Contents

List of Figures

Chapter 1

Motion and Depth

The perception of motion, like the perception of color, is a visual inference. The images encoded by the photoreceptors are merely changing two-dimensional patterns of light intensity. Our visual pathways *interpret* these two-dimensional images to create our perception of objects moving in a three-dimensional world. How it is possible, even in principle, to infer three-dimensional motion from the two-dimensional retinal image is one important puzzle of motion; how our visual pathways actually make this inference is a second.

In this chapter we will try first to understand abstractly the retinal image information that is relevant to the computation of motion estimation. Then, we will review experimental measurements of motion perception and try to make sense of these measurements in the context of the computational framework. Many of the experiments involving judgment of motion suggest that visual inferences concerning motion use information about objects and their relationships, such as occlusion and texture boundaries, to detect and interpret motion. Finally, we will review a variety of experimental measurements that seek to understand how motion is computed within the visual pathways. Physiologists have identified a visual stream whose neurons are particularly responsive to motion; in animals, lesions of this pathway lead to specific visual deficits of motion. The specific loss of motion perception, without an associated loss of color or pattern sensitivity, suggests that a portion of the visual pathways is specialized for motion perception.

It is useful to begin our review of motion with a few of the many behavioral observations that show motion is a complex visual inference, and not a simple representation of the physical motion. One demonstration of this, called *apparent motion*, can be seen in many downtown districts. There you will find displays consisting of a sequence of flashing marquis lights that appear to be moving, drawing your attention to a theater or shop. Yet, none of the individual lights in the display are moving; the lights are fixed in place, flashing in sequence. Even though there is no physical motion, the stimulus evokes a sense of motion in us. Even a single pair of lights, flashing in alternation, can provide a distinct visual impression of motion.

A second example of a fixed stimulus that appear to move can be found in *motion aftereffects*. Perhaps the best known motion aftereffect is the waterfall illusion, described in the following passage.

> Whilst on a recent tours of the highlands of Scotland, I visited the celebrated Falls of Foyers near Loch Ness, and there noticed the following phaenomenon.
>
> Having steadfastly looked for a few seconds at a particular part of the cascade, admiring the confluence and decussation of the currents forming the liquid drapery of waters, and then suddenly directed my eyes to the left, to observe the vertical face of the sombre age-worn rocks immediately contiguous to the waterfall, I saw the rocky surface as if in motion upwards, and with an apparent velocity equal to that of the descending water, which the moment before had prepared my eyes to behold this singular deception (Addams, 1834).

Notice that in the waterfall illusion, the object that appears to move (the rock) does not have the same shape or texture as the object that causes the motion aftereffect (the waterfall). The waterfall Addams described is shown in Figure 1.1.

A third way to convince yourself that motion is an inference is to consider the fact that many behavioral experiments show that perceived velocity, unlike physical velocity, depends on the color and contrast of an object. We know that the color of an object or its contrast relative to the background, are not good cues about motion. Indeed, the physical definition of motion does not depend on color or contrast at all. Yet, there are many motion demonstrations that perceived velocity depends on contrast and color. Some of the effects are quite large. For example, by the proper selection of the color and pattern a peripheral target moving at 1 degree per second can be made to appear as though it were still. These effects show that the visual inference of motion is imperfect because it is influenced by image features that are irrelevant to physical motion (Cavanagh and Anstis, 1991).

Motion computations and stereo depth perception are closely related. For example, as an observer moves the local motion of image points contain useful visual information about the distance from the observer to various points in the scene. As the observer moves, points in the image change in a predictable way that depends on the direction of the observer's motion and the distance of the point from the observer. Points that are further away generally move smaller amounts than points that are closer; and, points along the direction of heading move less than points far away from the direction of heading. Information from the image motion informs us

Addam’s
Waterfall

Figure 1.1: *The Falls of Foyers* where Addams observed the motion aftereffect called the waterfall illusion. The illusion demonstrates that perceived motion is different from physical motion; we see motion in the aftereffect although there is no motion in the image (Source: N. Wade).

about the position of objects in space relative to the viewer (Gibson, 1950).

The collective motion of points in an image from one moment to the next is called the *motion flow field*. By drawing our attention to this source of depth information, Gibson (1950) established an important research paradigm that relates motion and depth perception: define an algorithm for measuring the motion flow field, and then devise algorithms to extract information about observer motion and object depth from the motion flow field. In recent years, this problem has been substantially solved. It is now possible to use motion flow fields to estimate both an observer's motion through a static scene and a depth map from the observer to different points within the scene (Heeger and Jepson, 1992, Tomasi and Kanade, 1992; Koenderink, 1990).

These computational examples show that motion and stereo depth algorithms are related by their objectives: both inform us about the positions of objects in space and the viewer's relationship to those objects. Because motion and stereo depth estimation have similar goals, they use similar types of stimulus information. Stereo algorithms use the information in our two eyes to recover depth, relying on the fact that the two images are taken from slightly different points of view. Motion algorithms use a broader range of information that may include multiple views obtained as the observer moves or as objects change their position with respect to the viewer. Most of this chapter is devoted to a review of the principles in the general class of motion estimation algorithms. In a few places, because the goals of motion and stereo depth are so similar, I have inserted some related material concerning stereo depth perception.

1.1 Computing Motion

Figure 1.2 shows an overview of the key elements used in motion estimation algorithms. Figure 1.2a shows the input data, a series of images acquired over time. Because the observer and objects move over time, each image in the sequence is a little different from the previous one. The image differences are due to the new geometric relationship between the camera, objects, and light source. The goal of most motion and depth estimation algorithms is to use these changes to infer motion of the observer, the motion of the objects in the image, or the depth map relating the observer to the objects.

The arrows drawn in Figure 1.2b show the motion flow field. These arrows represent the changes in the position of points over small periods of space and time. The direction and length of the arrows correspond to the local motions that occur when the observer travels forward, down the road.

Figure 1.2c is a list of several quantities that can be estimated from the motion flow

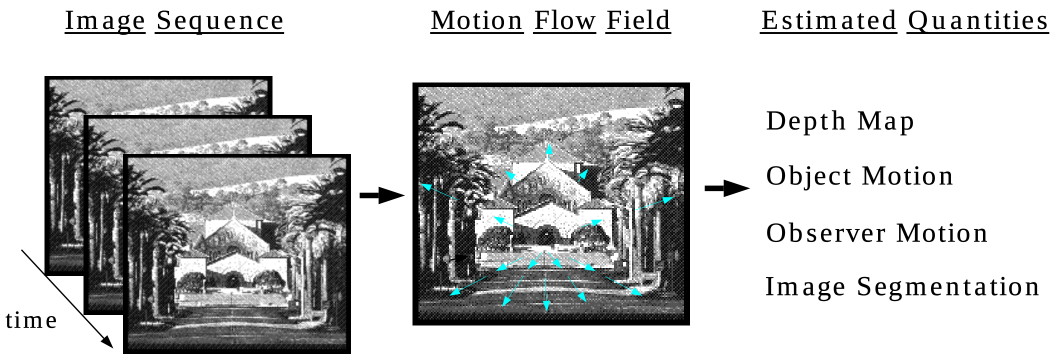

Figure 1.2: *The overall logic of motion estimation algorithms.* (a) The input stimulus consists of an image sequence. (b) The motion flow field at each moment in time is computed from the image sequence. (c) Various quantities relating to motion and depth can be calculated from the motion flow fields.

fields. As I have already reviewed, the motion flow field contains information relevant to the depth map and observer motion. In addition, the motion flow field can be used to perform, *image segmentation*, that is to identify points in the image are part of a common object. In fact, we shall see that the motion flow field defined by a set of moving points but devoid of boundaries and shading is sufficient to give rise to an impresson of a moving three-dimensional object.

Finally, the motion flow field contains information about object motion. This information is important for two separate reasons. First, as we have already reviewed, an important part of motion algorithms is establishing the spatial relationship between objects and the viewer. Second, object motion information is very important for guiding *smooth pursuit* eye movements. The structure of our retina, in which only the fovea is capable of high spatial resolution, makes it essential that we maintain good fixation on targets as we move or as the object moves in our visual field. Information derived from motion estimation is essential to guide the eye movement system as we track moving objects in a dynamic environment (Movshon and Lisberger, 1990; Komatsu and Wurtz, 1989; Schiller and Lee, 1993).

Stimulus Representation: Motion Sampling

We begin this review of motion algorithms by asking a simple question about the input data. In any practical system, the input stimulus is a sampled approximation of the continuous motion. The sampling rate should be rapid enough so that the input images provide a close approximation to the true continuous motion. How finely do we need to sample temporally the original scene in order to create a convincing visual display?

Beyond its significance for the science of motion perception, this question is also of great practical interest to people who design visual displays. A wide variety of visual display technologies represent motion by presenting a temporal sequence of still images. For example, movies and television displays both present the observer with a set of still images, each differing slightly from the previous one. In designing these display systems, engineers must select a temporal sampling rate for the display so that the observer has the illusion of smooth motion[1].

To answer the question of how finely we should sample an image sequence, we must include specifications about the image sequence and the human observer. First, consider why the answer must depend on the image sequence. Suppose that we are trying to represent a scene in which both objects and observer are still. In that case, we only need to acquire a single image. Next, suppose the image sequence contains a rapidly moving object. In that case, we need to acquire many images in order to capture all of the relevant scene information. If the image sequence contains rapidly moving objects, or if the observer moves rapidly, we must use high temporal sampling rates.

We will analyze the information in an image sequence using several simple representations shown in Figure 1.3. When we analyze the image sequence in order to estimate motion, we can call the image sequence the *motion signal* or *motion sequence.* Figure 1.3a represents the motion sequence as a three-dimensional data set: the volume of data includes two spatial dimensions and time (x, y, t). Each point in this volume sends an amount of light to the eye, $I(x, y, t)$. The data in Figure 1.3a illustrates an image sequence consisting of a dark bar on a light background moving to the right.

Next, consider two simplified versions of this three-dimensional data. Figure 1.3b shows a cross-section of the data in the (x, y) plane. This image represents the dark bar at a single moment in time. Figure 1.3c shows a cross-section of the motion volume in the (t, x) plane at a fixed value of y. In this plot time runs from the left (past) to the right (future) of the graph. The present is indicated by the dashed vertical line. The image intensity along the x direction is shown as the gray bar across the vertical axis.

When the spatial stimulus is one-dimensional, (i.e., constant in one direction) we can measure only the motion component perpendicular to the constant spatial dimension. For example, when the stimulus is an infinite vertical line, we can

[1] Different display technologies solve this problem using various special purpose methods. For example television in the U.S. displays a sequence of thirty static images per second. Each image is displayed in two frames, even numbered lines in the image are presented in one frame and odd numbered lines in the second frame. Hence, the display shows 60 frames (30 images) per second. Movie projectors display only twenty-four images per second, but each image is flashed three times so that the temporal flicker rate is 72 frames per second. Modern computer displays present complete images at seventy-two frames per second or higher. This rate is fast enough that they rarely need special tricks to avoid temporal flicker artifacts for typical image sequences.

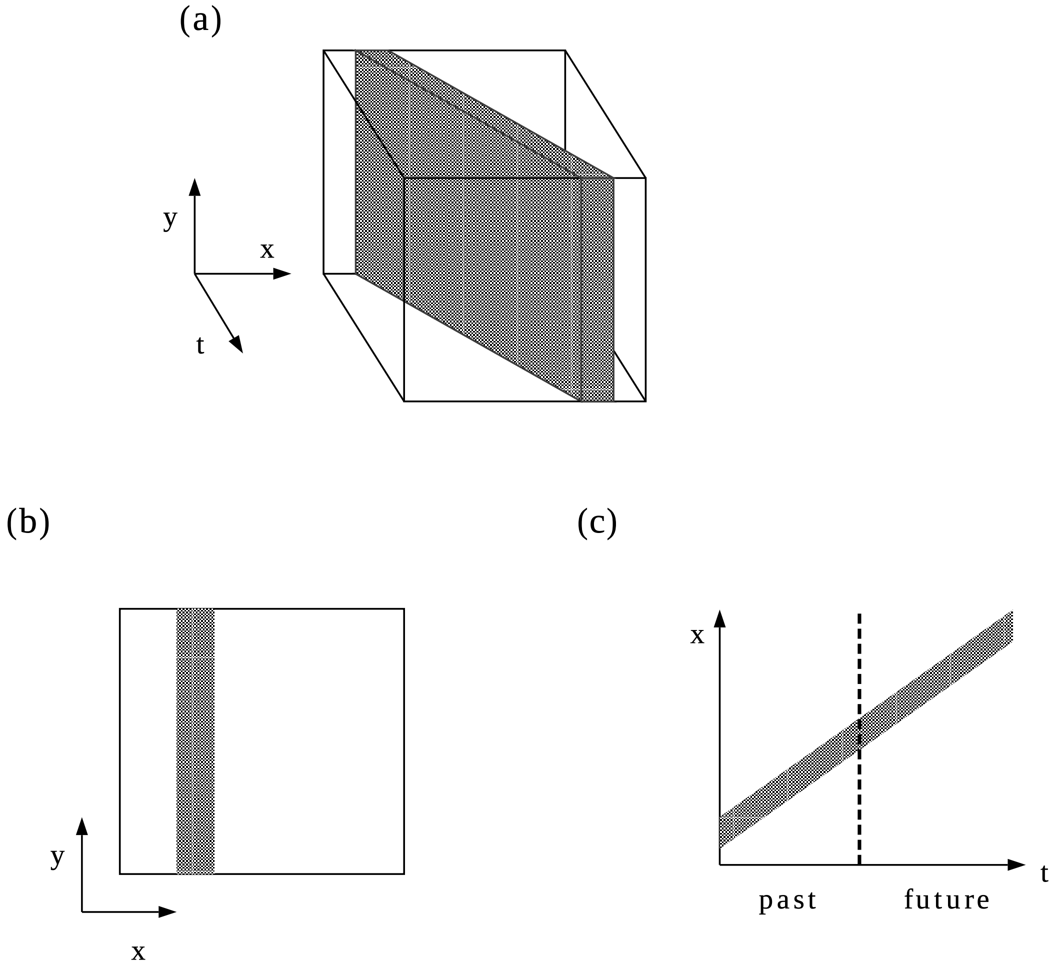

Figure 1.3: *A motion sequence* is a series of images measured over time. (a) The motion sequence of images can be grouped into a three-dimensional volume of data. Cross sections of the volume show the spatial pattern at a moment in time (panel b) time (t) plotted against one dimension (x) of space (panel c). When the spatial pattern is one-dimensional, the (t, x) cross-section provides a complete representation of the stimulus sequence.

estimate only the motion in the horizontal direction. In this case, the (t, x) cross-section is the same at all levels of y so that the (t, x) cross-section contains all of the information needed to describe the object's motion. In the motion literature, the inability to measure the motion along the constant spatial dimension is called the *aperture problem*[2].

