To appear: *The Visual Neurosciences*, Chapter 109
Editors: Leo M. Chalupa and John S. Werner
MIT Press, 2003.

Local Analysis of Visual Motion

Eero P. Simoncelli

We inhabit an ever-changing environment, in which sensing, processing and acting upon these changes can be essential for survival. When we move or when objects in the world move, the visual images projected onto each of our retinas change accordingly. The psychologist J. J. Gibson noted that important environmental information is embedded in this pattern of local retinal image velocities (Gibson, 1950), and thus initiated a scientific quest to understand the mechanisms that might serve to estimate and represent these velocities. Since that time, visual motion perception has been the subject of extensive research in perceptual psychology, visual neurophysiology, and computational theory.

There is an abundance of evidence that biological visual systems – even primitive ones – devote considerable resources to the processing of motion. A substantial proportion of the effort in the field of computer vision has also been devoted to the problem of motion estimation. Although the processing constraints in a biological system are somewhat different from those in an artificial vision system, each must extract motion information from the same type of brightness signal. This chapter adopts the philosophy that in order to understand visual motion processing in the brain, one should understand the nature of the motion information embedded in the visual world, and the fundamental issues that arise when one attempts to extract that information (Marr & Poggio, 1977). I'll develop the most basic computational solution to the motion estimation problem, and examine the qualitative relationship between aspects of this solution and the properties of neurons in the motion pathway.

• Center for Neural Science, and Courant Institute of Mathematical Sciences, New York University, New York, NY 10003
• The author is supported by the Howard Hughes Medical Institute, and the Sloan-Swartz Center for Theoretical Visual Neuroscience at NYU.
• Thanks to J. Pillow and A. Stocker for helpful comments on the manuscript.

Local motion

Images are formed as projections of the three-dimensional world onto a two-dimensional light-sensing surface. This surface could be, for example, a piece of photographic film, an array of light sensors in a television camera, or the photoreceptors in the back of a human eye. At each point on the surface, the image brightness is a measurement of how much light fell on the surface at that spatial position at a particular time (or over some interval of time). When an object in the world moves relative to this projection surface, the two-dimensional projection of that object moves within the image. The movement of the projected position of each point in the world is referred to as the *motion field*.

The estimation of the motion field is generally assumed to be the first goal of motion processing in machine vision systems. There is also evidence that this sort of computation is performed by biological systems. The motion field must be estimated from the spatiotemporal pattern of image brightness. This is usually done by assuming that the brightness generated by points in the world remain constant over time. In this case, the estimated motion of constant-brightness points (known as the *optical flow*) is also an estimate of the motion field. But as many authors have shown, the optical flow is *not* always a good estimate of the motion field (e.g., Horn, 1986; Verri & Poggio, 1989). For example, when a shiny object moves, specular highlights often move across the surface of the object. In this situation, the optical flow (corresponding to the highlight motion) does not correspond to the motion of any point on the object. Nevertheless, estimates of optical flow are almost universally used as approximations of the motion field.

In estimating optical flow, we cannot ask about the motion of an isolated point without considering the con-

text surrounding it. That is, we can only recognize the motion of local patterns of brightness. But our ability to estimate a unique velocity at a given image location depends critically on the structure of the image in the neighborhood of that location. Consider first the simplest situation, in which an object is moving horizontally, perpendicular to the line of sight. Figure 1 depicts three prototypical situations that can arise. First, the local brightness might be constant (over both spatial position and time). In this case, the local measurements places no constraint on the velocity. We will refer to this as the *blank wall* problem.

Second, the local brightness might vary only in one direction – that is, the spatial pattern could be striped. In this case, only the velocity component that is perpendicular to the stripes is constrained. Any component along the stripes will not create a change in the image, and thus cannot be estimated. This is typically known in the literature as the *aperture problem* (Wallach, 1935; Fennema & Thompson, 1979; Marr & Ullman, 1981). The expression refers to the fact that the motion of a moving one-dimensional pattern viewed through a circular aperture is ambiguous. The problem is not really due to the aperture, but arises from the one-dimensionality of the signal.

Finally, the local brightness may vary two-dimensionally, in which case the optic flow vector is uniquely constrained. But because of the occurrence of underconstrained regions (blank wall and aperture problems), a full solution for the motion problem in which all image points are assigned a velocity vector seems to require the integration of information across spatial neighborhoods of the image (and perhaps over time as well). This concept has been studied and developed for many years in computer vision (e.g., Horn & Schunck, 1981; Lucas & Kanade, 1981; Hildreth, 1984).

