NOTE Communicated by Geoffrey Hinton

Optimal Smoothing in Visual Motion Perception

Rajesh P. N. Rao
Department of Computer Science & Engineering, University of Washington, Seattle, WA 98195, U.S.A.

David M. Eagleman
Sloan Center for Theoretical Neurobiology, Salk Institute for Biological Studies, La Jolla, CA 92037, U.S.A.

Terrence J. Sejnowski
Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92037, U.S.A., and Department of Biology, University of California at San Diego, La Jolla, CA 92037, U.S.A.

When a flash is aligned with a moving object, subjects perceive the flash to lag behind the moving object. Two different models have been proposed to explain this "flash-lag" effect. In the motion extrapolation model, the visual system extrapolates the location of the moving object to counteract neural propagation delays, whereas in the latency difference model, it is hypothesized that moving objects are processed and perceived more quickly than flashed objects. However, recent psychophysical experiments suggest that neither of these interpretations is feasible (Eagleman & Sejnowski, 2000a, 2000b, 2000c), hypothesizing instead that the visual system uses data from the future of an event before committing to an interpretation. We formalize this idea in terms of the statistical framework of optimal smoothing and show that a model based on smoothing accounts for the shape of psychometric curves from a flash-lag experiment involving random reversals of motion direction. The smoothing model demonstrates how the visual system may enhance perceptual accuracy by relying not only on data from the past but also on data collected from the immediate future of an event.

1 Introduction

When subjects are presented with a flash that is aligned with a moving object, they perceive the flash to lag behind the moving object (MacKay, 1958; Nijhawan, 1994) (see Figure 1a). In order for subjects to perceive the flash as aligned with the moving object, the flash must be presented at a spatial location ahead of the moving object. A recent flurry of experiments has sparked interest in models that can explain this phenomenon (Ni-

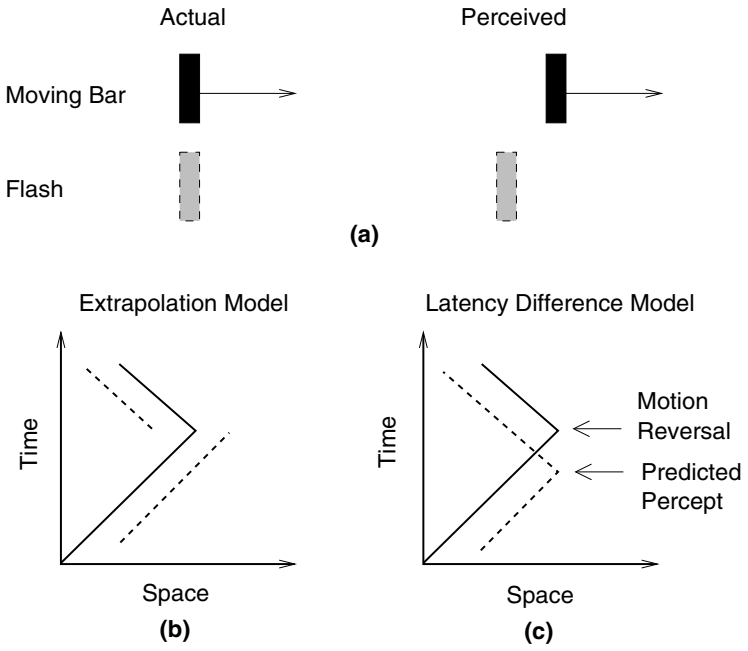

Figure 1: Flash-lag effect and predictions of two previous models. (a) Flash-lag effect. The moving bar (black) is perceived to be ahead of the flashed bar (shaded) even though the retinal images are physically aligned. (b) Actual (solid line) and perceived (dashed line) positions of the moving bar, as predicted by the motion extrapolation model for the motion reversal experiment. (c) Actual (solid line) and perceived (dashed line) positions, as predicted by the latency difference model.

jhawan, 1994; Baldo & Klein, 1995; Khurana & Nijhawan, 1995; Nijhawan, 1997; Lappe & Krekelberg, 1998; Purushothaman, Patel, Bedell, & Ogmen, 1998; Whitney & Murakami, 1998; Whitney, Murakami, & Cavanagh, 2000; Krekelberg & Lappe, 2000; Brenner & Smeets, 2000; Eagleman & Sejnowski, 2000a, 2000b, 2000c).

