r' \mid \Delta x) \parallel p(r, r' \mid \Delta x^*)\right] \tag{S44}$$

This is the Kullback-Leibler divergence between the probability distribution of spike counts given Δx and their distribution given Δx^*. Assuming that pixels are statistically independent, and using the instantaneous response property of the neurons, we obtain:

$$d_{\mathrm{KL}}\left[p(r, r' \mid \Delta x) \parallel p(r, r' \mid \Delta x^*)\right] = n^2 \begin{cases} 0 & , \quad \Delta x = \Delta x^* \\ \hat{d}_{\mathrm{KL}} & , \quad \Delta x \neq \Delta x^* \end{cases} \tag{S45}$$

where

$$\hat{d}_{\mathrm{KL}} = \sum_{r_1=0}^{\infty}\sum_{r_2=0}^{\infty}\left[\sum_{s} p(s)p(r_1 \mid s)p(r_2 \mid s) - \sum_{s_1,s_2} p(s_1)p(s_2)p(r_1 \mid s_1)p(r_2 \mid s_2)\right] \times \log\left[\sum_{s'} p(s')p(r_1 \mid s')p(r_2 \mid s')\right] \tag{S46}$$

In this expression $p(s)$ is the prior probability for pixels intensities which, in the following, we assume is uniform, and the spike statistics are Poisson:

$$p(r_i \mid s) = \frac{\mathrm{e}^{-\lambda(s)\Delta t}\left[\lambda(s)\Delta t\right]^{r_i}}{r_i!} \tag{S47}$$

For binary images we can write $\hat{d}_{\mathrm{KL}}$ as

$$\hat{d}_{\mathrm{KL}} = \frac{1}{4}\left[J_{0,0} + J_{1,1} - 2J_{0,1}\right] \tag{S48}$$

where

$$J_{s_1,s_2} \equiv \sum_{r_1=0}^{\infty}\sum_{r_2=0}^{\infty} p(r_1 \mid s_1)p(r_2 \mid s_2)\log\left[\frac{p(r_1 \mid 0)p(r_2 \mid 0) + p(r_1 \mid 1)p(r_2 \mid 1)}{2}\right] \tag{S49}$$

The decoder estimates whether the image has moved by correlating the spike trains in the two time intervals. In the limit of small Δt these firing patterns are sparse, and we expect the information coming from the correlations to scale as Δt^2. A precise expansion in powers of Δt yields

$$\hat{d}_{\mathrm{KL}} = \alpha(\lambda_0, \lambda_1)\Delta t^2 + \ldots \tag{S50}$$

where

$$\alpha(\lambda_0, \lambda_1) = \frac{1}{4}\log\left[\frac{2(\lambda_0^2 + \lambda_1^2)}{(\lambda_0 + \lambda_1)^2}\right](\lambda_1 - \lambda_0)^2. \tag{S51}$$

This expression is valid if $\lambda_{0,1}\Delta t \lesssim 1$. The decoder can track the image accurately if $n^2\hat{d}_{\mathrm{KL}} \gg 1$. We thus obtain the requirement that

$$D \ll \frac{n\alpha^{1/2}}{4} \tag{S52}$$

This result suggests that the value of D, beyond which performance starts to degrade, should scale as n, the square root of the number of pixels, rather than by n^2 as suggested by the tracking of a known image (Sec. II A). Indeed, when the accuracy of pixels inferred by the factorized decoder is plotted as a function of D/n, the traces are seen to be roughly independent of D/n, Fig. S2 B.