In addition to the blank wall and aperture problems, in which the velocity is underconstrained, there are often locations in an image at which multiple velocity signals interact. In particular, this can occur at occlusion boundaries of objects, where any local spatial neighborhood must necessarily include some portion of both object and background which are typically moving differently. Another example occurs in the presence of transparently combined surfaces or specular highlights. In each of these cases, the local motion description requires

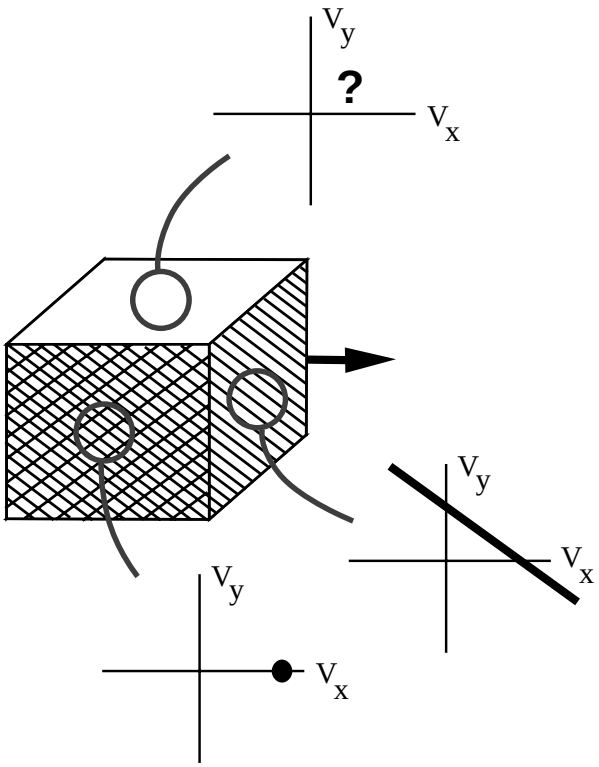

Figure 1. Conceptual illustration of motion estimation in three different regions of an image of a horizontally translating cube. In a region of constant brightness (top face of the cube), the local velocity is completely unconstrained since the observed image is not changing over time. We refer to this as the *blank wall* problem. In a region where the brightness varies only along a unique spatial direction (striped side of the cube), the brightness changes are consistent with a one-dimensional set of velocities: one can determine the motion perpendicular to the stripes, but *not* the motion along the stripes. This is known as the *aperture problem*. Finally, in a region where the brightness changes in all spatial directions (hatched side of the cube), a unique velocity is consistent with the observed brightness changes.

more than one velocity, along with some sort of assignment of which image content belongs to which velocity.

Solutions for the multiple-motion problem have been developed fairly recently. Specifically, a number of authors have proposed that one should simultaneously decompose the image into consistently moving layers of brightness content, and estimate the motions within those layers (Wang & Adelson, 1994; Darrell & Pentland, 1995; Ayer & Sawhney, 1995; Weiss & Adelson, 1996). Some authors have suggested that this sort of solution might be implemented biologically (Darrell & Simoncelli, 1994; Nowlan & Sejnowski, 1995; Koechlin *et al.*, 1999).

We'll return to this point in the Discussion section, but for most of the chapter, we'll restrict our attention to the simple case of translational motion over a local patch of the image, ignoring the possibility of multiple motions. As in much of the computational motion literature, we view this as a building block, that could be combined with further processing to provide a more complete solution for the analysis of motion. The goal of this chapter is to introduce a framework for thinking about motion, and to interpret the components of that framework physiologically.

Computational framework

The problem of motion estimation may be formalized using well-developed tools of estimation theory. Specifically, we adopt a Bayesian framework, one of the simplest and most widely used in the engineering literature. Bayesian approaches have also been used to model various aspects of visual perception (e.g., see Knill & Richards, 1996).

Brightness constancy. In any estimation problem, the most essential ingredient is the relationship between the thing one is trying to estimate and the measurements that one makes. In most motion estimation schemes, this relationship comes from the *brightness constancy assumption* (Horn & Schunck, 1981): changes in image brightness are a result of translational motion in the image plane. When expressed mathematically, this gives a relationship between image values and the local velocity:

$$I(x + u\Delta t, y + w\Delta t, t + \Delta t) = I(x, y, t),$$

where u and w are the horizontal and vertical components of the image velocity, at position (x, y) and time t. Assuming that the temporal interval Δt is small enough that the left side may be approximated by a Taylor series expansion up to first order, we can replace this with a differential version of the brightness constancy assumption:

$$I_x u + I_y w + I_t = 0, \tag{1}$$

where (I_x, I_y) are the spatial derivatives and I_t the temporal derivative of the image brightness. Note that this equation still corresponds to a particular position and moment in time: we've only dropped the arguments (x, y, t) to simplify the notation.

Although the signal I is usually considered to represent the image brightness, the formulation may be generalized. In particular, one can use any function locally derived from the image brightness. For example, one could prefilter the image with a bandpass filter to enhance information at certain frequencies, one could use a point nonlinearity (e.g., logarithm) to reduce the dynamic range of the input, or one could compute a measure of local contrast by dividing by the local mean. Under such modifications, equation (1) should be interpreted as expressing the constancy of some other (non-brightness) attribute. As such, the usual brightness-based motion estimator may be easily converted into a so-called second-order or *non-Fourier* motion estimator, as has been proposed by a number of authors (e.g. Chubb & Sperling, 1988; Wilson & Kim, 1994; Fleet & Langley, 1994).