The motion extrapolation model (see Figure 1b) (Nijhawan, 1994) assumes that the visual system extrapolates the location of moving objects to compensate for propagation delays as signals are transmitted from the retina to higher cortical areas. If left uncompensated, such delays would cause the perceived location of the moving object to lag significantly behind the actual location. Nijhawan suggested that extrapolation allows objects to be perceived at their actual location. In the latency difference model (Baldo & Klein, 1995; Purushothaman et al., 1998; Whitney & Murakami, 1998; Whitney et al., 2000), it is hypothesized that the moving object is perceived to

be ahead of the flash due to shorter neural propagation delays for moving objects as compared to flashed objects (see Figure 1c).

Recent experiments have revealed the shortcomings of both these models (Eagleman & Sejnowski, 2000a, 2000b, 2000c). In particular, both the extrapolation model and the latency difference model fail to provide a complete explanation for the psychometric curves obtained when the direction of motion of the moving object is abruptly reversed. We show that a model based on the engineering technique of optimal smoothing (Bryson & Ho, 1975) overcomes the limitations of both these models. In the smoothing model, perception of an event is not online but rather is delayed, so that the visual system can take into account information from the immediate future before committing to an interpretation of the event.

2 Motion Reversal Experiments

In an experiment designed to test the motion extrapolation model, Whitney and Murakami (1998) reversed the motion of a horizontally translating bar at a random time and location along its trajectory. A flash could appear at various times before or after motion reversal. The study tested where the flash needed to be placed in order to be perceived as aligned with the moving bar. According to the extrapolation model, at the point of motion reversal, the moving bar should be perceived at its extrapolated location as depicted in Figure 1b (recall that the time of reversal is random and unknown to the subject). Contrary to this prediction, Whitney and Murakami reported that the perceived position of the moving bar never overshot the reversal point. Rather, the perceived location of the moving bar began deviating significantly from that predicted by the motion extrapolation model at approximately 60 to 75 milliseconds before the time of reversal. If extrapolation were indeed occurring, the bar's reversal must have been known before the actual reversal took place, an impossibility. Whitney and Murakami therefore concluded that their results supported the latency difference model.

However, the latency difference model cannot by itself explain the rounding of the curve observed near the time of reversal, predicting instead a sharp reversal in the perceived location, as shown in Figure 1c. Whitney and Murakami (1998) suggested that the rounding may be due to neural delay variability or a spatiotemporal averaging filter, but other experiments have revealed more serious flaws in the latency difference model (Eagleman & Sejnowski, 2000a, 2000b, 2000c). For example, the flash-lag effect is preserved in the case where the bar starts moving at the same time t_0 as the flash that is aligned with it. In this "flash-initiated" paradigm (Khurana & Nijhawan, 1995; Eagleman & Sejnowski, 2000a), there is no past history of bar motion at time t_0 for a spatiotemporal filter to operate over. The moving bar should suffer the same initial processing delay as the flashed stimulus: how could it still be perceived ahead of the flash? This suggests that the visual system is using motion information occurring after time t_0 to make a

judgment about the perceived location at time t_0, in effect using information from the immediate future to estimate a quantity in the recent past, a form of "postdiction." This interpretation has recently been proposed by Eagleman and Sejnowski (2000a), who show that their psychophysical results are best explained by the postdiction hypothesis. We show here that this hypothesis can be framed succinctly within the statistical framework of optimal smoothing (Bryson & Ho, 1975).