We can use the (t, x) plot to represent the effect of temporal sampling. First, consider the (t, x) representation of a smoothly moving stimulus shown in Figure 1.4a. Now, suppose we sample the image intensities regularly in time. The sampled motion can be represented as a series of dashes in the (t, x) plot, as shown in Figure 1.4b. Each dash represents the bar at a single moment in time, and the separation between the dashes depends on the temporal sampling rate and target velocity. If the sampling rate is high, the separation between the dashes is small and the sequence will appear similar to the original continuous image. If the sampling rate is low, the separation between the dashes is large and the sequence will appear quite different from the continuous motion.

The separation between the dashes in the sampled representation also depends on the object's velocity. As the bar pattern moves faster, the dashes fall along a line of increasing slope. Thus, for increasing velocities the separation between the dashes will increase. Hence, the difference between continuous motion and sampled motion is larger for faster moving targets.

The Window of Visibility

Next, we will use measurements of visual spatial and temporal sensitivity to predict the temporal sampling rate at which a motion sequence appears indistinguishable from continuous motion. The basic procedure reviewed here has been described by several independent sources; Watson, Ahumada and Farrell (1983) call the method *the window of visibility* (see also Pearson, 1975; Fahle and Poggio, 1981)

The window of visibility method begins by transforming the images from the (t, x) representation into a new representation based on spatial and temporal harmonic functions. We can convert from the (t, x) representation to the new representation by using the Fourier transform (see Chapter **??** and the Appendix). In the new representation, the stimulus is represented by its intensity with respect to the spatial and temporal frequency dimensions, (f_t, f_x). We convert the motion signal from the (t, x) form to the (f_t, f_x) form because, as we shall see, it is easy to represent the

[2] The name "aperture problem" is something of a misnomer. It was selected because experimentally it is impossible to create infinite one-dimensional stimuli. Instead, subjects are presented finite one-dimensional patterns, such as a line-segment, through an aperture that masks the terminations of the line segment and making the stimulus effectively one-dimensional. The source of the uncertainty concerning the direction of motion, however, is not the aperture itself but rather the fact that the information available to the observer is one-dimensional.

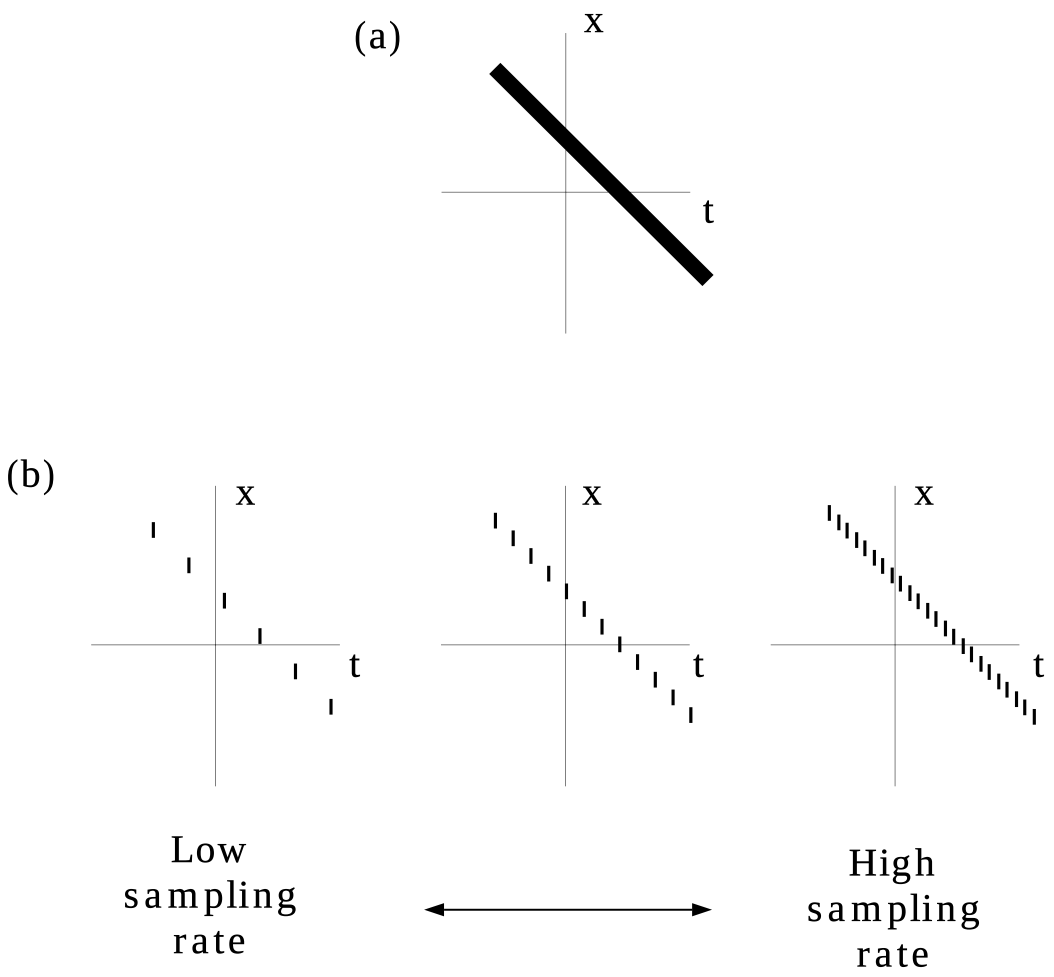

Figure 1.4: *The representation of continuous and temporally sampled motion.* (a) The continuous space-time plot of a moving line. (b) Temporal sampling of the continuous stimulus is represented by a series of dashes. As the temporal sampling rate increases (from left to right), the sampled stimulus becomes more similar to the continuous stimulus.

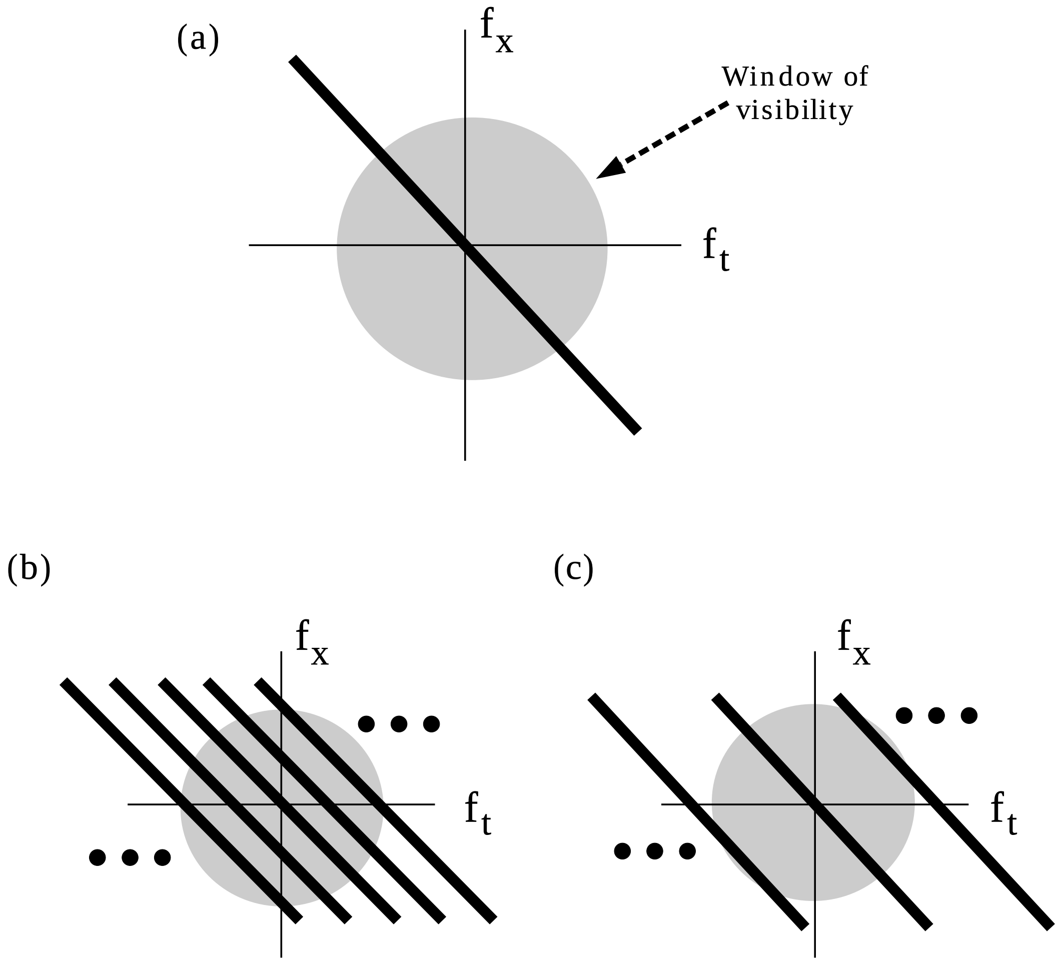

Figure 1.5: *A graphical method to decide when sampled motion can be discriminated from continuous motion.* The axes of the graphs represent the spatial and temporal frequency values of the stimulus. (a) The solid line represents the spatial and temporal frequency content of a vertical line moving continuously. The shaded area represents the range of visible spatial and temporal frequencies, that is, the window of visibility. (b,c) Temporally sampling the image introduces replicas. The spacing of the replicas depends on the temporal sampling rate. When the rate is low, the replicas are closely spaced and there is significant energy inside the window of visibility. When the sampling rate is high, the replicas are widely spaced and there is little energy inside the window of visibility. If the replicas fall outside the window, then the sampled and continuous motion will be indistinguishable.

effect of sampling in the latter representation.

In the (t, x) graph, a moving line is represented by a straight line whose slope depends on the line's velocity. In the (f_t, f_x) graph, a moving line is also represented by a straight line whose slope depends on the line's velocity. You can see how this comes about by considering the units of spatial frequency, temporal frequency, and velocity. The units of f_t and f_x are $cycles/sec$ and $cycles/deg$, respectively. The units of velocity, v, are deg/sec. It follows that spatial frequency, temporal frequency and velocity are related by the linear equation $f_t = v f_x$.

Now, a still image has zero energy at all non-zero temporal frequencies. Suppose an object begins to translate at velocity v. Then, each spatial frequency component associated with the object moves at this same velocity and creates a temporal modulation at the frequency $v f_x$. Consequently, an object moving at v will be represented by its spatial frequencies, f_x, and the corresponding temporal frequencies, $f_t = v f_x$. In the (f_t, f_x) graph this collection of points defines a line whose slope depends on the velocity, as shown in Figure 1.5a [3].

We use the (f_t, f_x) representation because it is easy to express the sampling distortion in that coordinate frame. A basic result of Fourier Theory is that temporally sampling a continuous signal creates a set of replicas of the original continuous signal in the (f_t, f_x) representation[4]. The temporal sampling replicas are displaced from the original along the f_t axis; the size of the displacement depends on the sampling rate. When the sampling is very fine, the replicas are spaced far from the original. When the sampling is very coarse, the replicas are spaced close to the original.

Figure 1.5bc contain graphs representing temporally sampled motion. In panel b, the temporal sampling rate is low and the replicas are spaced close to the original signal. At this sampling rate, the sampled motion is plainly discriminable from the continuous motion. In panel c, the temporal sampling rate is high and the replicas are far from the original signal. To make the sampled image appear like continuous motion, we must set the temporal sampling rate high enough so that the sampling distortion is invisible. The problem that remains is to find the sampling rate at which the replicas will be invisible.