Furthermore, the computation of derivatives of discretized images requires one to first perform some local integration by prefiltering with a lowpass filter. The differentiation and filtering operations may be combined, so that derivatives are effectively computed by convolution with the derivative of the prefilter (Simoncelli, 1993). In computer vision, for example, derivatives are often computed using the derivatives of a Gaussian function.

Local combination of constraints. Equation (1) cannot be solved for the two components of velocity, since it imposes only a single linear constraint. This is simply a differential manifestation of the aperture problem: the derivative measurements characterize the local brightness values in terms of their variation along a single direction (the gradient direction), and thus only the com-

ponent of velocity in this direction can be estimated. In order to completely estimate the velocity, we must impose further constraints on the problem. A simple solution is to assume that the velocity field is smooth (Horn & Schunck, 1981), or that it is constant over a small spatial neighborhood surrounding the location of interest (Lucas & Kanade, 1981), and to combine the constraints over that region. Alternatively, or in conjunction with the spatial combination, one can combine constraints arising from differently filtered versions of the image (e.g. Nagel, 1983). Many authors have described more sophisticated choices of combination rule (e.g., Hildreth, 1984; Black, 1992), but we will stick with this simplest form of spatial combination for this chapter.

Measurement noise. Next, we assume that the derivative measurements are corrupted by a small amount of noise, as must be true in any real system. For computational convenience, we assume the noise is added to the temporal derivative, and that the values are distributed according to a Gaussian probability density. Although this simple noise model is unlikely to be correct in detail (in particular, it is not a very accurate description of the noise in neural responses), it is sufficient to illustrate the fundamental issues of motion estimation.

Mathematically, we write:

$$I_x u + I_y w + I_t = n,$$

where n is a random variable representing the noise. This enables us to write the probability of observing the spatio-temporal image structure assuming a given velocity:

$$P(I|u,w) \propto \exp[-(I_x u + I_y w + I_t)^2/(2\sigma_n^2)],$$

where σ_n indicates the standard deviation of the noise variable n. Empirically, de Ruyter measured the relationship between gradient measurements and translational velocity in video footage gathered from a head-mounted camera, and found that it is roughly consistent with this formulation (de Ruyter, 2002).

Now, as described previously, we combine the constraints over a small spatial region, over which we assume the velocity vector is constant and the noise variables are independent:

$$P(I|u,w) \propto \exp\left(-\sum (I_x u + I_y w + I_t)^2/(2\sigma_n^2)\right), \quad (2)$$

where the sum combines the measurements at all locations within the neighborhood.

Prior probability distribution on velocity Equation (2) describes the *likelihood* function: the probability of observing a spatio-temporal image pattern given velocity (u, w). In Bayesian estimation, one reverses this conditionalization using Bayes' Rule, in order to get the *posterior* distribution:

$$P(u,w|I) = \frac{P(I|u,w)P(u,w)}{\int du \int dw P(I|u,w)P(u,w)} \quad (3)$$

The probability density $P(u, w)$ is known as the *prior*: it describes the probability of observing each velocity (independent of any particular image measurements).

For the prior probability distribution, we again make a choice of mathematical convenience: we assume that $P(u, w)$ is Gaussian with zero mean:

$$P(u,w) \propto \exp\left[-(u^2+w^2)/2\sigma_v^2\right]. \quad (4)$$

As in the choice of the noise model, this may not be an accurate characterization of the distribution of image velocities that an organism would encounter in the world, but it suffices to capture the main aspects of a reasonable solution to the motion estimation problem. The intuition is that in the absence of any specific image information (e.g., in a dark room), one should assume that things are not moving. More generally, the model proposes that slow velocities occur more frequently than fast ones. This hypothetical property of the environment is mirrored by the long-standing hypothesis that the perceived motion of stimuli corresponds to the slowest motion consistent with the observed image information (see references in Ullman, 1979). In particular, Wallach (1935) proposed that humans tend to see line segments as moving in a direction normal to their orientation because the velocity associated with that direction is the slowest of all velocities consistent with the image information.