3 The Optimal Smoothing Model

The strategy of estimating a value in a time series based on future values (in addition to past values) is known as smoothing in the engineering literature (Bryson & Ho, 1975). On the other hand, estimating a current value based only on past values is called filtering, e.g., Kalman filtering (see Kalman, 1960). We have simulated the experiments of Whitney and Murakami using a simple dynamical model describing the linear motion and reversal of the bar in the presence of gaussian noise:

$$x(t+1) = x(t) + c(t)y(t) + n(t) \tag{3.1}$$

where $x(t)$ denotes the position of the bar at time t, $c(t)y(t)$ is the increment or decrement in position for the next time step ($c(t) = +1$ initially, switching to -1 at a random time of reversal), and $n(t)$ is zero-mean gaussian noise with variance σ^2. The increment amount $y(t)$ is assumed to be constant except for additive zero-mean gaussian noise: $y(t+1) = y(t) + w(t)$. Finally, the position x is assumed to be corrupted by measurement noise $m(t)$ before being observed by the subject: $z(t) = x(t) + m(t)$, where m is again a gaussian noise process with zero-mean and variance σ_m^2.

An optimal linear filter (the Kalman filter; Kalman, 1960) was derived from the motion model above to estimate the most likely position $\widehat{x}$ of the bar at time t given information about the current and past positions of the moving bar (see Bryson & Ho, 1975, for a derivation):

$$\widehat{x}(t) = \overline{x}(t) + g(t)(z(t) - \overline{x}(t)) \tag{3.2}$$

$$\overline{x}(t) = \widehat{x}(t-1) + c(t-1)\widehat{y}(t-1) \tag{3.3}$$

where $g(t)$ is a gain term (see Bryson & Ho, 1975) and $\widehat{y}(t-1) = \overline{y}(0) = a$ (a determines the velocity of the bar, assumed to be constant in this case).

Equations 3.2 and 3.3 can be explained as follows. At any given time t, the filter maintains an estimate $\overline{x}(t)$ of bar position x *before* a new measurement $z(t)$ is obtained. This estimate is our best estimate of position using all previous measurements $z(t-1), \ldots, z(0)$ and the motion model in equation 3.1. Note that $\overline{x}(t)$ is computed from $\widehat{x}(t-1)$, which in turn is computed from $\overline{x}(t-1)$. Once the measurement $z(t)$ is obtained, the filter computes a new estimate $\widehat{x}(t)$ by correcting the old estimate $\overline{x}(t)$ using the mismatch

error $(z(t) - \overline{x}(t))$. Thus, $\widehat{x}(t)$ represents our best estimate of bar position *after* measuring $z(t)$.

The filter estimate for time t was smoothed recursively using the estimates from the next time steps. This increases accuracy by allowing data from time steps in the future of t to influence and possibly correct the filter estimate at time t (Bryson & Ho, 1975):

$$x_{sm}(t) = \widehat{x}(t) + h(t)(x_{sm}(t+1) - \overline{x}(t+1)) \tag{3.4}$$

where $x_{sm}(t)$ is the smoothed estimate for time t given position information from time steps $1, \ldots, N$ $(N > t)$, and $h(t)$ is a gain term (see Bryson & Ho, 1975). Note that since $x_{sm}(t)$ depends not only on $\widehat{x}(t)$ but also on $x_{sm}(t+1)$, which in turn depends on $x_{sm}(t+2)$ and so on, the smoothed estimate at time t relies not only on measurements from the past but also on measurements from future time steps relative to t. Smoothing corrects each position estimate $\widehat{x}(t)$ by adding the error term $h(t)(x_{sm}(t+1) - \overline{x}(t+1))$, which represents the mismatch between the smoothed and the filtered estimates at time $t+1$.

The smoothing model described above can be used to account quantitatively for the flash-lag results involving motion reversal. Before doing so, it is useful to make a distinction between the following three times associated with an event: (1) event time, which is the time at which an event occurs in the real world; (2) neural activity time, which is the time at which a representation of the event is formed at a particular neural level; and (3) represented time (or subjective time) of the occurrence of the event. To see how these three times may differ, consider the case of recalling visual memories, say, of an event that occurred during college and an event that occurred in childhood. Clearly the event times are different for these two events, as are the represented times, both of which have a temporal order (childhood events before college events). However, the neural activity time which is the time of recall of these memories, does not need to follow this temporal order. For the flash-lag effect, the latency difference model assumes that the neural activity time is the same as the represented time and that the neural activity time for the flash is later than that for the moving bar. The extrapolation model, on the other hand, assumes that for the moving bar, neural delays can be counteracted such that the represented time of the moving bar is equal to its event time.