III. TEMPORAL RESPONSE OF RETINAL GANGLION CELLS

We consider the situation where a temporal filter is involved in the response of RGCs. As before, we assume that each RGC fires as an inhomogeneous Poisson process but instead of Eq. (S4) we have

$$P\left[r_i\left(t\right) \mid s, X\right] = \left(1 - r_i\right)\left[1 - \lambda_i\left(s, X\right)\Delta t\right] + r_i \lambda_i\left(s, X\right)\Delta t \tag{S53}$$

where the rate $\lambda_i(s, X)$ is given by

$$\lambda_i(s, X) = \lambda_0 + \Delta\lambda \int d\tau f(\tau) s_{i-x(t-\tau)}. \tag{S54}$$

Here $f(\tau)$ is the temporal filter and we adopt the notation that X with a capital letter denotes a full trajectory, and $x(t)$ denotes the image position at a particular time t.

A. Ideal Bayesian Filter

We denote by $P(s, X; t)$ the posterior probability of the image s and the trajectory X given all the spikes emitted from time 0 up to time t. Between spikes,

$$\frac{\partial P(s, X; t)}{\partial t} = \left(\sum_i \lambda_i(s, X) - R_i\right) P(S, \mathbf{x}; t) + \sum_{X'} T\left(X \mid X'\right) P(s, Y; t) \tag{S55}$$

where

$$R_i = \sum_{X'} \sum_{s'} \lambda_i(s', X') P(s', X'; t). \tag{S56}$$

is the expected firing rate of neuron i. When neuron i spikes at time $t = t_i$,

$$P(s, X; t_i^+) = \frac{\lambda_i(s, X) P(s, X; t_i^-)}{Z} \tag{S57}$$

where Z is a normalization factor, chosen such that the sum

$$\sum_{s'} \sum_{X'} P(s', X'; t_i^+) = 1. \tag{S58}$$

B. Factorized approximation

We can apply the factorized approximation while keeping track of probabilities for full trajectories instead of only the current position:

$$P(s, X; t) \simeq \prod_i P_i(s_i; t) P(X; t). \tag{S59}$$

The update rules for $p_i(s_i;t)$ and $P(X;t)$ are obtained from Eqs (S55), (S57) by marginalizing over s_i and over the trajectory.

We begin with the updates between spikes. Evaluating the update rule for $P(X;t)$ involves averaging of the total firing rate over s given X. While this in general is complicated, it is simple in the linear case,

$$\sum_s P(s;t)\sum_i \lambda_i(s,X) = \lambda_0 + \Delta\lambda \int \mathrm{d}\tau f(\tau)\sum_i \sum_s P(s;t)s_{i-x(t-\tau)} \tag{S60}$$

where we used the notation:

$$P(s;t) = \prod_i P_i(s_i;t) \tag{S61}$$

The last term reduces to

$$\sum_i \sum_s P(s;t)s_{i-x(t-\tau)} = \sum_i m_{i-x(t-\tau)}(t) \tag{S62}$$

where $m_i(t)$ is the mean of s_i with respect to $P(s;t)$. This quantity is independent of X, and this leads to

$$\sum_s P(s;t)\sum_i \lambda_i(s,X) = \lambda_0 + \Delta\lambda f_t \sum_i m_i \tag{S63}$$

where

$$f_t = \int \mathrm{d}\tau f(\tau). \tag{S64}$$

Hence only the diffusion term survives,

$$\frac{\partial P(X,t)}{\partial t} = \sum_Y T(X \mid Y)P(Y,t). \tag{S65}$$

To compute the dynamics of $P_i(s_i;t)$, we denote by S^i the vector of all the s_j except s_i, and write,

$$\int d\tau f(\tau)\sum_{S^i} P(S^i;t)\sum_j s_{j-x(t-\tau)} = f_t\left(\sum_j m_j + s_i - m_i\right) \tag{S66}$$

Hence,

$$\frac{\partial P_i(s_i;t)}{\partial t} = \Delta\lambda f_t(s_i - m_i)P_i(s_i;t) \tag{S67}$$

We next consider the update following a spike in RGC i. Here we need to compute:

$$\sum_s P(s;t)\lambda_i(s,X) = \lambda_0 + \Delta\lambda \int \mathrm{d}\tau\, f(\tau)m_{i-x(t-\tau)} \tag{S68}$$