Combining equations (2) and (4) as specified in equation (3) gives the posterior distribution:

$$P(u,w|I) \propto \exp\left[-(u^2+w^2)/2\sigma_v^2 - \sum (I_x u + I_y w + I_t)^2/2\sigma_n^2\right]. \quad (5)$$

This specifies a Gaussian distribution over (u, w) whose behavior is illustrated in figure 2. Note that the distribution captures the uncertainties shown in the conceptual

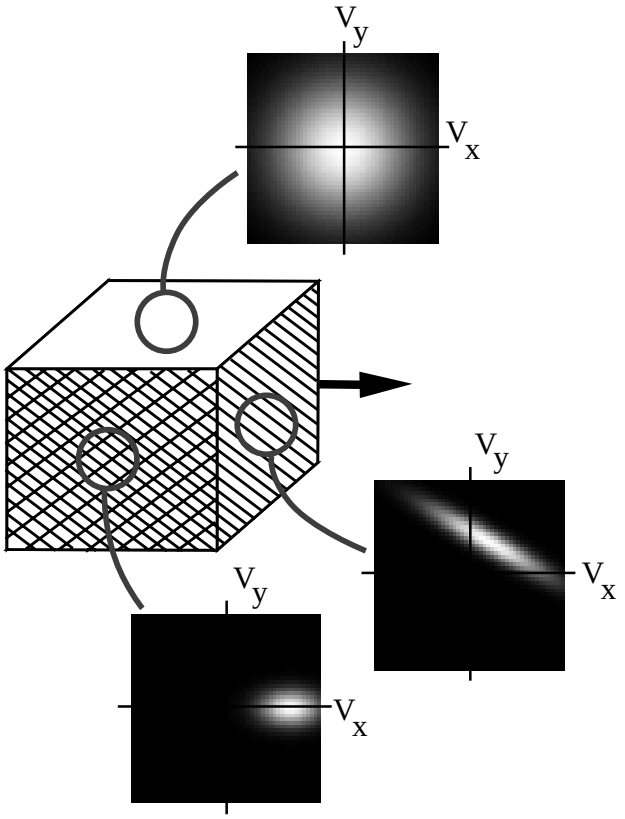

Figure 2. Illustration of posterior probability densities for the example shown in Figure 1.

illustrations of figure 1. In particular, the posterior density for the blank region is quite broad (it is identical to the prior). In the striped region, the density is elongated and lies along the line that defines the set of velocities consistent with the the image constraints. The width along this line is determined by the prior variance, σ_v^2. The width perpendicular to the line is determined by the noise variance, σ_n^2, as well as the image contrast. And in the patterned region, the density is much more compact, and centered on the correct horizontal velocity.

Physiological interpretation

A Bayesian model is appealing because of its generality, and its construction from a relatively small number of realistic assumptions. It can be implemented in a distributed analog network (Stocker, 2001), and has been used in machine vision applications. In previous work, we have shown that the mean (equivalently, the maximum) of the Gaussian posterior distribution described by equation (5) provides a surprisingly good match to human perception of velocity for a variety of translating stimuli (Simoncelli, 1993; Heeger & Simoncelli, 1993; Weiss, 1998; Weiss *et al.*, 2002). In this section, I'll describe the relationship between the elements of this framework and the functional properties of neurons that lie in the so-called *motion pathway* of mammals such as cats or monkeys. This should not be thought of as a quantitative physiological model of motion processing, but as a more qualitative assignment of computational function to neural populations along the motion pathway.

We'll work backward through the system, starting from the posterior distribution of equation (5). If we hold the velocity (u, w) fixed, then this function is tuned for the velocity of the input image. That is, it is maximized when the underlying image structure is consistent with the velocity (u, w), and it decreases as the image motion deviates from that velocity. We may identify this basic property with those neurons in visual area MT that are known as *pattern selective* (Movshon *et al.*, 1986) (see Britten, Chapter XX, this volume). These neurons are tuned for retinal image velocity; they respond vigorously to a visual stimulus moving with a particular speed and direction, and are relatively indifferent to the stimulus' spatial pattern (Maunsell & van Essen, 1983; Movshon *et al.*, 1986; Rodman & Albright, 1987).

A number of models posit that a population of such neurons provides a distributed representation of velocity (Heeger, 1987; Koch *et al.*, 1989; Grzywacz & Yuille, 1990). Here, we assume that the responses within this population correspond directly to samples of the posterior density. That is, the response of each neuron corresponds to equation (5) evaluated at a different *preferred* velocity (u, w). Note that in this interpretation, no single neuron encodes the velocity of the stimulus, and an actual estimate of the velocity can be obtained only by combining information across the population. For example, one could compute the mean velocity by summing the preferred velocities of all of the neurons in the population, each weighted by the response of that neuron. Also note that in this interpretation, the prior distribution, $P(u, w)$, provides only a gain adjustment (multiplicative scale factor) on each of the neural responses. The integral over (u, w) in the denominator of equation (3) corresponds to a sum over the responses of all neurons in the population. We have previously shown that this kind of divisive normalization operation

is consistent with the response properties of MT neurons (Simoncelli & Heeger, 1998).