The smoothing model, in contrast, is illustrated in Figure 2. Suppose that the event time of the flash is t_0. We assume that the neural activity time of the flash is $t_0 + \Delta$, where Δ is the neural propagation delay. Then, according to the smoothing model, the subject's perceived location of the bar at the time of the flash is given by the smoothed estimate $x_{sm}(t_0 + \Delta)$. Note that this estimate includes information from time steps up to $t_0 + \Delta + f$, where f is the amount of time in the future used for smoothing. Thus, the estimate of an event that happened at time t_0 is retrospectively assigned after a minimum duration of $\Delta + f$. The flash-lag effect occurs because the subject reports the

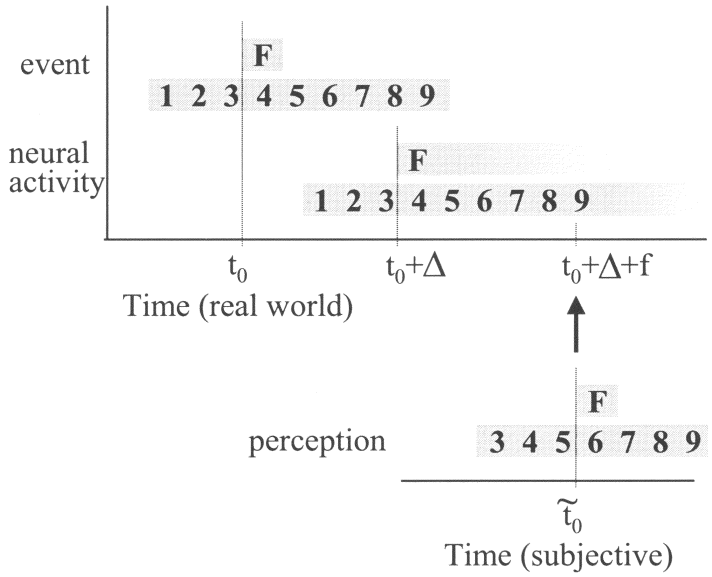

Figure 2: Event time, neural activity time, and represented time. F represents the flash; the strip of numbers represents successive positions of the moving object. In this example, the flash occurs at time t_0 (event time) when the moving object occupies position 4. After a neural propagation delay of Δ, neural activity pertaining to the flash begins at a particular level of the visual system ("neural activity" on the ordinate). After processing of further information, the results of the smoothing filter become available to consciousness at time $t_0 + \Delta + f$. The represented time cannot be displayed on the same axis (real-world time); instead, it must be displayed on its own axis (subjective-world time). All studies of the flash-lag effect measure only the relative timing between flashed and moving objects, informing us in no way about real-world time. In the figure, $\tilde{t}_0$ is the perceived moment of the flash which occurred at time t_0 (the graphs are offset because the represented time $\tilde{t}_0$ cannot exist until smoothing is complete). In subjective time, the flash is aligned with position 6 (the smoothed position estimate for time $t_0 + \Delta$). This misalignment is the flash-lag illusion. The absence of positions 1 and 2 in the perception represents the Frohlich effect, in which the initial positions of a moving object are not perceived.

location of the moving bar to be the smoothed estimate at time $t_0 + \Delta$ (see Figure 2).

For the simulations, the following parameter values were used: $g(t) = 0.7$, $h(t) = 0.5$, $\sigma = 0.01$, $\sigma_m = 0.01$, $a = 1$, $\overline{x} = 0$, $N = 50$, $x_{sm}(N) = \widehat{x}(N)$, and $\Delta = 45$ milliseconds (two time steps in the simulations). Similar results were obtained when parameter values, such as the noise variances, were varied in the neighborhood of the values given above. The gain terms $g(t)$

and $h(t)$ can be made a time-varying function of the variances σ and σ_m (Bryson & Ho, 1975), but in the simulations, constant values, such as the ones specified above, were found to be sufficient for modeling the psychophysical results. (Matlab code for running these simulations is available online at http://www.cnl.salk.edu/~rao/smoothing.m.)