The shaded area in each panel of Figure 1.5 shows a region called the *window of visibility*. The window describes the spatio-temporal frequency range that is

[3] Speaking more precisely, the Fourier Transform maps the intensities in the (t, x) plot into a set of complex numbers. The plot shown in this figure represents the locations of the nonzero components of the data after applying the Fourier Transform.

[4] I do not have a simple intuitive reason why the distortions are arranged in this way. Ordinarily, this effect of sampling is proven by appeal to properties of the Fourier Transform that are slightly beyond the scope of this volume. But, the interested reader can find a proof of the consequences of sampling, and many other useful properties of the Fourier Transform, in Bracewell (1978) and Oppenheim et al., 1983).

detectable to human observers. The boundary of the window is a coarse summary of the limits of space-time sensitivity that we reviewed in Chapter **??**. Spatial signals beyond roughly 50 cycles per degree, or temporal frequency signals beyond roughly 60 cycles per second, are beyond our perceptual limits. If the sampling replicas fall outside the window of visibility, they will not be visible and the difference between the continuous and sampled motion will not be perceptible. Hence, to select a temporal sampling rate at which sampled motion will be indiscriminable from continuous motion, we should select a temporal sampling rate such that the replicas fall outside of the window of visibility.

The window of visibility method is a very useful approximation to use when we evaluate the temporal sampling requirements for different image sequences and display technologies. But, the method has two limitations. First, the method is very conservative. There will be sampled motion signals that fail to contain energy within the window and yet the sampled motion will still appear to be continuous motion. This will occur because the energy that unwanted sampling energy that falls within the window of visibility may be masked by the energy from the continuous motion (see Chapter **??** for a discussion of masking).

Second, the method is a limited description of motion. By examining the replicas, we can decide that the stimulus does look the same as the original continuous motion. But, the method doesn't help us to decide the motion looks like, i.e., the velocity and direction. We analyze how to estimate motion from image sequences next.

Image Motion Information

What properties of the data in an image sequence suggest that motion is present? We can answer this question by considering the representation of a one-dimensional object, say a vertical line, in the (t, x) plot. When the observer and object are still, the intensity pattern does not change across time. In this case the (t, x) plot of the object's intensity is the same for all values of t and is simply a horizontal line. When the object moves, its spatial position, x, changes across time so that in the (t, x) the path of the moving line includes segments that deviate from the horizontal. The value of the orientation of the trajectory in the (t, x) plot depends on the object's velocity. Large velocities are near the vertical; small velocities are near the horizontal orientation of a still object. Hence, *orientation* in the (t, x) representation informs us about velocity (Adelson and Bergen, 1985; Watson and Ahumada, 1985).

The connection between orientation in the (t, x) plot and velocity sensitivity suggests a procedure for estimating motion. Suppose we wish to create a neuron that responds well to motion at a particular velocity but responds little to motion at other velocities. Then, we should create a neuron whose space-time receptive field is sensitivity to signals with the proper orientation in the (t, x) plot.

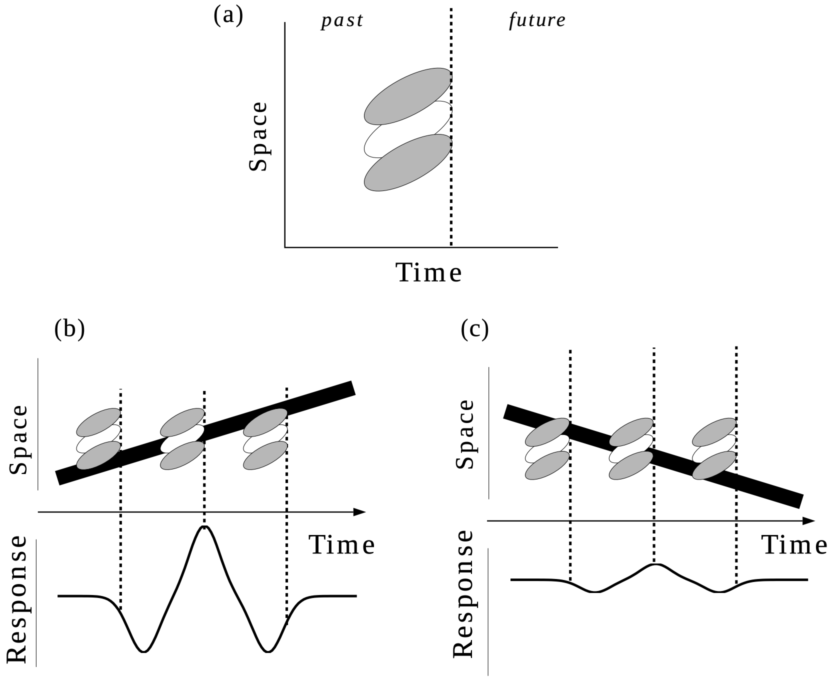

Figure 1.6: *Space-time oriented receptive field.* (a) The space-time receptive field of a neuron is represented on an (t, x) plot. The neuron always responds to events in the recent past, so the receptive field moves along the time axis with the present. The dark areas show an inhibitory region and the light area shows an excitatory region. (b) The upper portion of the graph shows an (t, x) plot of a moving line and the space-time receptive field of a linear neuron. The neuron's receptive field is shown at several different moments in time, indicated by the vertical dashed line. The common orientation of the space-time receptive field and the stimulus motion produce a large amplitude response, shown in the bottom half of this graph. (c) When the same neuron is stimulated by a line moving in a different direction, the stimulus motion aligns poorly with the space-time receptive field. Consequently, the response amplitude is much smaller.

Figure 1.6 shows the idea of a space-time oriented receptive field. In Figure 1.6a, I have represented the space-time receptive field of the neuron using the same conventions that we used to represent the space-time receptive field of neurons in Chapters **??** and **??**: The excitatory regions of the receptive field are shown as the light area and the inhibitory regions are shown as the shaded area. The horizontal axis represents time, and the dashed line represents the present moment in time. The receptive field of the neuron always responds to recent events in the past, and so it travels along with the line denoting the present, trailing just behind. This graph has much in common with the ordinary notation showing an oriented two-dimensional spatial receptive field. The graph is somewhat different from the conventional spatial receptive field because the space-time receptive field always travels in time just behind the present.

Neurons with oriented space-time receptive fields respond differently to stimuli moving in different directions and velocities. Figure 1.6b shows the response to a line moving in a direction that is aligned with space-time receptive field. The upper portion of the graph shows the relationship between the stimulus and the neuron's receptive field at several points in time. Because the stimulus and the receptive field share a common orientation, the stimulus fills the inhibitory, excitatory, and then inhibitory portions of the receptive field in turn. Consequently, the neuron will have a large amplitude response to the stimulus, as shown in the lower portion of Figure 1.6b.

Figure 1.6c shows the response of the same neuron to a stimulus moving in a different direction. The space-time orientation of this stimulus is not well-aligned with the receptive field, so the stimulus falls across the inhibitory and excitatory receptive field regions simultaneously. Consequently, the response amplitude to this moving stimulus is much smaller. Just as neurons with oriented receptive fields in (x, y) respond best to bars with the same orientation, so too neurons with oriented receptive fields in (t, x) respond best to signals with a certain velocity.

It is possible, in principle, to create neurons with space-time oriented receptive fields by combining the responses of the simple cells in area V1 of the cortex. One of many possible methods is shown in Figure 1.7a. Consider an array of neurons with adjacent spatial receptive fields. The spatial receptive fields are shown at the top of panel (a). The responses of these neurons are added together after a temporal delay, δt. The sum of these responses drives the output neuron shown at the bottom.

Figure 1.7b shows the (t, x) receptive field of the output neuron. In panel (b), the receptive field plotted along the x dimensional are the one-dimensional receptive fields of the neurons in panel (a), that is the receptive fields measured using a one-dimensional stimulus that is constant along the y axis. The spatial receptive fields of the input neurons are adjacent to one another, so they are shifted along the x dimension of the graph. The temporal response of the neurons, measured at each point in the receptive field is also shown in panel (b). Since the input neurons are

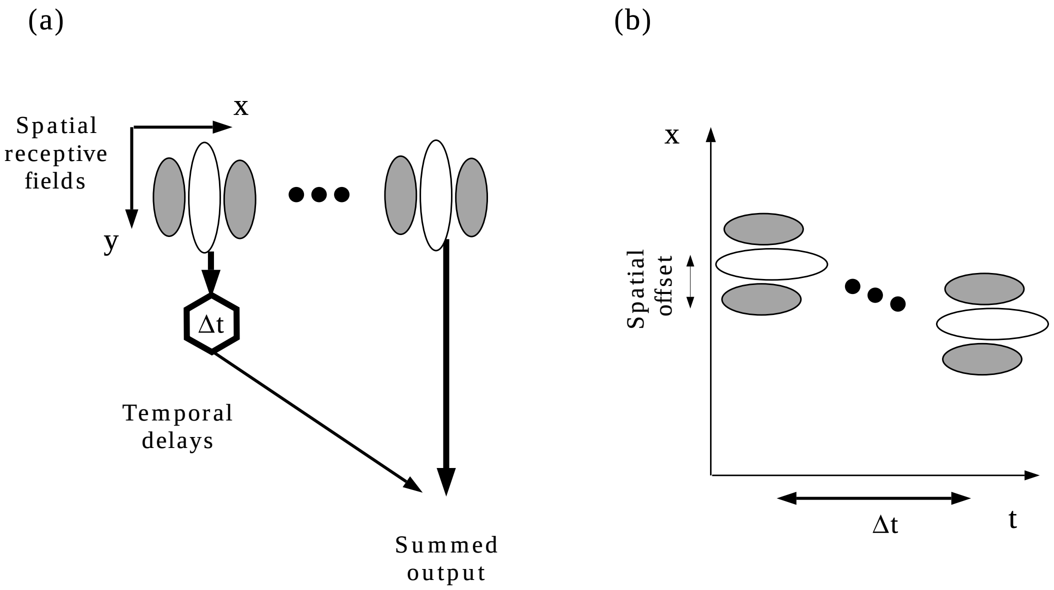

Figure 1.7: *A method for creating a space-time oriented receptive field.* (a) A pair of spatial receptive fields, displaced in the x-direction, is shown at the top. The response of the neuron on the left is delayed and then added to the response of the neuron on the right. (b) The (t, x) receptive field of the output neuron in panel (a) is shown. The temporal response of the neuron on the left is delayed compared to the temporal response of the neuron on the right. The combination of spatial displacement and temporal delay yield an output neuron whose receptive field is oriented in space-time.

delayed prior to being summed, the temporal receptive fields are shifted along the t dimension. The shift in both the x and t dimensions yield an output receptive field that is oriented in the space-time plot.

When a neuron has a space-time oriented receptive field, its response amplitude varies with the image velocity (see Figure 1.6). Thus, to measure stimulus velocity we need to compare the response amplitudes of an array of neurons, each with its own preferred motion. There are various methods of computing the amplitude of the time-varying response of a neuron and comparing the results among different neurons. Generally, these methods involve simple squaring and summing operations applied to the responses. In recent years, several specific computational methods for extracting the amplitude of the neural responses have been proposed (Adelson and Bergen, 1985; Watson and Ahumada, 1985; van Santen and Sperling, 1985). We will return to this topic again after considering a second way to formulate the problem of motion estimation.

The Motion Gradient Constraint

The (t, x) representation of motion clarifies the requirements for a linear receptive field that discriminates among motion in different directions and velocities. There is a second way to describe the requirements for motion estimation that provides some additional insights. In this section, we will derive a motion estimation computation based on the assumption that in small regions of the image motion causes a displacement of the point intensities without changing the intensity values. This assumption is called the *motion gradient constraint*.

The motion gradient constraint is an approximation, and sometimes not a very good approximation. As the relative positions of the observer, objects and light sources change, the spatial intensity pattern of the reflected light changes as well. For example, when one object moves behind another the local intensity pattern changes considerably. Or, as we saw in Chapter **??**, if we are viewing a specular surface the spatial distribution of the light reflected to our eye varies as we change position. As a practical matter, however, there are often regions within an image sequence where the motion gradient constraint is a good approximation. In some applications the approximation is good enough so that we can derive useful information.