We can make a qualitative comparison of this model to typical neural responses in area MT. The lower left panel of figure 3 shows polar plots of the response of a model MT neuron with preferred velocity $u = 0.3, w = 0$ pixels per frame, as a function of the normal direction of a drifting sinusoidal grating. As with MT pattern cells (Movshon *et al.*, 1986), the response is tuned for the direction of motion, and is largest when this direction matches the preferred direction. The lower right panel shows the response as a function of direction for a *plaid* stimulus, constructed as the sum of two drifting gratings 120 degrees apart. The response is largest when this plaid stimulus is moving rightward, a configuration in which the two constituent gratings are moving at ± 60 degrees from rightward. This response to the direction of pattern motion (rather than to the two component directions) is a characteristic used to identify MT pattern cells (Movshon *et al.*, 1986).

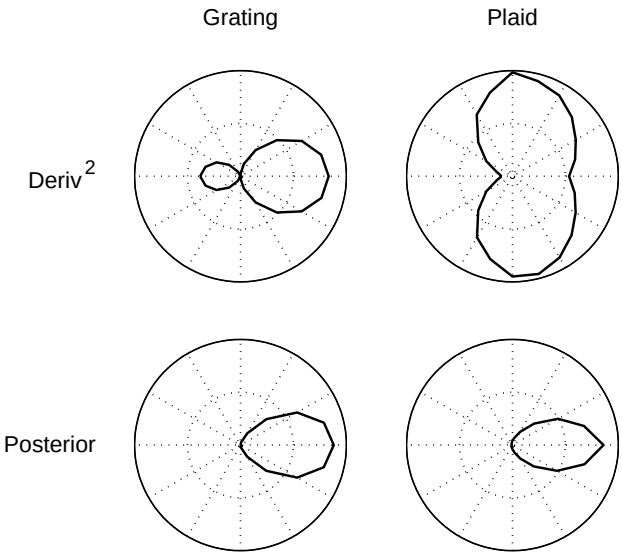

Figure 3. Direction tuning curves for model components. Upper left: Model V1 neuron (squared directional derivative) response to drifting sinusoidal grating; Upper right: Model V1 neuron response to drifting sinusoidal plaid; Lower left: Model MT neuron (posterior probability) response to grating; Lower right: Model MT neuron response to plaid.

Figure 4 shows the posterior function plotted as a function of stimulus speed (again for a drifting sinusoidal grating). The maximum response occurs at the 0.3 pixels per frame, the speed corresponding to the chosen u. At half height, the curve spans a speed range of about two octaves, which is roughly comparable to the responses of some MT cells (Maunsell & van Essen, 1983).

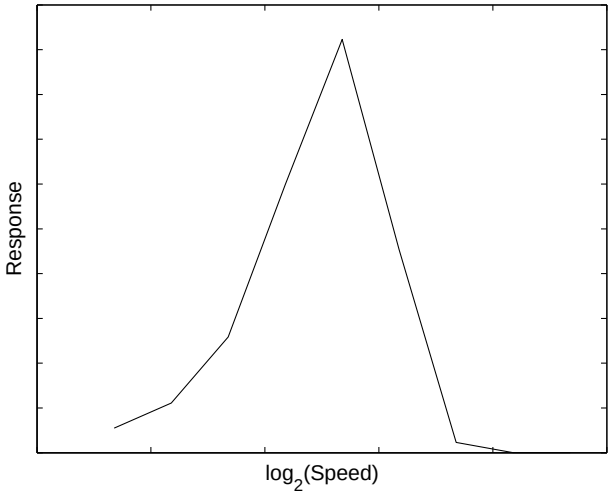

Figure 4. Speed tuning curve for posterior distribution (model MT neuron) to a drifting sinusoidal grating. The absolute scale of the horizontal axis is arbitrary, but the tick marks correspond to increments of speed by multiples of 2.

Having identified the posterior distribution with MT pattern responses, we now want to consider the components (i.e., the afferents) from which those responses are generated. The exponent of equation (5) may be expanded (dropping a factor of two) as follows:

$$\begin{aligned} f(I,u,w) = & \\ & -\sum [u^2 I_x^2 + 2uwI_xI_y + w^2 I_y^2 + 2uI_xI_t + 2wI_yI_t + I_t^2]/ \\ & -[u^2+w^2]/\sigma_v^2, \end{aligned}$$

where the sum is taken over a small spatial neighborhood. Many of the terms contain squared image derivatives, and those that contain products of image derivatives may be written as a difference of two squared derivatives. For example:

$$\begin{aligned} I_xI_y &= [(I_x+I_y)^2 - (I_x-I_y)^2]/4 \qquad (6) \\ &= [I_{d1}^2 - I_{d2}^2]/4, \end{aligned}$$

where I_{d1} and I_{d2} are derivatives at angles 45 degrees and -45 degrees, respectively. In general, a derivative