Hence,

$$P(X, t_i^+) = \frac{\lambda_0 + \Delta\lambda \int \mathrm{d}\tau f(\tau) m_{i-x(t-\tau)}}{R_i} P(X, t_i^-) \tag{S69}$$

where

$$R_i = \lambda_0 + \Delta\lambda \int \mathrm{d}\tau f(\tau) \sum_x m_{i-x} p_\tau(x; t_i^-) \tag{S70}$$

Here $p_\tau(x;t)$ equals the probability with respect to $P(X;t)$ that $x(t-\tau) = x$:

$$p_\tau(x;t) = \sum_X P(X;t) \delta_{X(t-\tau),x} \tag{S71}$$

For the probability of s_k, we write

$$\sum_X P(X;t) \sum_{S^k} P(S^k, t) s_{i-x(t-\tau)} = p_\tau(i-k;t)(s_k - m_k) + R_i \tag{S72}$$

so that,

$$P_k(s_k; t_i^+) = \left[1 + \frac{\Delta\lambda}{R_i} \tilde{P}(i-k; t_i^-)(s_k - m_k)\right] p_k(s_k; t_i^-) \tag{S73}$$

where

$$\tilde{P}(x;t) \equiv \int \mathrm{d}\tau \, f(\tau) p_\tau(x;t) \tag{S74}$$

We note that the update rules for $P_k(s_k;t)$ do not require knowledge of the full distribution over trajectories $P(X,t)$: Only the marginals $p_\tau(x;t)$ are required. Furthermore, the update rules for the pixels have precisely the same form as in the case without temporal filtering, if $P(x;t)$ is replaced by $\tilde{P}(x;t)$. (In the case without temporal filtering we assumed that the firing rate $\lambda(s_i)$ can be written as $\lambda(s_i) = \lambda_0 + \Delta\lambda s_i$). This is seen by comparing Eqs. (S67), (S70), and (S73) with Eqs. (S12), (S15), and (S17), respectively.

Known trajectory

If the trajectory is known, the dynamics between spikes are given by Eqs. (S67) and (S73), where in Eq.(S73) $\tilde{P}(x,t)$ is replaced by:

$$\tilde{P}(x,t) = \int \mathrm{d}\tau \, f(\tau) \delta_{X(t-\tau),x}. \tag{S75}$$

Rectification

So far we assumed that λ_0 is sufficiently large that $\lambda_i(t)$ remain positive at all times. In more realistic models of RGC responses, the firing rate involves rectification:

$$\lambda_i(s, X) = \phi\left[\lambda_0 + \Delta\lambda \int d\tau f(\tau) s_{i-x(t-\tau)}\right]. \tag{S76}$$

Here we assume linear rectification:

$$\phi(\lambda) = \begin{cases} \lambda & , \quad \lambda > \lambda_c \\ \lambda_c & , \quad \lambda < \lambda_c \end{cases} \tag{S77}$$

where λ_c is a (typically very small) cutoff firing rate. The precise treatment of rectification within the factorized approach leads to complicated update rules. Instead, we use an approximation, which reduces to the precise update rules derived earlier when there is no rectification.

To explain the approximation we consider the update rule between spikes. To derive the update rule for $p_k(s_k; t)$ we need to calculate the following quantity,

$$A_k = p_k(s_k) \prod_{j \neq k} \left(\sum_{s_j} p_j(s_j) \right) \sum_i \phi\left[\lambda_0 + \Delta\lambda \int \mathrm{d}\tau\, f(\tau) s_{i-x(t-\tau)}\right] \tag{S78}$$

The derivative of $p_k(s_k; t)$ between spikes can then be written in terms of A_k as

$$\frac{\mathrm{d}}{\mathrm{d}t} p_k(s_k; t) = -A_k + \sum_j A_j p_k(s_k; t) \tag{S79}$$