To estimate local velocity using the motion gradient, we reason as follows. We describe the image intensities in the sequence as $I(a, b, t)$, the intensity at location (a, b) and time t. Suppose the velocities in the x and y directions at a point (a, b, t) are described by the *motion vector*, (v_x, v_y). Further, suppose that images in the motion signal are separated by a single unit of time. In that case the intensity at point (a, b) will have shifted to a new position in the next frame,

$$I(a, b, t) = I(a + v_x, b + v_y, t + 1). \tag{1.1}$$

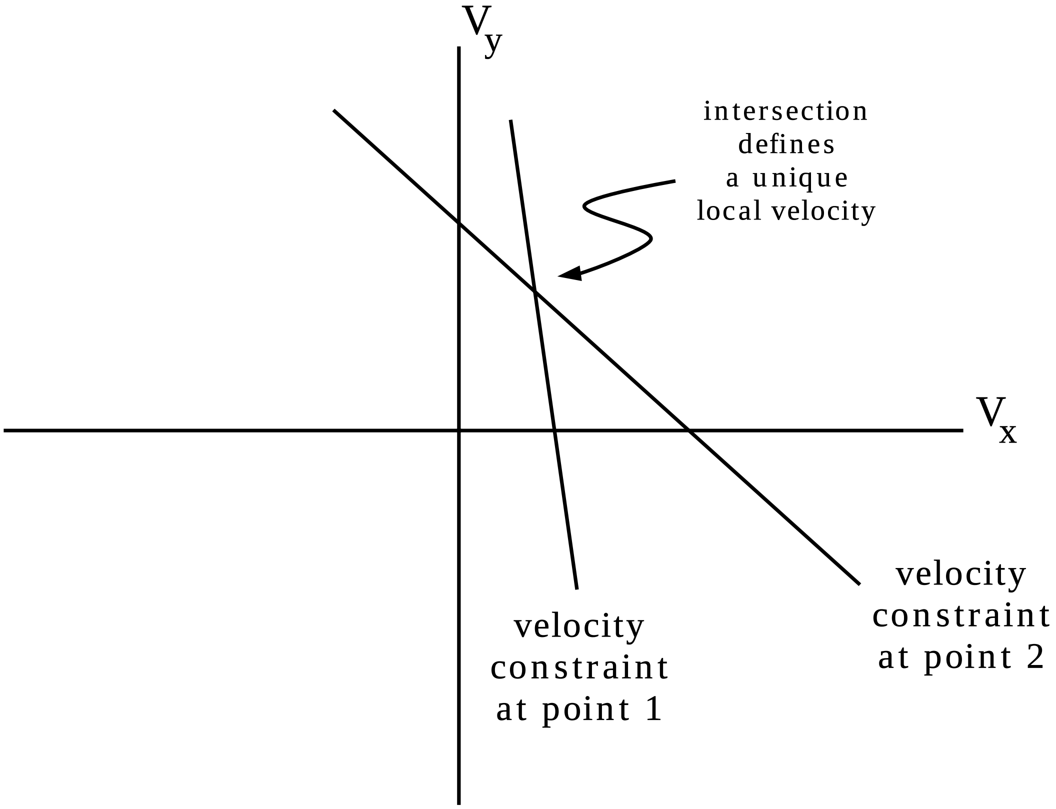
V_y
intersection defines a unique local velocity
V_x
velocity constraint at point 1
velocity constraint at point 2

Figure 1.8: *A graphical representation of the motion gradient constraint.* According to the motion gradient constraint, the spatial and temporal derivatives at each image point constrain the local velocities to fall somewhere along a line. The intersection of the constraint lines derived from nearby points yields a local velocity estimate that is consistent with the motion of all the local points.

(Remember that v_x and v_y depend on the spatial position and the moment in time, (a, b, t).)

Our goal is to use the changing image intensities to estimate the motion vector, (v_x, v_y), at each position. We expand the right hand side of Equation 1.1 in terms of the partial derivatives of the intensity pattern with respect to space and time,

$$I(a + v_x, b + v_y, t + 1) \approx I(a, b, t) + v_x \frac{\partial I}{\partial x} + v_y \frac{\partial I}{\partial x} + \frac{\partial I}{\partial t}. \tag{1.2}$$

The terms $\frac{\partial I}{\partial x}$, and $\frac{\partial I}{\partial y}$ are the partial derivatives of $I(x, y, t)$ in the spatial and temporal dimensions, respectively. Grouping Equations 1.1 and 1.2, we obtain the *gradient constraint equation*.

$$v_x \frac{\partial I}{\partial x} + v_y \frac{\partial I}{\partial y} + \frac{\partial I}{\partial t} \approx 0. \tag{1.3}$$

Equation 1.3 is a linear relationship between the space-time derivatives of the image intensities and the local velocity. Since there is only one linear equation and there are two unknown velocity components, the equation has multiple solutions. All of the solutions fall on a *velocity constraint line* shown in the graph in Figure 1.8.

Because the data at a single point do not define a unique solution, to derive an estimate we must combine the velocity constraint lines from a number of nearby points. There will be a unique solution, that is all of the lines will intersect in a single point, if (a) the motion gradient constraint is true, (b) nearby points share a common velocity, and (b) there is no noise in the measurements. If the lines do not come close to a single intersection point, then the motion gradient constraint is a poor local description of the image motion or the nearby points do not share a common motion. In the Appendix, I discuss some of the issues related to finding the best solution to the motion gradient constraint equations of multiple points in the presence of measurement noise.

Space-time filters and motion gradient

Early in this chapter, we found that we can use the response amplitudes of space-time oriented linear filters to measure local image velocity. Now, studying the motion gradient constraint, we find that we can combine the spatio-temporal derivatives of the image intensities to measure the local velocity. These two methods of measuring local motion are closely linked as the following argument shows (Simoncelli, 1993).

To be able to explain the relationship using pictures, let's consider only one-dimensional spatial stimuli and use the (t, x) graph. With this simplification, the motion gradient constraint has the reduced form

$$v_x \frac{\partial I}{\partial x} + \frac{\partial I}{\partial t} = 0. \tag{1.4}$$

Since the stimuli are one-dimensional, we can only estimate a single directional velocity, v_x. How can we express the computation in Equation 1.4 in terms of the responses of space-time receptive fields? To perform this computation, we need to compute the local spatial and temporal derivatives of the image. A receptive field that computes the spatial derivative can be computed in two steps. First, we compute a weighted average of the local image intensities over a small space-time region. We know that some space-time averaging of the image intensities must take place because of various unavoidable physiological factors, such as optical blurring and temporal sluggishness of the neural response. Suppose we describe this space-time averaging using a Gaussian weighting function $g(x, t)$. Second, we compute the spatial derivative by applying the spatial partial derivative operator to the averaged data. The space-time averaging followed by a partial derivative calculation can be grouped into a single operation, $\frac{\partial g}{\partial x}$. Similarly, we can express the temporal derivative operator as $\frac{\partial g}{\partial t}$. The space-time receptive fields that compute these two derivative operations are shown in (t, x) graphs at the top of Figure 1.9a.

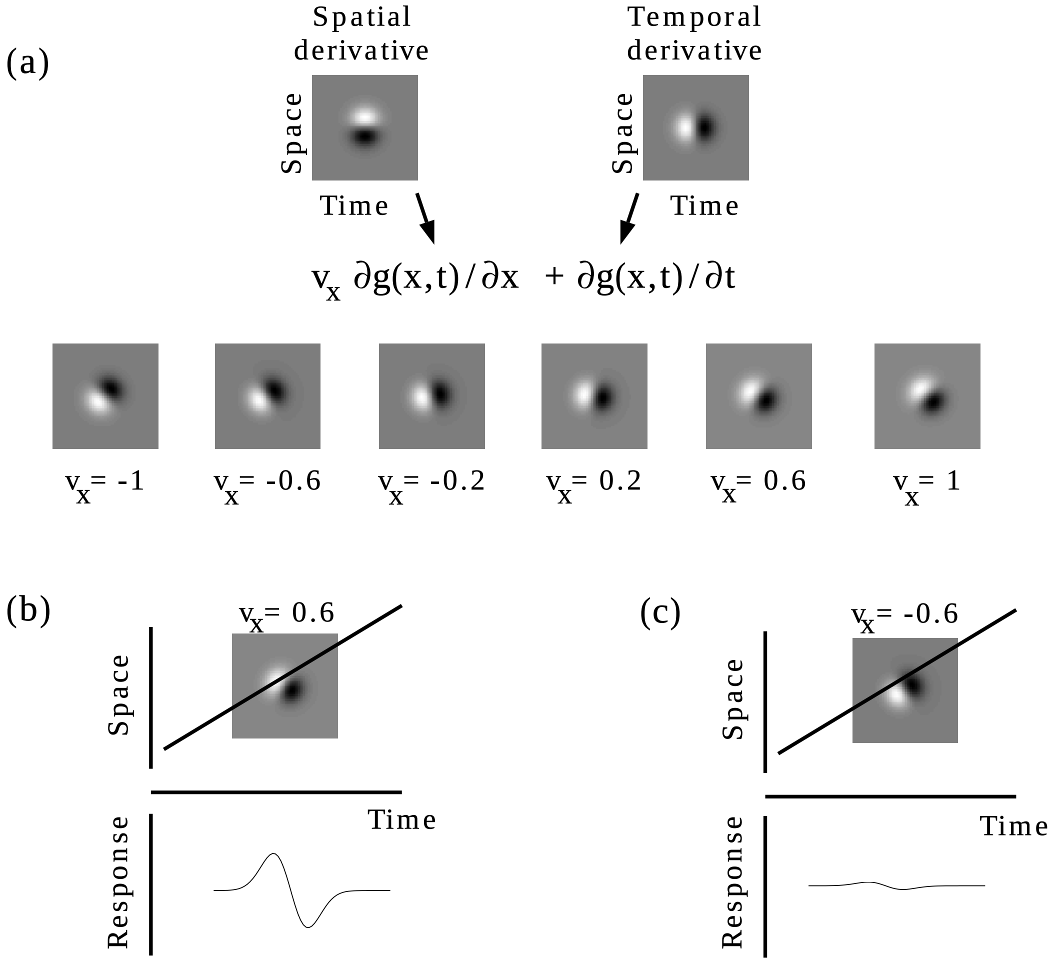

Figure 1.9: *The motion gradient constraint represented in terms of space-time receptive fields.* (a) The spatial and temporal derivatives can be computed using neurons whose (t, x) receptive fields are shown at the top. We can form weighted sums of these neural responses to create new receptive fields that are oriented in space-time. (b,c) The response amplitudes of these neurons can be used to identify the motion of a stimulus. The receptive field of the neuron represented in (b) responds strongly to the stimulus motion while the receptive field of the neuron in (c) responds weakly. By comparing the response amplitudes of the array of neurons, one can infer the stimulus motion.

Now, according to Equation 1.4, we should apply each of the two receptive fields to the image. The ratio of the responses is equal to the velocity, v_x. An equivalent way to perform this calculation is to create an array of neurons whose receptive fields are various weighted sum the two derivative operators for different values, v_x. The receptive fields of such an array of neurons is shown in Figure 1.9a. Each receptive field is oriented in space-time, and the orientation depends on the velocity used as a weight, v_x.

The pattern of response amplitudes of these neurons can be used to estimate the stimulus motion. For example, the neuron whose receptive field is shown in panel (b) has a receptive field that is aligned with the stimulus motion and has a large response amplitude. The neuron shown in panel (c) has a small response amplitude. We can deduce the local image velocity from the pattern of response amplitudes.

This set of pictures shows that the motion gradient constraint can be understood in terms of the responses of space-time oriented receptive fields. Hence, space-time oriented receptive fields and the motion gradient constraint are complementary ways of thinking about local motion.

Depth Information in the Motion Flow Field

We now have several ways of thinking about motion flow field estimation. But, remember that the motion flow field itself is not our main goal. Rather, we would like to be able to use the information in the motion flow field estimate to make inferences about the positions of objects in the scene. Much of the computational theory of motion and depth is concerned with how to calculate these quantities from the motion flow field. I do not provide a general review of such algorithms here. But, I do explain one principle concerning the computation of a depth map from observer motion, illustrated in Figure 1.10, that is important to understanding many of these algorithms (Longuet-Higgins and Prazdny, 1980; Koenderink and van Doorn, 1975).

We can partition each motion flow field vector into two components that are associated with different changes in the observer's viewpoint. The first component is due to a pure rotation of the viewpoint, and the second component is due to a pure translation of the viewpoint. Each motion flow field vector is the sum of a change due to the rotational and translational viewpoint changes.

These two flow field components contain different information about the distance of the point from the viewer. The change caused by a viewpoint *rotation* does not depend on the distance to the point. When the viewpoint rotates, the local flow field of all points, not matter what their distance from the observer, rotates by same amount. Hence, the rotational component of the motion flow field contains no information about the distance to different points in the image (Figure 1.10a).

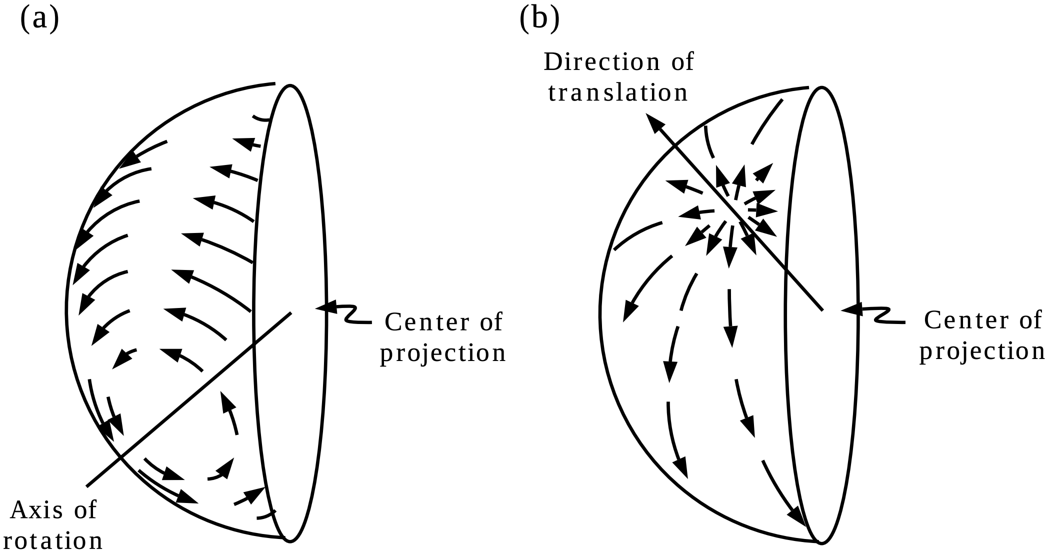

Figure 1.10: *The motion flow field components associated with observer motion.* A change in the observer's viewpoint causes two separate changes in the motion flow field. The total flow field vector is a sum of the rotation and translation components. (a) When the viewpoint rotates, the rotation component of the motion flow field is the same for all points, no matter their distance. (b) When the viewpoint translates, the motion flow vector the rotation component of the motion flow field varies with the image point distance from the observer and the direction of heading (After: Longuet-Higgins and Prazdny, 1980).