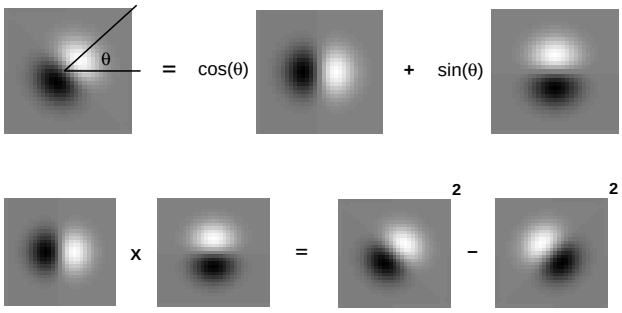

Figure 5. Illustration of two identities relating separable and oriented derivative operations. Top: a spatial derivative operator at any orientation may be computed as a linear combination of separable (axis-aligned) derivative operators (equation (7)). Bottom: A product of X and Y derivatives may be computed as the difference of two squared obliquely oriented derivatives (equation (6)). Operators in both cases are Gaussian derivatives, but the identities hold for derivatives of other functions as well.

at any orientation may be constructed via suitable combination of axis-aligned (separable) derivative operators (Freeman & Adelson, 1991):

$$I_\theta = \cos(\theta) I_x + \sin(\theta) I_y, \tag{7}$$

where I_θ is a derivative in direction θ. The mathematical relationships of equations (6) and (7) are illustrated in figure 5.

An analogous transformation allows us to write the term containing $I_x I_t$ as a difference of two derivatives. Since one of the axes is now time, the result is a difference of space-time oriented derivatives, $\{I_r, I_l\}$, which are most responsive to vertically oriented structures that are moving rightward/leftward. This construction, known as an *opponent motion* computation, has been previously proposed as an explanation for a number of psychophysical phenomena (Adelson & Bergen, 1985). Similarly, the product $I_y I_t$ results in a difference of squared upward and downward derivatives, $\{I_u, I_d\}$. Combining all of this allows us to write the exponent as:

$$\begin{aligned} f(I,u,w) = -\frac{1}{\sigma_n^2} \sum \Big[& u^2 (I_x^2 + \sigma_n^2/\sigma_v^2) + \\ & uw(I_{d1}^2 - I_{d2}^2)/2 + w^2(I_y^2 + \sigma_n^2/\sigma_v^2) + \\ & u(I_r^2 - I_l^2)/2 + w(I_u^2 - I_d^2)/2 + I_t^2 \Big]. \end{aligned} \tag{8}$$

The purpose of these transformations is to show that the computation of the posterior is based on a sum of terms that could arise in the responses of primary visual cortical neurons. The receptive fields of so-called *simple cells* in primary visual cortex (area V1) of cats and monkeys are selective for stimulus position and orientation (Hubel & Wiesel, 1962) (see Ferster, Chapter XX, this volume). Many simple cells are also direction-selective: they give stronger responses for stimuli moving in one direction than the opposite direction (see DeAngelis and Anzai, Chapter XX, this volume). Many authors have characterized simple cell responses as the halfwave rectified (and sometimes squared) responses of linear receptive fields (e.g. Campbell *et al.*, 1968; Campbell *et al.*, 1969; Movshon *et al.*, 1978; Daugman, 1985; Heeger, 1992). In these models, the neuronal response is derived from a weighted sum (over local space and recently past time) of the local stimulus contrast. This type of model can explain the primary properties of these cells, including selectivity for stimulus orientation, spatial frequency, temporal frequency, and direction.

The linear derivative filters used in our Bayesian motion estimator (as shown in figure 5) bear some similarity to the response properties of simple cells: they are tuned for spatial orientation, spatial frequency, temporal frequency and direction. Example direction tuning curves of squared derivative operators for both gratings and plaids are shown in the top row of figure 3. Note that, as in V1 neurons, the response to the plaid is bimodal: unlike the posterior response shown on the bottom row, the operator responds to each of the components of the plaid rather than the pattern as a whole.

In general, tuning of first derivative operators for most variables is significantly broader than that of most simple cells. We have shown in previous work that this inconsistency is easily corrected through the use of higher-order derivative operators (Simoncelli, 1993; Heeger & Simoncelli, 1993; Simoncelli & Heeger, 1998). The Bayesian solution given above may be re-derived using such operators, and the form of the result is essentially the same despite the proliferation of terms in the equations.

Also prevalent in the primary visual cortex are *complex cells*, which exhibit similar selectivity for orientation, spatial frequency, temporal frequency and direction, but which are more nonlinear in their response properties. In particular, their responses are relatively insensitive to the precise location of stimuli within their receptive fields. Models of these these cells have been based on

local oriented *energy*, computed as the sum of squared responses of even- and odd-symmetric oriented linear filters (e.g. Adelson & Bergen, 1985). In the Bayesian model of this paper, this type of construction could be achieved by combining the constraints obtained from differently filtered versions of the input image. Alternatively, this translation-invariance of responses is approximately achieved with a sum of squared linear responses over a local spatial neighborhood, as in equation (8). Summarizing, the posterior distribution of our Bayesian model is constructed from building blocks that resemble direction-selective complex cells. This parallels the physiology, where it has been shown that the neurons that project from area V1 to MT are direction-selective complex cells (Movshon & Newsome, 1996).