4 Results

Figures 3a and 3b show the perceived location of the moving bar for a human subject (data points) and for the optimal smoothing model (data points labeled x_{sm}), respectively. The perceived location estimated by the optimal filtering model ($\overline{x}$) is given by the dotted line. The data shown were averaged over 100 trials with a single random reversal of motion in each trial.

As seen in the figure, the smoothing model reproduces the rounding of the curve observed in human subjects (see Figure 3a), while the filtering model, which uses only past positions of the bar, overshoots at the point of reversal before correcting its estimate at subsequent time steps. This overshoot is avoided by the smoothing model because data from the immediate future are taken into account, producing a more accurate estimate of bar position.

How many data points from the future are taken into account in the model? To answer this question, we computed the impulse response functions of the filter and the smoother that were used in the simulations. As seen in Figures 3c and 3d, both the filter and smoother use input data from the current and approximately four previous time steps. However, the smoother also takes into account data from about four to five time steps into the future. In the model, this corresponds to a time interval of approximately 90 to 112 milliseconds (one time step $\approx$ 22.5 milliseconds), which is in the range of the time window of approximately 80 milliseconds reported by Eagleman and Sejnowski (2000a).

5 Discussion

Why should the visual system delay its perception of an event to integrate information from the future? The smoothing model suggests that this is done in order to enhance perceptual accuracy in the presence of uncertainty and noise. It has long been known in the engineering community (see, e.g., Bryson & Ho, 1975) that the limitations of filtering (signal estimation based on the past) can be overcome by smoothing techniques that take some or all future data in a time series into account for optimal estimation of signal properties. Our results suggest that the visual system may be employing this strategy for accurate estimation of visual motion. An additional advantage of smoothing is that smoothed estimates make learning more reliable. For example, in the case of hidden Markov models (HMMs),