The sum over i is the total firing rate from the whole population of RGCs. Due to the nonlinearity it is difficult to calculate precisely the sum over s_j. Our approximation is to replace, for each i, the argument inside ϕ by an estimate based on the expected firing rate,

$$\phi\left[\lambda_0 + \Delta\lambda \int \mathrm{d}\tau\, f(\tau) s_{i-x(t-\tau)}\right] \simeq \Theta_i \times \left[\lambda_0 + \Delta\lambda \int \mathrm{d}\tau\, f(\tau) s_{i-x(t-\tau)}\right] \tag{S80}$$

where

$$\Theta_i = \Theta\left[\lambda_0 + \Delta\lambda \int \mathrm{d}\tau\, f(\tau) m_{i-x(t-\tau)} - \lambda_c\right] \tag{S81}$$

and Θ is the Heaviside function. In other words, the decoder estimates for each RGC whether its output is rectified, based on its current estimate of the pixels. After making this approximation, it is straightforward to evaluate A_k and in the binary case we get

$$\frac{\partial m_k(t)}{\partial t} = -\Delta\lambda m_k(t) \left[1 - m_k(t)\right] \sum_x \tilde{P}(x; t) \Theta_{x+k} \tag{S82}$$

where we use the notation $m_k(t) = P_k(s_k = 1; t)$. A similar procedure yields an approximation rule for the update following a spike in RGC i,

$$m_k(t_+) = \frac{q_1}{[1 - m_k(t_-)]\, q_0 + q_1 m_k(t_-)} m_k(t_-) \tag{S83}$$

where

$$q_1 = \min\left\{ R_k + \Delta\lambda \tilde{P}(i-k)\,[1 - m_k(t_-)]\,, \lambda_c \right\}, \tag{S84}$$

$$q_0 = \min\left\{ R_k - \Delta\lambda \tilde{P}(i-k) m_k(t_-), \lambda_c \right\} \tag{S85}$$

and where R_k is given by Eq. (S70).

Comparison with the ideal decoder

In the case of a known trajectory [Eq. (S75)] and for a very small image (4 x 4 pixels) we can compare performance of this decoder with the ideal Bayesian decoder, Fig. S3 B. The factorized decoder in this case matches almost precisely the ideal Bayesian decoder, and therefore provides an estimated upper bound for performance in the case of an unknown trajectory. Further, we expect performance for a known trajectory to depend only weakly on image size. This expectation is confirmed by comparing Fig. 3S B (red trace) with Fig. 3 A (dashed trace).

C. Unknown trajectory - factorized decoder with trajectory filtering

In the full problem where the decoder jointly estimates the trajectory and the filter, we considered an approximate scheme, which we call the factorized decoder with trajectory filtering. The decoder estimates the position of the image using the naive rules of Eqs. (S11) and (S14). Even though the naive decoder ignores the temporal filter, it tracks the position of the image, with a small delay $\delta t \simeq 15\,\text{ms}$ that matches the peak time of $f(\tau)$, Fig. S3 C. The decoder then generates an estimate of $\tilde{P}(x,t)$ as follows,

$$\tilde{P}(x;t) = \int \mathrm{d}\tau\, f(\tau) P(x; t-\tau) \tag{S86}$$

This estimate is used to update the pixel estimates $m_i(t)$ using Eqs. (S67) and (S73) using Eqs. (S82) and (S83). The network architecture that could implement this decoding strategy is shown schematically in Fig S3 D. Because the estimate of $\tilde{P}(x;t)$ is delayed by δt, we introduce a compensating delay in the spikes when updating $P_i(s_i;t)$. Therefore the process of spike estimation

starts only after a delay δt. In order to improve the trajectory estimate during the initial δt period, we update $P_i(s_i)$ during this period using the naive rules, Eqs. (S12) and (S17). After the initial δt period, pixel estimation starts anew using Eqs. (S82) and (S83).

The factorized decoder with trajectory filtering performs significantly better than the naive factorized decoder that ignores temporal filtering altogether, as demonstrated in Fig. S3E.