Finally, we need to address the construction of the spatiotemporally oriented V1 responses from the primary afferents arriving from the lateral geniculate nucleus. These responses are generally not orientation-tuned and they are not directional (in cats and monkeys). Specifically, the receptive fields are described by a product of a spatial (center-surround) weighting function and a temporal weighting function (see Sherman, Chapter XX, this volume). In the context of the Bayesian model, spatial derivative computation (to produce orientation-selective receptive fields) can be obtained by spatial combination of lateral geniculate nucleus receptive fields, as has been suggested physiologically (Hubel & Wiesel, 1962; Reid & Alonso, 1995). A number of authors have suggested that directional responses may be constructed physiologically by superimposing an appropriate set of space-time separable responses (e.g., Fahle & Poggio, 1981; Watson & Ahumada, 1983; Adelson & Bergen, 1985). In the case of the directional derivatives used in this chapter, one need only sum a spatial and temporal derivative operator (with the correct weighting) to obtain a spatiotemporally oriented derivative operator selective for any desired orientation. This construction is made explicit in equation (7), and illustrated in the top panel of figure 5.

Discussion

We have provided a Bayesian analysis for local motion based on a minimal set of assumptions: (1) brightness conservation, (2) a simple model of measurement noise, and (3) a prior preference for slower speeds. Given these assumptions, the components of the optimal solution can be seen to have properties qualitatively matching those of neurons in the mammalian motion pathway.

In previous work, we have shown that this model accounts for a surprising number of psychophysical results demonstrating non-veridical perception of motion stimuli (Weiss *et al.*, 2002). We have also shown that an elaborated variant of this model can be fit more precisely to neural response properties (Simoncelli & Heeger, 1998). In that model, narrower V1 stage tuning curves are achieved through use of higher-order derivatives, and nonlinear properties of V1 responses are incorporated using divisive normalization.

The particular Bayesian model described in this chapter is the simplest of its kind. A more correct model should include a more realistic model of uncertainty in photoreceptors, as well as in subsequent neural responses. It should also include a prior that more accurately reflects the velocity distribution in the visual environment (although this is quite difficult to model, given that it depends not just on the environment, but also on the motion of the organism). Such modifications are unlikely to lead to qualitatively different behaviors of the solution, but they may produce a more accurate account of the physiology.

Finally, the formulation of the motion estimation problem using brightness constancy is simplistic in assuming that a single translational velocity accounts for the motion in each local region. As described earlier, this assumption is violated in real scenes near occlusion boundaries and in the presence of transparent surfaces. Studies in computer vision have suggested that segmentation or grouping of the scene must be tightly integrated with the motion estimation solution, and a number of authors have proposed joint solutions (e.g., Darrell & Pentland, 1995; Wang & Adelson, 1994; Ayer & Sawhney, 1995; Weiss & Adelson, 1996). These solutions are invariably recurrent, perhaps suggesting that physiological implementations will require recurrent lateral or feedback projections between two neural populations computing velocity and grouping.

References

Adelson, E. H. & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*, 284–299.

Ayer, S. & Sawhney, H. (1995). Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. In *Proc Int'l Conf Computer Vision*, Cambridge, MA.

Black, M. J. (1992). A robust gradient method for determining optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition*, Champagne-Urbana.

Campbell, F. W., Cleland, B. G., Cooper, G. F., & Enroth-Cugell, C. (1968). The angular selectivity of visual cortical cells to moving gratings. *Journal of Physiology (London)*, *198*, 237–250.

Campbell, F. W., Cooper, G. F., & Enroth-Cugell, C. (1969). The spatial selectivity of visual cells of the cat. *Journal of Physiology (London)*, *203*, 223–235.

Chubb, C. & Sperling, G. (1988). Drift-balanced random stimuli: a general basis for studying non-fourier motion perception. *Journal of the Optical Society of America A*, *5*, 1986–2006.

Darrell, T. & Pentland, A. (1995). Cooperative robust estimation using layers of support. *IEEE Trans Pattern Analysis and Machine Intelligence*, *17*(5), 474–487.

Darrell, T. & Simoncelli, E. (1994). Separation of transparent motion into layers using velocity-tuned mechanisms. In Eklundh, J. O., editor, *Third European Conf on Computer Vision, Stockholm*. Springer-Verlag.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, *2*(7), 1160–1169.

de Ruyter, R. (2002). Personal communication.

Fahle, M. & Poggio, T. (1981). Visual hyperacuity: spatiotemporal interpolation in human vision. *Proc. Royal Society of London, B*, *213*, 451–477.

Fennema, C. L. & Thompson, W. (1979). Velocity determination in scenes containing several moving objects. *Computer Graphics and Image Processing*, *9*, 301–315.