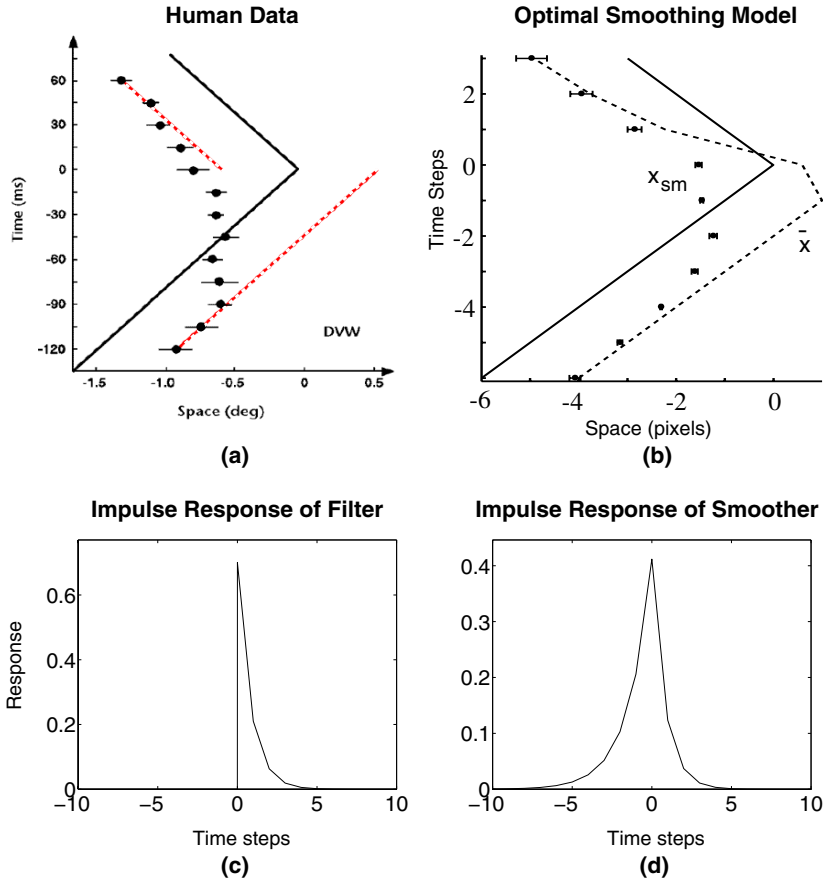

Figure 3: Flash-lag effect interpreted as optimal smoothing of visual motion estimates. (a) Data from a human subject showing the perceived location of a moving bar, as revealed by aligning the flash (reproduced from Whitney & Murakami, 1998). Lines through data points are 95% confidence intervals. Dotted line = prediction of the extrapolation model. (b) Data from the optimal smoothing model, where the perceived location is taken to be the smoothed position estimate x_{sm}. Lines through data points are one standard deviation above and below average values computed over 100 trials. The filtered estimate $\bar{x}$ is shown as a dotted line for comparison. Note the overshoot at the time of reversal for the filtered but not the smoothed estimate. (c) Impulse response of the optimal linear filter used in the simulations. (d) Impulse response of the optimal smoother used in the simulations. Note that the impulse response function for the smoother includes weights for past, current, and future data, whereas the impulse response for the filter considers only current and past data.

both the forward (filtering) and the backward (smoothing) procedures are used for computing the likelihood in the Baum-Welch algorithm for learning model parameters (Rabiner & Juang, 1993; see also Dayan & Hinton, 1996). Filtering and smoothing are similarly used in algorithms for learning the parameters of continuous-state linear dynamical systems (Shumway & Stoffer, 1982; Ghahramani & Hinton, 1996).

Given the natural trade-off between the amount of perceptual delay required for smoothing and the need for real-time computation, an interesting open question is whether the delay of 80 to 100 milliseconds inferred from psychophysical experiments (Eagleman & Sejnowski, 2000a) represents an optimal balance between perceptual accuracy and real-time inference. A related question is whether this delay can be adapted according to the task at hand. These questions remain the subject of ongoing investigations. The model presented here also assumes that the subject possesses a model of the moving stimulus as given by equation 3.1. Such a model could have been acquired as a result of prior experience with moving stimuli and fine-tuned during training before collection of data or, alternately, could have been learned directly during training. The latter possibility is supported by several algorithms that have recently been suggested for learning the parameters of linear dynamical systems directly from input data (Shumway & Stoffer, 1982; Ghahramani & Hinton, 1996; Rao & Ballard, 1997).

An interesting question is whether the smoothing model can predict the effect of varying the luminance of the flash and the moving bar. We expect such an experimental manipulation to change the signal-to-noise ratio in the input channels and, hence, the gain terms $g(t)$ and $h(t)$ in the filter and smoother, respectively, thereby changing the shape of their impulse response functions (see Figures 3c and 3d). This could result in a flash-lead effect under some circumstances, as observed experimentally (Purushothaman et al., 1998).

It is known that the flash-lag effect is reduced when the flash becomes more predictable (Eagleman & Sejnowski, 2000b). For the simulations reported here, we used a minimal internal model for the flash: the flash is detected by the subject after some amount of processing delay. A more general approach is to use a dynamical model for the flash in addition to the dynamical model for the moving object. Such an extended model would allow smoothed estimates of both the moving and flashed bars to be computed; the smoothed position of the flashed bar would then be compared to the smoothed position of the moving bar. In the case of multiple predictable flashes (e.g., stroboscopically moving flashes), such a model would be expected to produce a reduction in flash lag (due to smoothing of the flashed bars), in accordance with previous experimental findings (Lappe & Krekelberg, 1998; Eagleman & Sejnowski, 2000b). Testing this hypothesis remains an interesting direction for future research.