IV. PIECEWISE STATIC DECODER

The piecewise static decoder (Fig. 4C, gray trace) is defined as follows. Time is split into intervals of duration T. The spikes emitted in each one of these time intervals are analyzed separately to generate a likelihood estimate for each of the patterns s^α (the 26 letters). This estimate is given by

$$p_\alpha(t) = \frac{1}{Z(t)} \sum_x \prod_i p\left[r_i(t) \mid s^\alpha_{i+x}\right] \tag{S87}$$

where s^α_i is the intensity of pixel i in pattern α, and the sum is over all possible translations of the pattern. The decoder assumes that within a time interval the position of the image is static, and all possible locations (represented in the sum by x) are equally likely. The spike count statistics are Poisson,

$$p\left[r \mid s\right] = \frac{\exp[-\lambda(s)T]\left[\lambda(s)T\right]^r}{r!} \tag{S88}$$

Finally, the decoder treats positions in different time intervals as if they are independent. Hence the likelihood for each pattern is given by:

$$\log P_\alpha = \sum_t \log p_\alpha(t). \tag{S89}$$

In a discrimination task, only the relative magnitude of P_α for different α is important. Therefore it is sufficient for the decoder to evaluate the following quantities

$$L_\alpha = \sum_t \log\left\{\sum_x \exp\left[-a_\alpha + \sum_i r_i(t) \log\lambda\left(s^\alpha_{i+x}\right)\right]\right\} \tag{S90}$$

where

$$a_\alpha = \sum_i \lambda(s^\alpha_i)T \tag{S91}$$

which represent the log likelihood up to an additive constant which is independent of α, and to choose the pattern α for which L_α is maximal.

V. ONLINE METHODS

Initialization

At time $t = 0$ all estimates in the *what* population are set to $m_i(0) = 0.5$, *i.e.* to implement a prior that *on* and *off* occur with equal probabilities. We arbitrarily label the position of the image at time 0 as $x(0) = 0$. Therefore we set $p(x = 0, t = 0) = 1$. For all other x, $p(x, t = 0) = 0$.

Accuracy measurements

The representation in *what* cells is stabilized in time, but in different trials with the same image it may converge at different spatial shifts. To accommodate for these shifts when measuring accuracy, we first find the shift x_m such that $p\left(s \mid \{m_{i+x_m}\}\right)$ is maximized, where s is the true image and $p\left(s|\{m\}\right) \equiv \prod_i \left[s_i m_i + (1 - s_i)(1 - m_i)\right]$. To measure accuracy we compare the maximum-likelihood pattern (obtained by rectifying m_i) with the image s at the shift x_m.

Resolution of reconstruction

For reconstruction of images composed of pixels subtending 0.5 arcmins, a diffusion coefficient $D = 100\,\mathrm{arcmin}^2/\mathrm{s}$ corresponds to $400\,\mathrm{pixels}^2/\mathrm{s}$. To estimate performance on reconstruction of pixels spanning 1 arcmin, we modified our simulations in two ways: First, because four RGCs are available to report on the value of each pixel, we increased firing rates by a factor of 4. Second, we decreased the diffusion coefficient, in units of $\mathrm{pixels}^2/\mathrm{s}$, by the same factor.

Temporal filter

In all numerical simulations with a temporal filter, we used a biphasic kernel of the form [15,19]

$$f(t) = \frac{t^n}{\tau_1^{n+1}} \mathrm{e}^{-t/\tau_1} - \rho \frac{t^n}{\tau_2^{n+1}} \mathrm{e}^{-t/\tau_2} \tag{S92}$$

where $\tau_1 = 5\,\mathrm{ms}$, $\tau_2 = 15\,\mathrm{ms}$, $n = 3$, and $\rho = 0.8$.