Fleet, D. J. & Langley, K. (1994). Computational analysis of non-fourier motion. *Vision Research*, *22*, 3057–3079.

Freeman, W. T. & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Pat. Anal. Mach. Intell.*, *13*(9), 891–906.

Gibson, J. J. (1950). *The Perception of the Visual World*. Houghton Mifflin, Boston, MA.

Grzywacz, N. M. & Yuille, A. L. (1990). A model for the estimate of local image velocity by cells in the visual cortex. *Proc. Royal Society of London A*, *239*, 129–161.

Heeger, D. J. (1987). Model for the extraction of image flow. *J. Opt. Soc. Am. A*, *4*(8), 1455–1471.

Heeger, D. J. (1992). Half-squaring in responses of cat simple cells. *Visual Neuroscience*, *9*, 427–443.

Heeger, D. J. & Simoncelli, E. P. (1993). Model of visual motion sensing. In Harris, L. & Jenkin, M., editors, *Spatial Vision in Humans and Robots*, chapter 19, pages 367–392. Cambridge University Press.

Hildreth, E. C. (1984). Computations underlying the measurement of visual motion. *Artificial Intelligence*, *23*(3), 309–355.

Horn, B. K. P. (1986). *Robot Vision*. MIT Press, Cambridge, MA.

Horn, B. K. P. & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*, 185–203.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, *160*, 106–154.

Knill, D. & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Koch, C., Wang, H., & Mathur, B. (1989). Computing motion in the primate's visual system. *The Journal of experimental Biology*, *146*, 115–139.

Koechlin, E., Anton, J. L., & Burnod, Y. (1999). Bayesian inference in populations of cortical neurons: a model of motion integration and segmentation in area MT. *Biol. Cybernetics*, *80*, 25–44.

Lucas, B. D. & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proc. 7th Int'l Joint Conf. on Artificial Intelligence*, pages 674–679, Vancouver.

Marr, D. & Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, *15*, 470–488.

Marr, D. & Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proc. Royal Society of London B*, *211*, 151–180.

Maunsell, J. H. R. & van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey I. Selectivity for stimulus direction, speed, and orientation. *Journal of Neurophysiology*, *49*, 1127–1147.

Movshon, J. A., Adelson, E. H., Gizzi, M. S., & Newsome, W. T. (1986). The analysis of moving visual patterns. In Chagas, C., Gattass, R., & Gross, C., editors, *Experimental Brain Research Supplementum II: Pattern Recognition Mechanisms*, pages 117–151. Springer-Verlag, New York.

Movshon, J. A. & Newsome, W. T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *Visual Neuroscience*, *16*(23), 7733–7741.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology (London)*, *283*, 53–77.

Nagel, H. H. (1983). Displacement vectors derived from second order intensity variations in image sequences. *Computer Vision, Pattern Recognition, and Image Processing*, *21*, 85–117.

Nowlan, S. J. & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, *15*, 1195–1214.

Reid, R. C. & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature*, *378*, 281–284.

Rodman, H. R. & Albright, T. D. (1987). Coding of visual stimulus velocity in area MT of the macaque. *Vision Research*, *27*, 2035–2048.

Simoncelli, E. P. (1993). *Distributed Analysis and Representation of Visual Motion*. PhD thesis, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Simoncelli, E. P. & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743–761.

Stocker, A. A. (2001). *Constraint Optimization Networks for Visual Motion Perception - Analysis and Synthesis*. Ph.d. thesis no. 14360, Swiss Federal Institute of Technology ETHZ, Zürich, Switzerland.

Ullman, S. (1979). *The Interpretation of Visual Motion*. MIT Press, Cambridge, MA.

Verri, A. & Poggio, T. (1989). Motion field and optical flow: Qualitative properties. *IEEE Pat. Anal. Mach. Intell.*, *11*(5), 490–498.

Wallach, H. (1935). Über visuell whargenommene bewegungrichtung. *Psychologische Forschung*, *20*, 325–380.

Wang, J. Y. & Adelson, E. H. (1994). Representing moving images with layers. *IEEE Trans on Image Processing*, *3*(5), 625–638.

Watson, A. B. & Ahumada, A. J. (1983). A look at motion in the frequency domain. In Tsotsos, J. K., editor, *Motion: Perception and representation*, pages 1–10. Assoc. for Computing Machinery, New York.

Weiss, Y. (1998). *Bayesian motion estimation and segmentation*. PhD thesis, Dept. Brain and Cognitive Sciences, Massachusetts of Technology.

Weiss, Y. & Adelson, E. H. (1996). A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. In *Proc IEEE Conf Computer Vision and Pattern Recognition*, pages 321–326.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

Wilson, H. R. & Kim, J. (1994). A model for motion coherence and transparency. *Visual Neuroscience*, *11*, 1205–1220.