The flow field component caused by a *translation* of the viewpoint depends on the distance to the image point in two ways. First, along each line of sight the points closer to the viewpoint are displaced more than distant points. Second the direction of the local flow field depends on the direction of the translation (Figure 1.10b). You can demonstrate these effects for yourself by closing one eye and looking at a pair of objects, one near and one distant. As you move your head side-to-side, the image of the nearby point shifts more than the image of the distant point (i.e., motion parallax). Hence, the translational component of the motion flow field vectors contains information about the distance between the viewer and the point.

1.2 Experimental Observations of Motion

The main theoretical ideas we have reviewed concerning motion and depth each has an experimental counterpart. For example, there are behavioral studies that analyze the role of the motion gradient constraint in visual perception (Adelson and Movshon, 1982). And, the cat visual cortex contains neurons with space-time oriented receptive fields (McLean and Palmer 1994, DeAngelis, Ohzawa and Freeman, 1993).

In addition to confirmations of the importance of the computational ideas, the experimental literature on motion perception has provided new challenges for computational theories of motion perception. Most computational theories are based on measurements of image intensities and their derivatives. Experimental evidence suggests that motion perception depends on more abstract image features, such as surfaces or objects, as well. In Chapter **??** we saw that surfaces and illuminants are an important component of our analysis of color vision. Similarly, the experimental literature on motion perception shows us that we need to incorporate knowledge about surfaces and objects to frame more mature theories of motion perception (Ramachandran, et al., 1988; Stoner et al., 1990; Anderson and Nakayama 1994; Hildreth et al., 1995; Treue et al., 1995).

The experimental work defines new challenges and guidelines for those working on the next generation of computational theories. As we review these results, we shall see that motion perception is far from perfect; we make many incorrect judgments of velocity and direction. Moreover, perception of surfaces and occlusion are an integral part of how we interpret motion and depth. A complete computational theory of motion perception will need to include representations of surfaces and objects, as well as explanations of why image features such as contrast and color influence motion perception.

Motion gradients: The intersection of constraints

Adelson and Movshon (1982) studied how the some of the ideas of the motion gradient constraint apply to human perception. Their experiment is a motion superposition experiment that measures how people integrate motion information from separate image features when observers infer motion. The principle behind Adelson and Movshon's measurements is shown in Figure 1.11.

The motion of a one-dimensional pattern is ambiguous. We cannot measure whether a one-dimensional pattern, say a bar, has moved along its length. The graph in Figure 1.11a shows three images from an image sequence of a moving bar, and the dashed line in Figure 1.11b shows the set of possible velocities that are consistent with the image sequence. The graph is called a *velocity diagram*, and the set of possible motions are called the *constraint line*. The image sequence constrains the bar's horizontal velocity, but the data tell us nothing about the bar's vertical velocity. Although the information in the image sequence is ambiguous, subjects' perception of the motion is not ambiguous: the bar appears to move to the right. This perceived velocity is indicated on the velocity diagram by the black dot.

Figure 1.11cd shows the image sequence and constraint line of a horizontal bar. In this case, the stimulus information is only informative about the the vertical motion. This stimulus defines a different constraint line in the velocity diagram, and in this case, subjects see the line moving upward.

Figure 1.11ef shows the superposition of the two lines. In this stimulus, each bar separately determines a single constraint line; the *intersection of constraints* is the only point in the velocity diagram that is consistent with the image sequence information. The only consistent interpretation that groups the two lines into a single stimulus is to see the pair of lines moving up and to the right. This is what subjects see.

Adelson and Movshon (1982) described a set of measurements in which they evaluated whether observers generally saw the motion of the superimposed stimuli moving in the direction predicted by the intersection of constraints. In their original study, they used one-dimensional sinusoidal gratings as stimuli, rather than bars[5]. They altered the contrast, orientation and spatial frequency of the two gratings. As parameters varied, observers generally saw motion near the direction of the intersection of constraints. Often, however, observers saw the two gratings as two different objects, sliding over one another, an effect called *motion transparency*. Transparency is a case in which the observer perceives two motions at each point in the image. This fact alone has caused theorists to scurry back to their chalkboards and reconsider their computational motion algorithms.

The behavioral experiments show that calculations like the

[5] The superposition of two gratings at different orientations looks like a plaid, so this design is often called a *motion plaid* experiment.

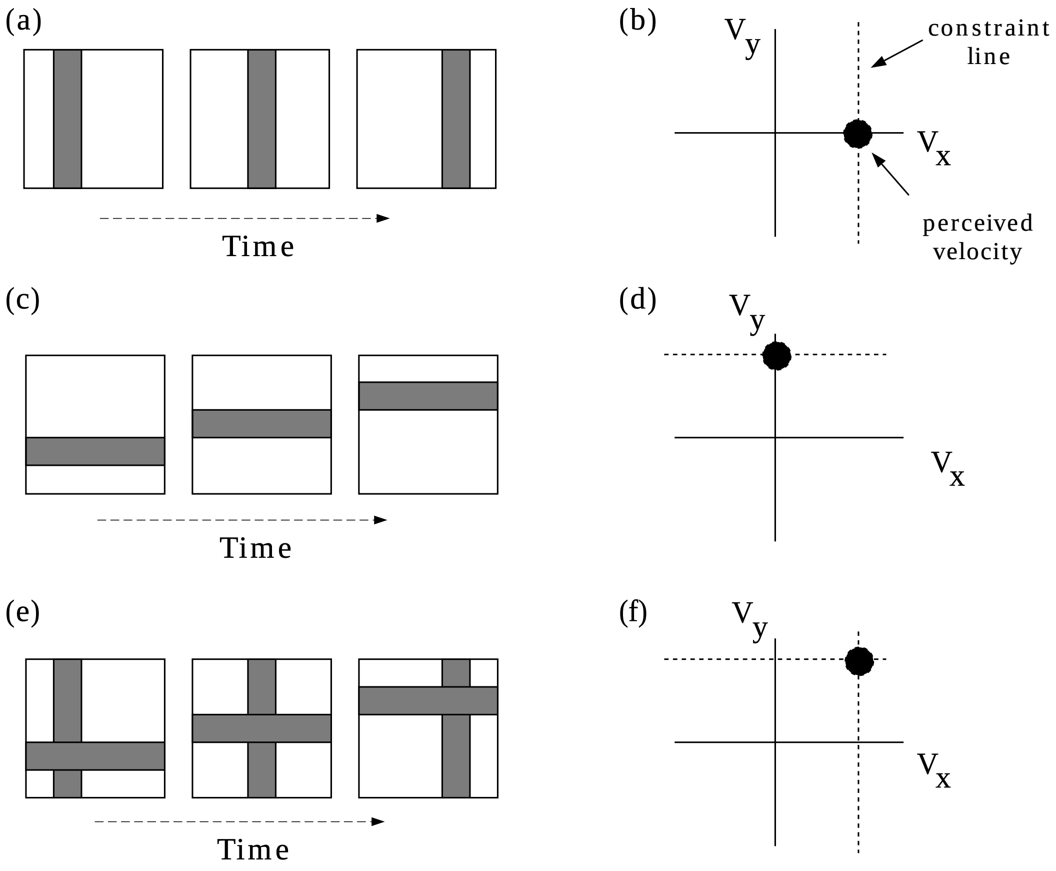

Figure 1.11: *The intersection of constraints.* The physical motion of a one-dimensional stimulus is inherently ambiguous. The physical motion in the display is consistent with a collection of possible velocities that plot as a line in velocity space. (a) A set of images of a vertical line moving to the right. (b) The set of physical motions consistent with the stimulus information plots along the velocity constraint line. The dot shows the perceived physical motion. (c,d) A similar set of graphs for a horizontal line. (e,f) When we superimpose the two lines, there is only a single physical motion that is consistent with the stimulus information. That motion plots at the intersection of the two velocity constraint lines.

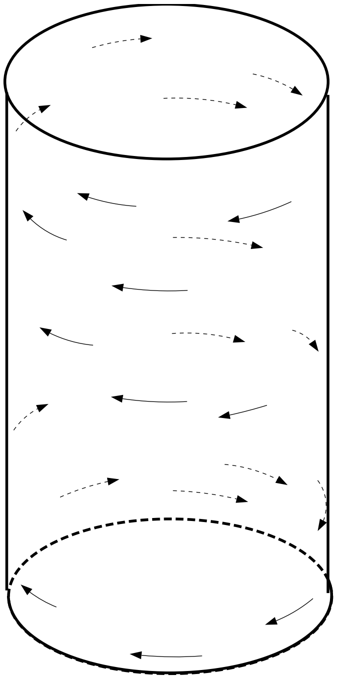

Figure 1.12: *Description of a random dot kinematogram.* Suppose that an observer views a random collection of dots, and that each dot is moving as if it is attached to the surface of a transparent cylinder. Observers perceive the surface of an implicit rotating cylinder, even though there is no shading or edge information in the image. Because the cylinder is transparent, dots move in both directions in each local region of the image. The dots are perceived as being attached to the near or far surface consistent with their direction of motion.

intersection-of-constraints are helpful in some cases. But, even in these simple mixture experiments observers see objects and surfaces, concepts that are not represented in the motion computations, are important elements of human motion perception. A second way to see the close relationship between motion an surfaces is through a visual demonstration, described in Figure 1.12, called a *random dot kinematogram*.

The demonstration described in Figure 1.12 consists of an image sequence in which each individual frame is an array of dots. From frame-to-frame, the dots change position as if they were painted on the surface of a moving object. In a single frame the observer sees nothing but a random dot pattern; the only information about the object is contained in the dot motions. The motions of the dots painted on the surface of a transparent cylinder are shown in Figure 1.12. The cylinder is also shown, though in the actual display the cylinder outline is not shown.

Random dot kinematograms reveal several surprising aspects of how the visual system uses motion information. First, the visual system seems to integrate local motions into globally coherent structures. The ability to integrate a set of seemingly independent local motions into a single coherent percept is called *structure from motion*. The demonstration described in Figure 1.12 is particularly impressive on this point. The image sequence contains a representation of dots on a transparent object. Because some of the dots are painted onto the front and some onto the back of the transparent object, each region of the image contains dots moving in opposite directions. Despite the apparent jumble of local motions, observers automatically segregate the local dot motions, and interpret the different directions and speeds, yielding the appearance of the front and back a rotating and transparent object.

Second, the ability to integrate these dot motions into an object seems to be carried out by a visual mechanism that infers the presence of a surface without being concerned about the stability of the surface texture. We can draw this conclusion from the fact that the stability and temporal properties of the individual dots has very little effect on the overall perception of the moving object. Single dots can be deleted after only a fraction of a second; new dots can be introduced at new surface locations on the implicit surface without disrupting the overall percept. Even as dots come and go, the observer sees a stable moving surface. The local space-time motions of the dots are important for revealing the object, but the object has a perceptual existence that is independent of any of the individual dots (Treue, 1991).

The shapes we perceive using random dot kinematograms are very compelling, but they do not look like real moving surfaces. The motion cue is sufficient to evoke the shape, but many visual cues are missing and the observer is plainly aware of their absence. Random dot kinematograms are significant because they seem to isolate certain pathways within the visual system. Random dot kinematograms may permit us to isolate the flow of information between specialized motion mechanisms and shape recognition. By studying motion and depth perception using these stimuli, we learn about special interconnections in the visual pathway. Studies with these types of stimuli have played a large role in the physiological analysis of the visual pathways, which we turn to next.

Contrast and Color

The color or contrast of an object is not a cue about the object's velocity. While velocity judgments should be independent of contrast and color, in fact perceived velocity depends on these stimulus variables. Models of human performance must be able to predict this dependence and to explain why judged velocity depends on these extraneous variables.