The idea that the visual system performs statistical or Bayesian inference based on its inputs has recently been proposed by several research groups

(e.g., Freeman, 1994; Hinton, Ghahramani, & Teh, 2000; Knill & Richards, 1996; Rao, 1999). Our results support this emerging model of visual perception and show how the visual system may base its inference about a particular event not only on past observations but also on observations from the immediate future. Such a model extends previous models of the visual cortex based on optimal filtering theory (Mumford, 1994; Rao & Ballard, 1997; Rao, 1999). It may additionally allow novel interpretations of other well-known visual phenomena, such as backward masking (Bachmann, 1994) and the color phi effect (Kolers & von Grunau, 1976), involving the effect of future stimuli on the perception of a preceding stimulus.

Acknowledgments

This research was supported by the Sloan Center for Theoretical Neurobiology at the Salk Institute and the Howard Hughes Medical Institute. We thank Geoffrey Hinton for providing the impetus to this work by pointing out the connections between smoothing and the flash-lag effect, and the reviewers for their helpful comments and suggestions that led to Figures 2, 3c, and 3d.

References

Bachmann, T. (1994). *Psychophysiology of visual masking*. Commack, NY: Nova Science Publishers.

Baldo, M. V., & Klein, S. A. (1995). Extrapolation or attention shift? *Nature*, *378*, 565–566.

Brenner, E., & Smeets, J. B. (2000). Motion extrapolation is not responsible for the flash-lag effect. *Vision Research*, *40*(13), 1645–1648.

Bryson, A. E., & Ho, Y.-C. (1975). *Applied optimal control*. New York: Wiley.

Dayan, P., & Hinton, G. E. (1996). Varieties of Helmholtz machine. *Neural Networks*, *9*(8), 1385–1403.

Eagleman, D. M., & Sejnowski, T. J. (2000a). Motion integration and postdiction in visual awareness. *Science*, *287*, 2036–2038.

Eagleman, D. M., & Sejnowski, T. J. (2000b). Response: The position of moving objects. *Science 289*, 1107a. Available online at: http://www.sciencemag.org/cgi/content/full/289/5482/1107a.

Eagleman, D. M., & Sejnowski, T. J. (2000c). Response: Latency difference, not postdiction. *Science, 290*, 1051a.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, 368, 542–545.

Ghahramani, Z., & Hinton, G. E. (1996). *Parameter estimation for linear dynamical systems* (Tech. Rep. No. CRG-TR-96-2). Toronto: Department of Computer Science, University of Toronto.

Hinton, G. E., Ghahramani, Z., & Teh Y. W. (2000). Learning to parse images. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 463–469). Cambridge, MA: MIT Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction theory. *Trans. ASME J. Basic Eng.*, *82*, 35–45.
Khurana, B., & Nijhawan, R. (1995). Extrapolation or attention shift? *Nature*, *378*, 565–566.
Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.
Kolers, P., & von Grunau, M. (1976). Shape and color in apparent motion. *Vision Research*, *16*, 329–335.
Krekelberg, B., & Lappe, M. (2000). A model of the perceived relative positions of moving objects based upon a slow averaging process. *Vision Research*, *40*(2), 201–215.
Lappe, M., & Krekelberg, B. (1998). The position of moving objects. *Perception 27*(12), 1437–1449.
MacKay, D. M. (1958). Perceptual stability of a stroboscopically lit visual field containing self-luminous objects. *Nature*, *181*, 507–508.
Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 125–152). Cambridge, MA: MIT Press.
Nijhawan, R. (1994). Motion extrapolation in catching. *Nature*, *370*, 256–257.
Nijhawan, R. (1997). Visual decomposition of colour through motion extrapolation. *Nature*, *386*, 66–69.
Purushothaman, G., Patel, S. S., Bedell, H. E., & Ogmen, H. (1998). Moving ahead through differential visual latency. *Nature*, *396*, 424.
Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.
Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, *39*, 1963–1989.
Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, *9*, 721–763.
Shumway, R. H., & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, *3*, 253–264.
Whitney, D., & Murakami, I. (1998). Latency difference, not spatial extrapolation. *Nature Neuroscience*, *1*, 656–657.
Whitney, D., Murakami, I., & Cavanagh, P. (2000). Illusory spatial offset of a flash relative to a moving stimulus is caused by differential latencies for moving and flashed stimuli. *Vision Research*, *40*, 137–149.

Received February 25, 2000; accepted September 25, 2000.