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.

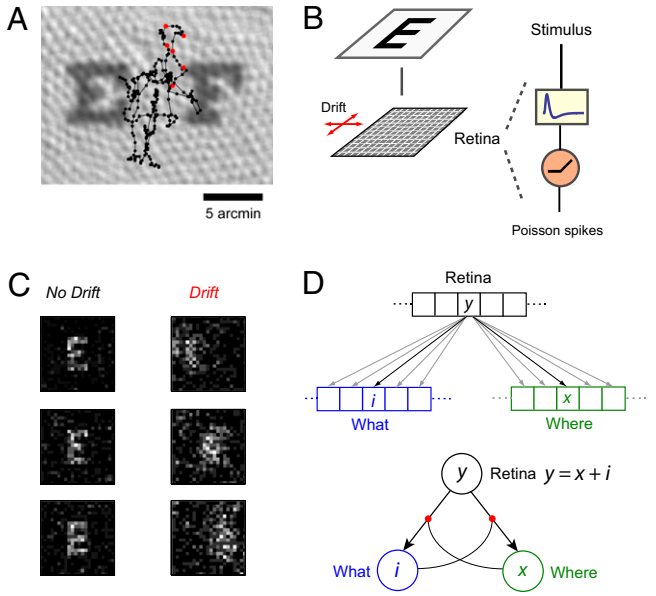

Fig. 1. (*A*) The letters E and F on the 20/20 line of the Snellen eye chart test, projected on an image of the foveal cone mosaic (photoreceptor image modified from ref. 39). The 1-arcmin features that distinguish the letters extend over only a few cones. Also shown is a sample fixational eye movement 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-

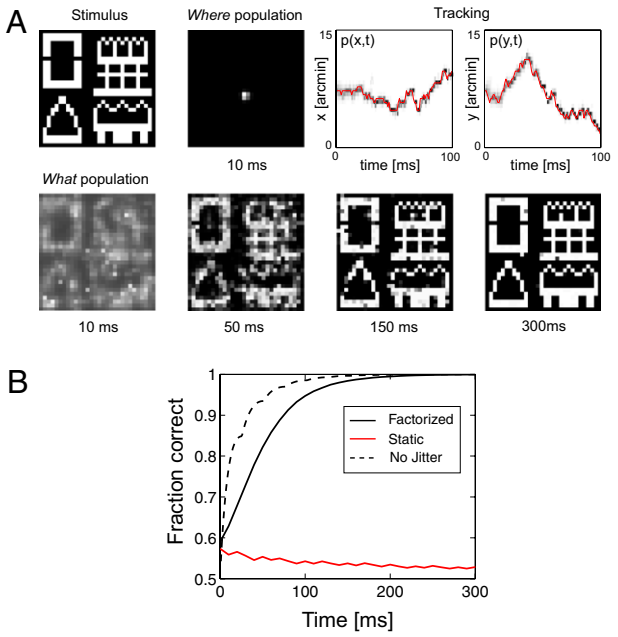

Fig. 2. (*A*) Example of image reconstruction by the factorized decoder. (*Upper*) From left to right: the stimulus; snapshot of activity in the *where* cell population at t = 10 ms; and tracking of horizontal and vertical image position over time, with probability (grayscale) compared with actual trajectory (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).

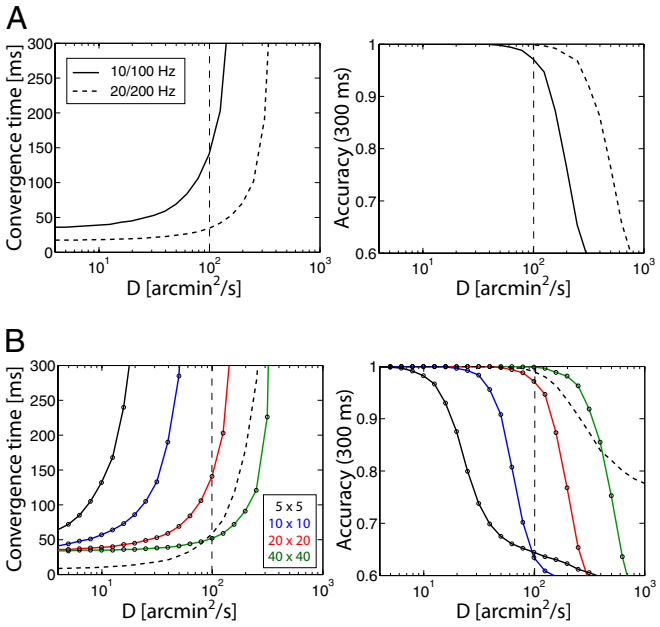

Fig. 3. (*A*) Performance as a function of D, averaged over 1,000 presentations of random images. The convergence time (at which 90% of pixels are estimated correctly) increases with D (*Left*) and the accuracy (fraction of correctly estimated pixels at t = 300 ms) decreases with D (*Right*). Results 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*).

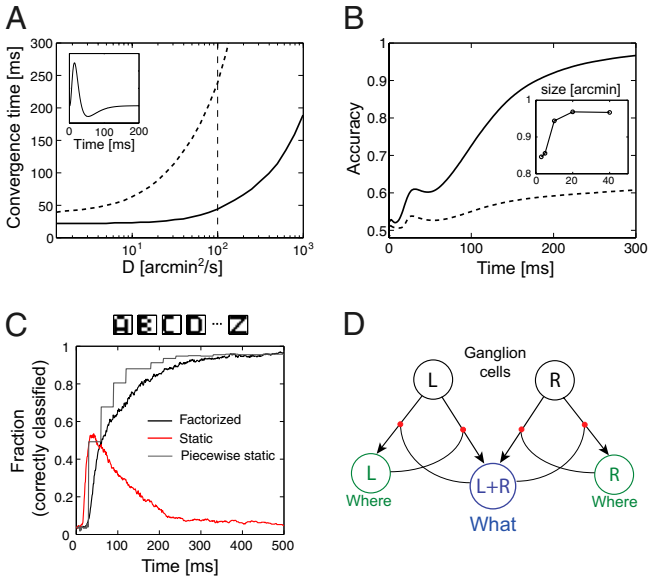

Fig. 4. Performance for spike trains generated with a temporal filter in RGC response. (*A*) Convergence time when the trajectory is known to the decoder. In contrast to the case of instantaneous response, performance depends on the diffusion statistics. Traces show the convergence time (for 90% accuracy), 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.