Stone and Thompson (1992) measured how perceived speed depends on stimulus contrast. Subjects compared the speed of a standard grating drifting at 2 deg/sec with the speed of a test grating. The data points measure the chance that the test

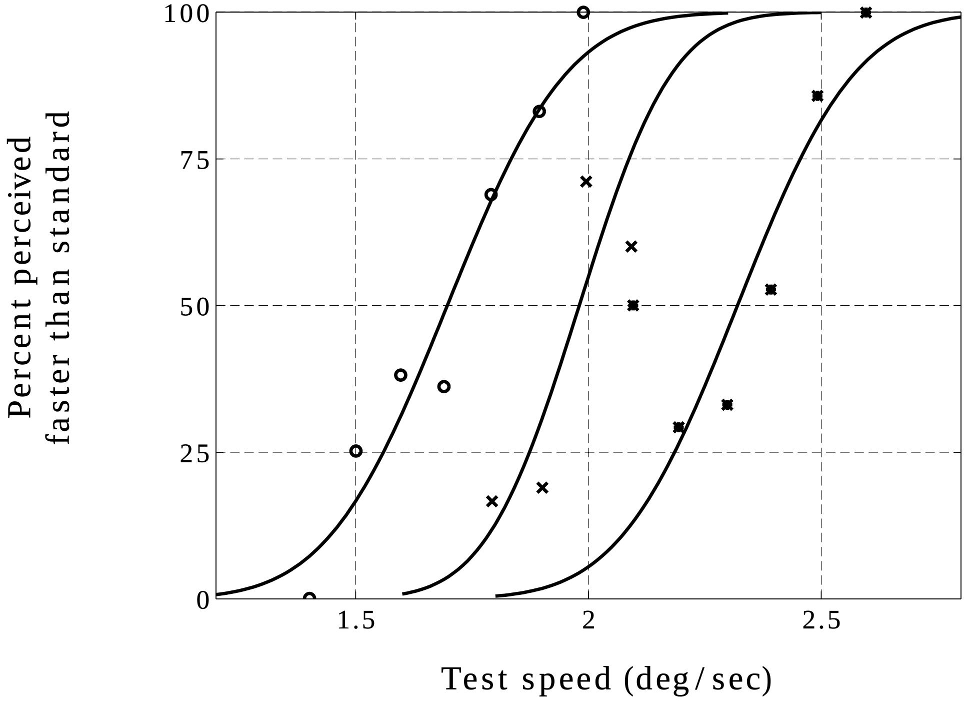

Figure 1.13: *Perceived speed depends on stimulus contrast.* The horizontal axis measures the velocity of a test grating. The vertical axis measures the probability that the test grating appears to be moving faster than a standard grating whose speed is always 2 deg/sec. The three separate curves show measurements for test gratings at three different contrasts. The curve on the left is for a test at 7 times the contrast of the standard; the curve on the middle is for a test at the same contrast as the standard; the curve on the right is for a test at one seventh the contrast of the standard. The spatial frequency of the test and standard were always 1.5 cpd. (Source: Stone and Thompson, 1992).

appeared faster than the standard as a function of the speed of the test grating. The data points near the curve in the middle of the figure are a control condition in which the test and standard had the same contrast. In the control condition, the test and standard had the same apparent speed when they had the same physical speed. The curve on the left shows the results when the test grating had seven times more contrast than the standard. In this case the test and standard had equal perceived speed when the test speed was 1.6 deg/sec, significantly slower than the standard. The data points near the curve on the right shows the results when the test grating had one-seventh the contrast of the standard grating. In this condition the test had equal perceived speed to the standard speed at 2.4 deg/sec, considerably faster than the standard. High contrast targets appear to move faster than low contrast targets.

Our velocity judgments also depend on other irrelevant image properties, such as the pattern of the stimulus (e.g. Smith and Edgars, 1991) and the color of the stimulus (e.g. Moreland, 1980; Cavanagh et al., 1984). Taken together these experiments suggest that some properties of the peripheral representation, intended to be helpful in representing color and form information, have unwanted side-effects on motion perception. The initial encoding of the signal by the visual system must be appropriate for many different kinds of visual information. Given these practical requirements, interactions between irrelevant image properties and motion estimation may be unavoidable.

Long and Short Range Motion Processes

Creative science often includes a clash between two opposing tendencies. The search to unify phenomena in a grand theory is opposed by the need to distinguish phenomena with separate root explanations. The tension is summarized in Einstein's famous remark: A theory should be as simple as possible, but no simpler. Perhaps the best known effort to classify motion into different processes is Braddick's (1974) classification into *short-range* and *long-range* motion processes.

Since Exner's (1875) original demonstrations, psychologists have known that even very coarsely sampled motion still generates a visual impression of motion. Such motion is easily distinguished from continuous motion, but there is no doubt that the something appears to be moving. The motion impression created by viewing coarsely sampled stimuli is called *apparent motion*[6]. The properties of apparent motion were studied by many of the Gestalt Psychologists, such as Wertheimer (1912) and Korte (1915), who describe the conditions under which the motion illusion is most compelling.

Braddick (1974) found that the spatial and temporal sampling necessary to perceive motion when using large spots is quite different from the limits using small displays

[6] This term is peculiar since all perceived motion is apparent.

with small spatial patterns, such as random dot kinematograms. Specifically, Braddick (1974) found that subjects perceive motion in a random dot kinematogram only when the spatial separations between frames are less than about 15 minutes of arc. When the spatial displacements are larger than this, observers see flickering dots but no coherent motion. This is a much smaller spatial limit than the sampling differences at which subjects report seeing apparent motion. The upper limit on the spatial displacement at which motion is perceived is called D_{max} in the motion literature.

Braddick also noted a second difference between the motion perceived in random dot kinematograms and apparent motion. Suppose we present an alternating pair of large spots to elicit apparent motion. We perceive the motion when the two spots are presented either to the same eye or when they are presented alternately to two eyes. Alternating frames of a random dot kinematogram between the two eyes, however, is less effective at evoking a sense of motion (Braddick, 1974). A wide range of experimental findings have been interpreted within this classification of motion (see e.g., Braddick, 1980; Anstis, 1980; Nakayama, 1984).

From our earlier analysis of temporal sampling of motion (see Figure 1.5) we learned that the ability to discriminate continuous from sampled motion will depend on the spatial and temporal properties of the image display. It is not too surprising, then, to find that the ability to perceive motion at all should depend on the spatiotemporal properties of the stimulus as well. Cavanagh and Mather (1989) review a set of measurements that suggest the difference between short- and long-range processes can be explained by the spatiotemporal organization of the visual system to stimuli of different spatial size and pattern, rather than by a subdivision within the visual pathways. For example, the value of D_{max} appears to be proportional to the size of the elements in the random-dot field for dot sizes larger than 15 min rather than an absolute value that divides the sensitivity of the two motion systems (Cavanagh et al., 1985).

Based on their review of a variety of experiments, Cavanagh and Mather conclude that the measurements that give rise to the definition of the long- and short-range processes are a consequence of differential visual sensitivity, not of motion classification. The long and short-range processes classification is still widely used, so understanding the classification is important. Because, I suspect that the classification will not last (see also, Chang and Julesz, 1985; Shadlen and Carney, 1986).

First and Second order motion

Several authors have proposed a classification of motion based on whether or not a stimulus activates neurons with space-time oriented receptive fields followed by simple squaring operations (Anstis, 1980; Chubb and Sperling, 1988). According to this classification, whenever a stimulus creates a powerful response for these types of sensors, the motion can be perceived by the *first-order* motion system. Chubb and Sperling (1988) show precisely how to create stimuli that are ineffective at stimulating the first-order system but that still appear to move. They propose that these moving stimuli are seen by a *second-order* motion system [7].

Figure 1.14 describes an example of a stimulus that is ineffective at stimulating space-time oriented filters and yet appears to move. At each moment in time, the stimulus consists of a single uniform texture pattern. The stimulus appears to contain a set of vertical boundaries because the local elements in different bands move in opposite (up / down) directions. In addition to this up / down motion, the positions of the bands drift from left-to-right. This stimulus does not evoke a powerful left-right response from space-time oriented filters. Yet, the pattern plainly appears to drift left-to-right (Cavanagh and Mather, 1989).

These second-order motion stimuli can also be interpreted as evidence that surfaces and objects play a role in human motion perception. The motion of the bars is easy to see because we see the bars as separate objects. We perceive these objects because of their local texture pattern, not because of any luminance variation. Computational theorists have not come to a consensus on how to represent objects and surfaces. Thus, the motion of these second-order stimuli cannot be easily explained by conventional theory. We might take the existence of these second-order stimuli, then, as a reminder that we need to extend current theory to incorporate a notion of perceived motions that includes concepts that connect local variations to broader ideas concerning surfaces and object (Fleet and Langley, 1994; Hildreth, et al. 1995).

Because we perceive the motion of borders, including borders defined by texture and depth, we must find ways to include signals derived from the outputs of texture in theories of motion perception. Cavanagh and Mather (1989) suggest that we might reformulate the classification into first and second-order processes as a different and broader question: Can motion perception can be explained by a single basic motion detection system that receives inputs from several types of visual subsystems, or are there multiple motion sensing systems each with its own separate input. There is no current consensus on this point. Some second-order phenomena can be understood by simple amendments to current theory (Fleet and Langley, 1994), But, phenomena involving transparency, depth, and occlusion may need significant new additions to the theory (Hildreth, et al., 1995). At present, then, I view the classification into first and second-order motion systems as a reminder that many different signals lead to motion perception, and that at present our analyses

[7] In their original paper, Chubb and Sperling (1988) called to the two putative neural pathways *Fourier* and *non-Fourier* motion systems. Cavanagh and Mather (1989) used the terms *first-* and *second-order* motion systems to refer to a broad class of similar motion phenomenona. This terminology seems to be gaining wider acceptance, and it has other advantages.

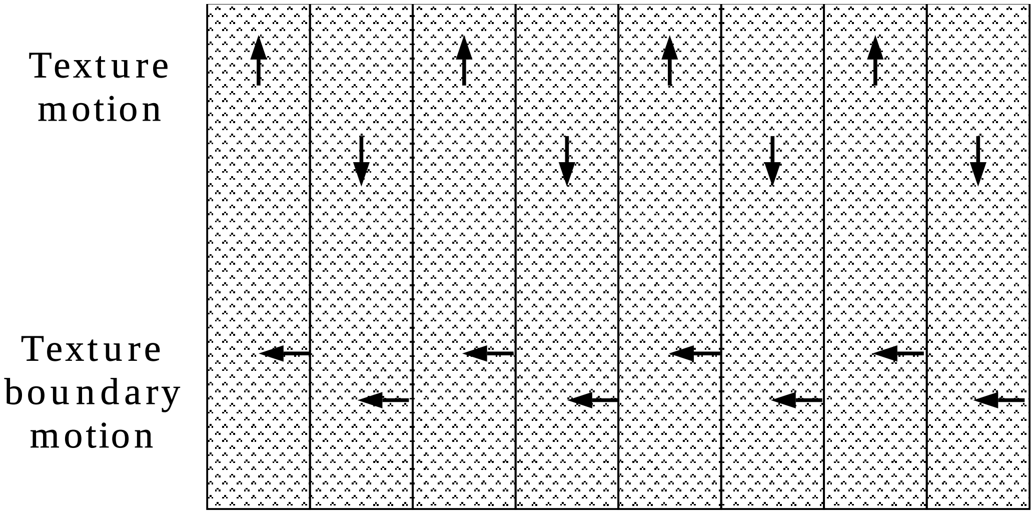

Figure 1.14: *A second-order motion stimulus.* The stimulus consists only of a set of texture patterns that are set into motion in two different ways. First, the texture in the separate bands moves up or down, alternately. The up and down motion of the texture bands defines a set of boundaries that are easily perceived even though there is no line separating the bands. (In the actual stimulus, there is no real edge to separate the texture bands. The lines are only drawn here to clarify the stimulus). The boundaries separating the texture bands moves continuously from right to left. This motion is also easily perceived. The leftward motion of the stimulus is very ineffective at creating a response from linear space-time oriented filters (Source: Cavanagh and Mather, 1989).

have only explored a few types of these signals.

1.3 Binocular Depth

> It will now be obvious why it is impossible for the artist to give a faithful representation of any near solid object, that is, to produce a painting which shall not be distinguished in the mind from the object itself. When the painting and the object are seen with both eyes, in the case of the painting two *similar* pictures are projected on the retinae, in the case of the solid object the two pictures are *dissimilar*; there is therefore an essential difference between the impressions on the organs of sensation in the two cases, and consequently between the perceptions formed in the mind; the painting therefore cannot be confounded with the solid object (Wheatstone, 1838, p. 66).

Wheatstone (1838) was the first to analyze thoroughly the implications of the simple but powerful fact that each eye images the world from a slightly different position. For near objects, the different perspective obtained by each eye provides us with an important cue to depth, namely *retinal disparity* (see Chapter **??**). The differences between a pair of stereo images, and the differences seen when the observer translates, have much in common. In the case of stereo depth we refer to the differences as retinal disparity, and in the case of motion sequence we refer to the differences as a motion flow field. In both cases the differences between the two images arise due to translation and rotations of the viewpoint associated with the different images.

Depth Without Edges

Just as there has been some debate concerning the role of surfaces and edges in motion perception, so too there has been a debate on the role of these different levels of representation in perceiving depth. Until the mid-1960s, psychologists generally supposed that the visual pathways first formed an estimate of surface and edge properties within each monocular image. It was assumed that disparity, and ultimately stereo depth, were calculated from the positions of the edge and surface locations estimated in each of the monocular images (Ogle, 1964).