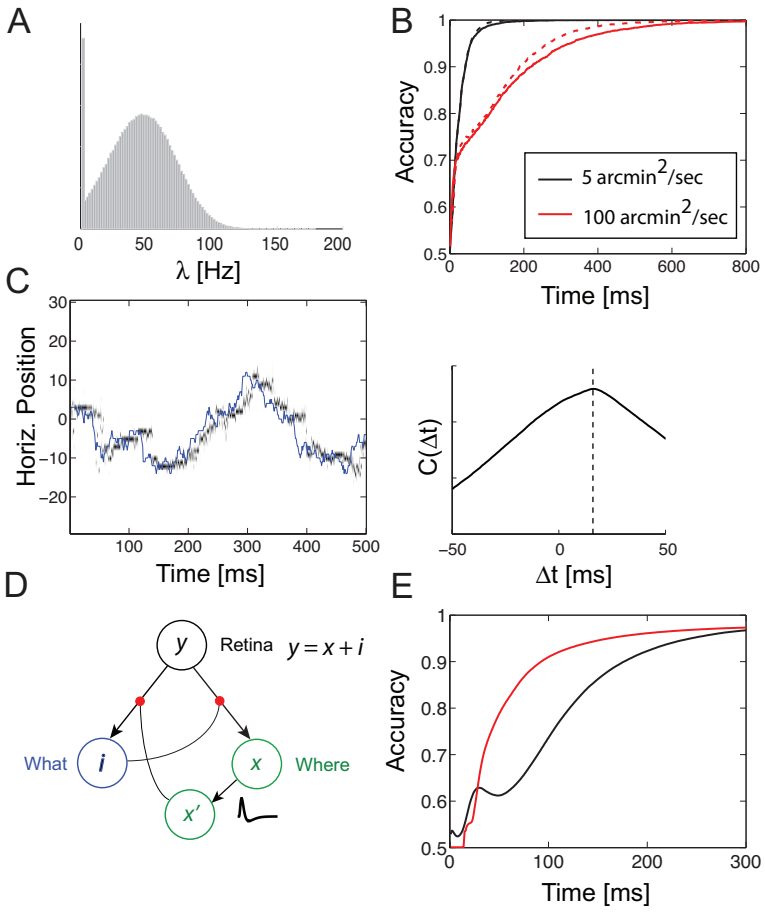

FIG. S3: **A** Distribution of firing rates [Eq. (6) in Methods] measured over a long presentation of an image containing 10 x 10 pixels and averaged over all RGCs. The diffusion coefficient $D = 100\,\mathrm{arcmin}^2/\mathrm{s}$. In all panels in this figure, $\lambda_0 = 20$ and $\Delta\lambda$ is set such that the maximum possible firing rate is 200 Hz. **B** For a known trajectory and spikes generated with a temporal filter we compare performance of the ideal Bayesian filter (dashed traces) and the factorized decoder of Eqs. (S82)–(S85) with known trajectory, Eq. (S75), which takes into account the structure of the temporal filter (solid traces). The full Bayesian decoder can only be implemented for very small images, hence the image contains only 4 x 4 pixels. Results are shown for two values of the diffusion coefficient (legend). The resolution is 0.5 arcmin. **C** Tracking of the image position by the naive factorized decoder which assumes that RGC response is instantaneous, when presented with spikes generated from RGCs with a non-instantaneous response. Left: The true position (blue trace), and tracking by the *where* cells: grayscale intensities represent the inferred position, marginalized over the vertical axis. Right: correlation function of the true position and the mean estimated position. Tracking lags behind the true position by about 16.5 ms (vertical dashed line). This lag corresponds approximately to the sharp peak in the temporal filter (Fig. 4A, inset). **D** Schematic architecture of a neural network that implements the factorized decoder with trajectory filtering (Supporting Text). **E** Performance of the factorized decoder with trajectory filtering (red trace), compared to the naive factorized decoder (black trace, as in Fig. 4C). Parameters: 30 x 30 arcmin image, 1 arcmin resolution, $D = 100\,\mathrm{arcmin}^2/\mathrm{s}$.