Julesz (1960, 1971) introduced a stimulus, the *random dot stereogram*, that proves that an object can be seen in stereo depth even though we cannot perceive the object's edges are invisible in the monocular images. The random dot stereogram consists of a pair of related random dot patterns as shown in Figure 1.15. Each image seen separately appears to be a random collection of black and white dots. Yet, when the

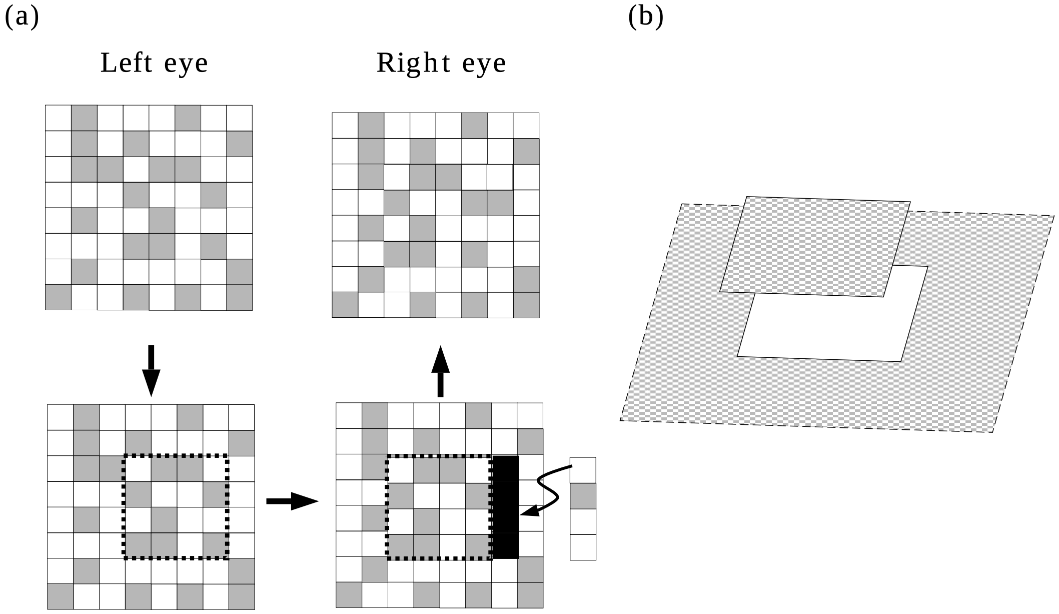

Figure 1.15: *Construction of a random dot stereogram.* (a) A random dot stereogram is created in a series of steps. First, a random dot pattern is created to present to, say, the left eye. The stimulus for the right eye is created by copying the first image, displacing a region horizontally, and then filling the gap with a random sample of dots. (b) When the right and left images are viewed simultaneously, the shifted region appears to be in a different depth plane from the other dots.

two images are presented separately to the two eyes, the relationship between the two collections of dots is detected by the visual pathways and the observer can perceive the surface of the object in depth.

Figure 1.15 shows how to create the two images comprising a random dot stereogram. First, create a sampling grid and randomly assign a black or white intensity to each position in the grid. This random image of black and white dots will be one image in the stereo pair. Next, select a region of the first image. Displace this region horizontally, over-writing the original dots. Displacing this region of dots leaves some unspecified positions; fill in these unspecified positions randomly with new black and white dots.

Random dot stereograms are a fascinating tool for vision science because the patterns we see in these stereo pairs are computed by signals that are carried separately by the two eyes. First, they demonstrate the simple but important point that even though we cannot see any monocular edge or surface information of the object, we can still see the object based on the disparity cue. Second, they provide an interesting tool for anatomically localizing different types of perceptual computations. Recall that the earliest binocular cells are in the superficial layers of area V1 (Chapter **??**). Hence, any part of the surface or edge computation that is performed in the monocular pathways prior to area V1 cannot play a role in the representation of edges and surfaces seen in random dot stereograms[8].

Depth With Edges

That observers perceive depth in random dot stereograms does not imply that edge detection or surface interpretation plays no role in depth perception. This is quite analogous to the experimental situation in motion perception. Observers perceive motion in random dot kinematograms, but surfaces and edge representations appear to be an important part of how we perceive motion.

We can see the relationship between surface representations and depth by considering the role of surface *occlusion*. Occlusion is one of the most powerful *monocular* cues to image depth since when one object blocks the view of another, it is a sure sign of their depth relationship. Shimojo and Nakayama (1990; He and Nakayama, 1994) have argued that occlusion relationships play an important role in judgments of *binocular* vision, too.

Their demonstrations of the role of occlusion in stereo depth are based on the simple physical observations shown in Figure 1.16a. Leonardo Da Vinci used this drawing in his *Trattato della Pittura* to describe the relationship between occlusion,

[8] Julesz (1971) calls the inference of anatomical localization from psychological study “psychoanatomy.”

half-occlusion, and transparency.

> if an object C be viewed by a single eye at A, all objects in the space behind it ... are invisible to the eye at A; but when the other eye at B is opened, part of these objects become visible to it; those only being hid from both eyes that are included ... in the double shadow CD cast by two lights at A and B. ... [Because] the angular space EDG beyond D being always visible to both eyes ... the object C seen with both eyes becomes, as it were, transparent, according to the usual definition of a transparent thing; namely, that which hides nothing beyond it. (Leonardo Da Vinci, 1906).

Da Vinci points out that when an observer looks at an object, each eye encodes a portion of the scene that is not encoded by the other eye. These are called *half-occluded* regions of the image (Belhumeur and Mumford, 1992). When one looks beyond a small object, there is a small region that is fully occluded, and another region that both eyes can see. When examining points in this furthest region, the small object is, effectively, transparent.

There are several simple rules that describe the location and properties of the half-occluded image regions. First, half-occluded regions seen by the left eye are always at the left edge of the near object, while half-occluded regions seen by the right eye are always at the right edge of the near object. The other two local possibilities (left eye sees an occlusion at the right edge, right eye sees an occluded region at the left edge) are physically impossible (see Figure 1.16b).

Second, the relative size of the half-occluded regions varies systematically with the distance of the near and far objects. As the near object is placed closer to the observer, the half-occluded region becomes larger(see Figure 1.16c). Hence, both the position and the size of half-occluded regions contain information that can be used to infer depth[9].

Shimojo and Nakayama (1990) found that surface occlusion information influences observers judgment of binocular depth. In their experiments, observers viewed a stereogram with physically unrealizable half-occlusions. They found that when the half-occlusion was, say, a pattern seen only by the right eye but near the left edge of a near object, observers suppressed the visibility of the pattern. When they presented the same pattern at the right edge of the near object, where it could arise in natural viewing, the pattern was seen easily. Anderson and Nakayama (1994)

[9] Random dot stereograms contains two half-occluded regions that are usually consistent with the image depth. When we displace the test region, we overwrite a portion of the original random dot image. The overwritten dots are half-occluded because they are only seen by the eye that views the first image. The dots that are added to complete the second image are half-occluded because they are only seen by the eye that views that second image.

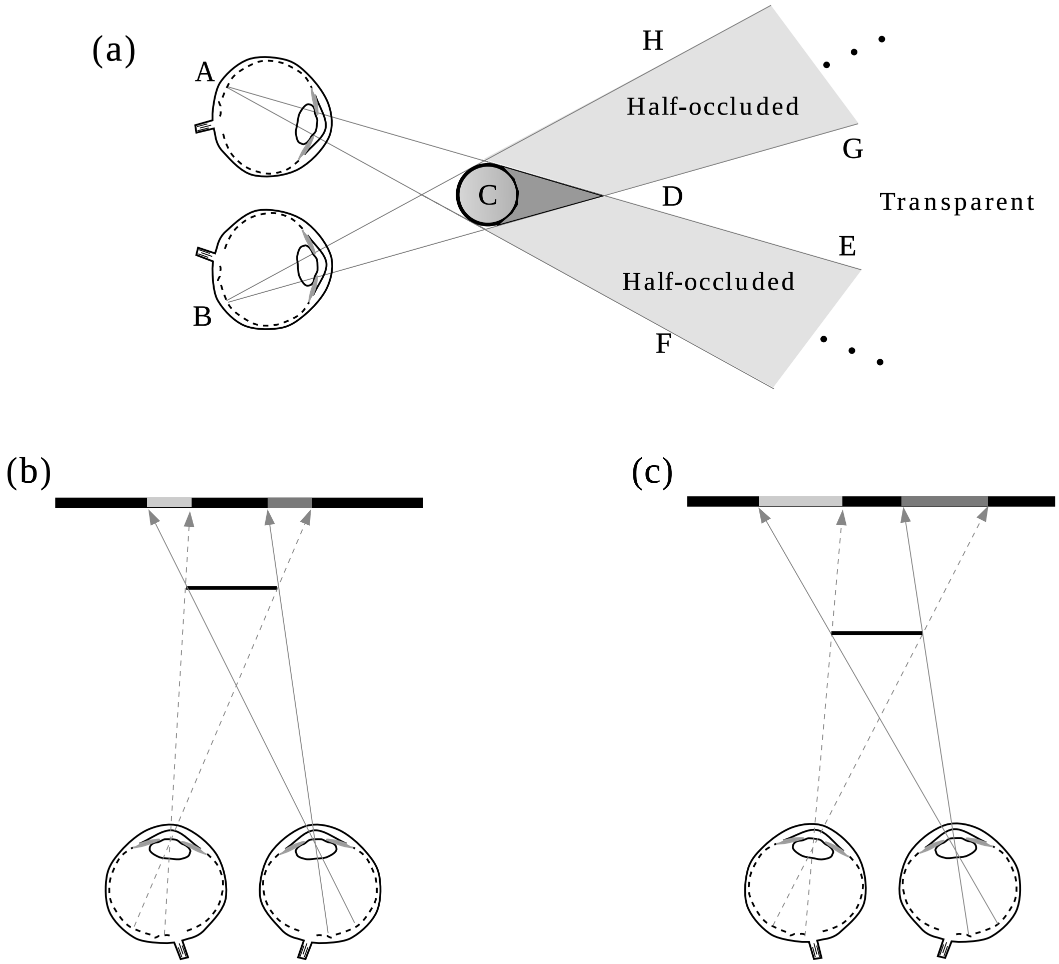

Figure 1.16: *Half-occluded regions.* In normal viewing, there will be regions of the image that are seen by both eyes, neither eye, or only one eye. (a) When viewing a small, nearby object, there is a fully occluded region just beyond the object (dark shading). There are a pair of half-occluded regions (light gray-shading). Well beyond the small object, both eyes see the image so that the object is, effectively, transparent (After: Da Vinci, 1906) (b) The half-occluded regions seen by the right eye fall near the right edge of the near object, while the half-occluded regions seen by the left eye fall near the left edge of the near object. (c) The size of the half-occluded region depends on the distance between the observer, the near object, and the far object.

summarize a number of related observations, and they conclude that occlusion configurations, that is a property of surfaces and objects, influence the earliest stages of stereo matching.

Perceived depth, like motion and color, is a visual inference. These results show that the visual inference of depth depends on a fundamental property of surfaces and objects, namely that they can occlude one another.

1.4 Head and Eye Movements

Our eyes frequently move, smoothly tracking objects or jumping large distances as we shift our attention. To judge the motion of objects in the image, it is essential for the visual pathways to distinguish motion present in the image from motion due to eye movements.

Helmholtz (1865) distinguishes several ways the visual system might incorporate information about eye position into judgments of motion. When we move our eyes, the motor system must generate a signal that is directed from the central nervous system to the eye muscles. This outgoing signal is one potential source of information about eye movements. Helmholtz referred to this signals as denoting the *effort of will*. He reasoned that a copy of this motor signal may be sent to the brain centers responsible for motion perception, and that this willful signal may be combined with retinal signals to estimate motion. This hypothetical signal is called the *corollary discharge*.

A second possible source of information are nerve cells that are attached to the muscles that control eye movements. Neural sensors may measure the tension on the muscles, or the force exerted by the muscles, and the responses of these sensors may be sent to the brain centers responsible for motion perception. These are incoming sources of information, so that we can distinguish these two theories with the names *outflow theory* and *inflow theory* (Gregory, 1990).

Helmholtz lists several simple experimental demonstrations in favor of outflow theory. First, when we rotate the eye by pushing on it with our finger, the world appears to move. In this case, the retinal image moves but there is no willful effort to rotate the eye. According to outflow theory we should interpret the field as rotating, and it does. Second, if we create a stabilized retinal image, say by creating an afterimage, rotating the eyeball does not make the afterimage appear to move. In this case there is no retinal image motion, and no willful effort to rotate the eye. Hence, no motion is expected. Third, Helmholtz finds support for the outflow theory in the experience of patients whose eye muscles have been paralyzed. He writes

> ... in those cases where certain muscles have suddenly been paralyzed,

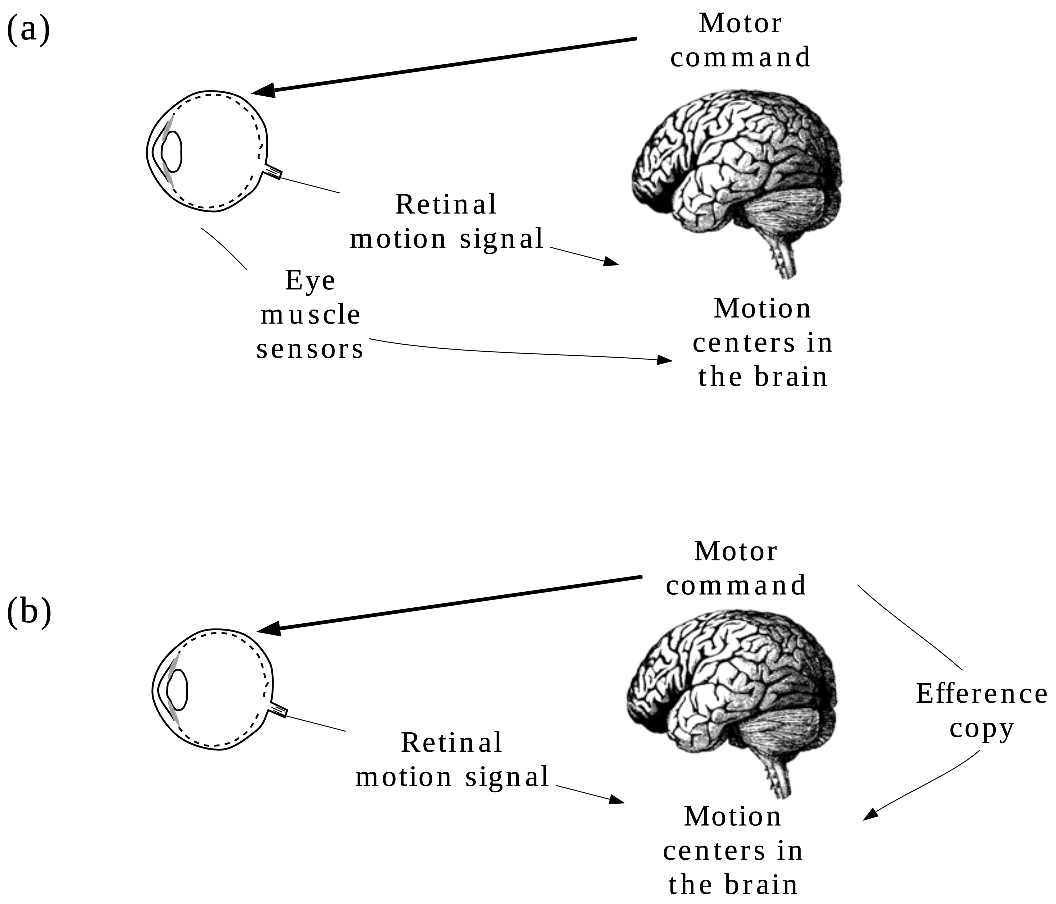

Figure 1.17: *Inflow and Outflow theories for discounting eye movement.* (a) According to inflow theory, signals from the retina and the muscles controlling eye movement are at the motion centers in the brain. By comparing these two signals, the motion centers discount eye movements and infer object motion. (b) According to outflow theory, signals from the retina and an efference copy of the motor signal are sent to motion centers in the brain. By comparing these two signals, the motion centers discount eye movements and infer object motion.

> when the patient tries to turn his eye in a direction in which it's powerless to move any longer, apparent motions are seen, producing double images if the other eye happens to be open at the time (Helmholtz, 1865, p. 245).

There is good support, then, for the basic outflow theory. This raises a second interesting question concerning how the visual system infers the changing position of the eye. The nervous system has two types of information about eye position. One type of information is based on the retinal image and computed by the motion flow field. As I reviewed earlier in this chapter and in the Appendix, it is possible to estimate the observer's motion from the motion flow field. Now, we find that it is also possible to estimate the motion of the eye from an efferent signal from the motor pathways. Which one does the visual system use?

The answer seems to be very sensible. There are times when the information about eye position from the motor system is more reliable than information from the motion flow field. Conversely, sometimes motion flow information is more reliable. Human experimental measurements suggest that under some conditions observers use motion flow information information alone to estimate heading; under other conditions extra-retinal signals, presumably from the oculomotor pathways, are combined with motion flow signals (Warren and Hannon, 1988; Royden and Banks, 1992).

Vision during saccadic eye movements

There is an interesting and extreme case in which the oculomotor system dominates retinal signals you can observe yourself. First, find a small mirror and a friend to help you. Ask your friend to hold the mirror close to his or her eyes. Then, have your friend switch gaze between the left and right eye repeatedly. As your friend shifts gaze, you will see both eyes move. Then, change roles. While your friend watches you, shift your gaze in the mirror from eye to eye. Your friend will see your eyes shift, but you will not be able to see your own eyes move. As your eyes saccade back and forth and watch in the mirror, you cannot see your own eyes move at all.

There have been several different measurements of sensitivity loss during saccades. To measure the loss of visual sensitivity one needs to separate out the visual effects from the simple effects having to do with the motion of the eye itself. The motion of the eye in, say, the horizontal direction changes the effective spatiotemporal signal in the direction of motion. By measuring sensitivity using horizontal contrast patterns, however, one can separate out the effect of the eye movement on the signal from the suppression by the visual pathway. The suppressive effects caused by neural, rather than optical, factors is called *saccadic suppression*.

Sensitivity loss during saccades shows two main features that relate to the motion

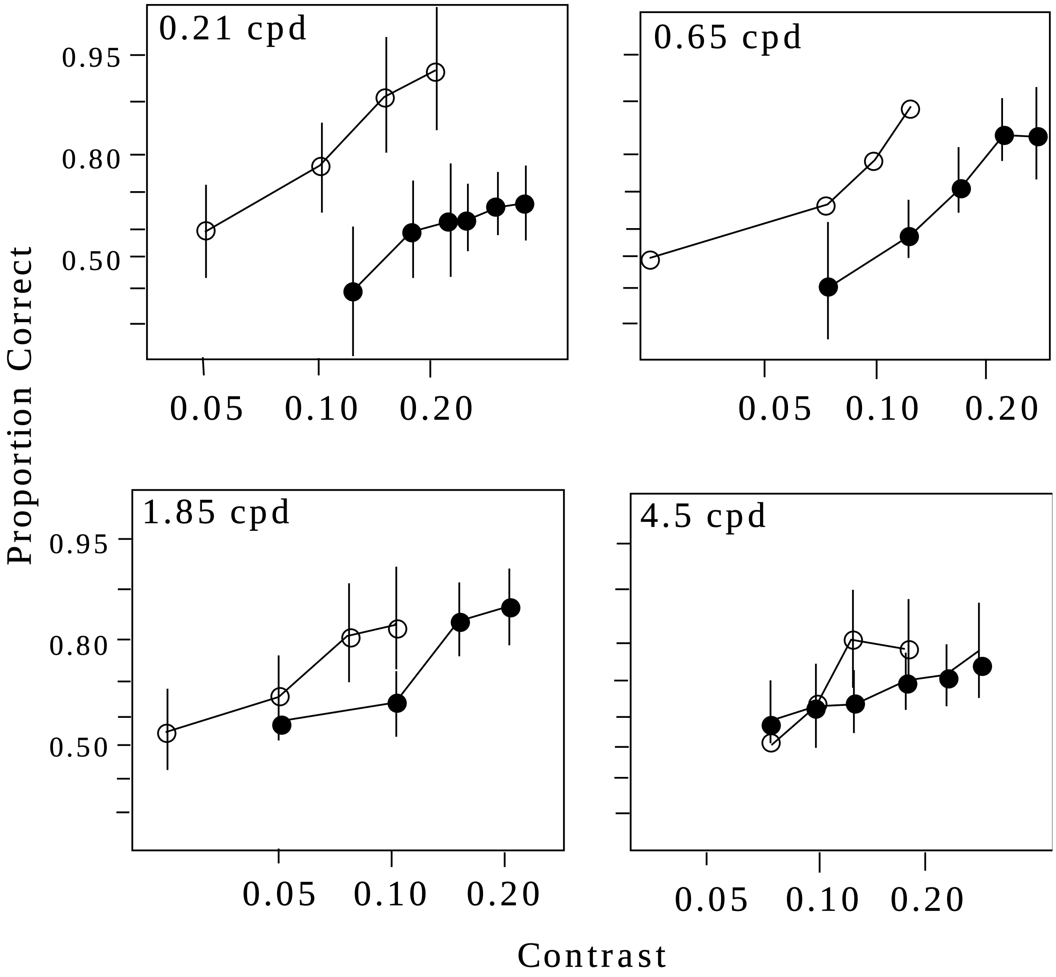

Figure 1.18: *Contrast sensitivity is suppressed during a saccade.* Each panel plots the probability of detection as a function of signal contrast while the observer is fixating (open symbols) or executing a saccade (filled symbols). The data are combined from three observers and the vertical lines through the points represent 95% confidence limits. The separate panels show measurements using patterns with different spatial frequencies. (Source: Volkman et al., 1978).

pathway. First, during saccades contrast sensitivity to low frequency light-dark patterns is reduced strongly, while sensitivity to high spatial frequency patterns is not significantly changed (Volkman, 1978; Burr et al, 1994). Second, there is no sensitivity loss to colored edges during a saccade (Burr et al, 1994).

The curves in Figure 1.18 measure probability of detecting a luminance grating as a function of the target contrast. The filled circles show measurements made during steady fixation and the open circles show measurement made during a 6 deg saccade. To see the target during the saccade, subjects need to increase the target contrast. Plainly, the suppression is strongest for the low spatial frequency targets. Suppression acts mainly on low spatial frequency targets, and there is little suppression of colored patterns. Lesion studies described in Chapter **??** suggested that these stimuli are detected mainly by signals on the M-pathway, a component of the motion pathway. Based on the parallel loss of visual sensitivity from these lesions and during saccadic eye movements, Burr et al., (1994) suggested that saccadic suppression takes place within the motion pathway.

The oculomotor and motion systems must work together to establish a visual frame of reference. Saccadic suppression illustrates that in some cases the visual system judges the retinal information to be unreliable and suppresses the retinal signal until a more stable estimate of the reference frame can be obtained. But, as we turn our head, the world remains visible and observers can detect image displacements as small as two or three percent. Hence, although we suppress quite large displacements during saccades, we remain sensitive to displacements as we turn our heads or move about (Wallach, 1987).

1.5 The Cortical Basis of Motion Perception

More than any other visual sensation, motion seems to be associated with a discrete visual portion of the visual pathway. A visual stream that begins with the parasol cells in the retina and continues through cortical areas V1 and MT seems to have a special role in representing motion signals. This *motion pathway*[10] has been studied more extensively than any other portion of the visual pathways, and so most of this section of the chapter is devoted to a review of the responses of neurons in the visual stream from the parasol cells to area MT.

Before turning to the physiological literature, I will describe an interesting clinical report of a patient who cannot see motion.

[10] While I will call this visual pathway a “motion pathway,” following common usage, the certainty implied by the phrase is premature. Other portions of the visual pathways may also be important for motion, and this pathway may have functions beyond motion perception.

Acquired Motion Deficits

Zihl et al. (1983) described a patient (LM) who, following a stroke, had great difficulty in perceiving certain types of motion. There have been a few reports of individuals with a diminished ability to perceive motion as a consequence of stroke, and transient loss of motion perception can even be induced by magnetic stimulation of the brain. But Zihl's patient, LM, has been studied far more thoroughly than the others and so we will focus on her case (Beckers et al, 1992; Vaina et al., 1990; Zeki, 1991).

LM's color vision and acuity remain normal, and she has no difficulty in recognizing faces or objects. But, she can't see the coffee flowing into a cup. Instead, the liquid appears frozen, like a glacier. Since she cannot perceive the fluid rising, she spills while pouring. LM feels uncomfortable in a room with several moving people, or on a street, since she cannot track changes in positions, "people were suddenly here or there but I have not seen them moving." She can't cross the street for fear of being struck by a moving car. "When I'm looking at the car first, it seems far away. But then, when I want to cross the road, suddenly the car is very near."

There are very few patients with a specific motion loss and so few generalizations are possible[11]. Patient LM succeeds at certain motion tasks but fail at others. Patient LM has a difficult time segregating moving dots from stationary dots, or segregating moving dots from a background of randomly moving dots. Patient LM has no difficulty with stereo. (Zihl, 1983; Hess et al., 1989; Baker et al., 1991)

Patient LM has a lesion that extends over a substantial region of visual cortex, so that this case does not localize sharply the regions of visual cortex that are relevant to her defects. However, it is quite surprising that the loss of motion perception can be dissociated from other visual abilities, such as color and pattern. This observation supports the general view that motion signals are represented on a special motion pathway. To consider the nature of the neural representation of motion further, we turn to experimental studies.

The motion pathway

The starting point for our current understanding of the motion pathway is Zeki's discovery that the preponderance of neurons in cortical area MT are *direction selective*; they respond vigorously when an object or a field of random dots move in one direction, and they are silent when the motion is in a different direction. These neurons are relatively unselective for other aspects of the visual stimulus, such as color or orientation (Dubner and Zeki, 1971; Zeki, 1974).

[11] Zeki (1991) calls the syndrome *akinetopsia*. His review makes clear that the evidence for the existence of this syndrome is much weaker than the evidence for dyschromatopsia.

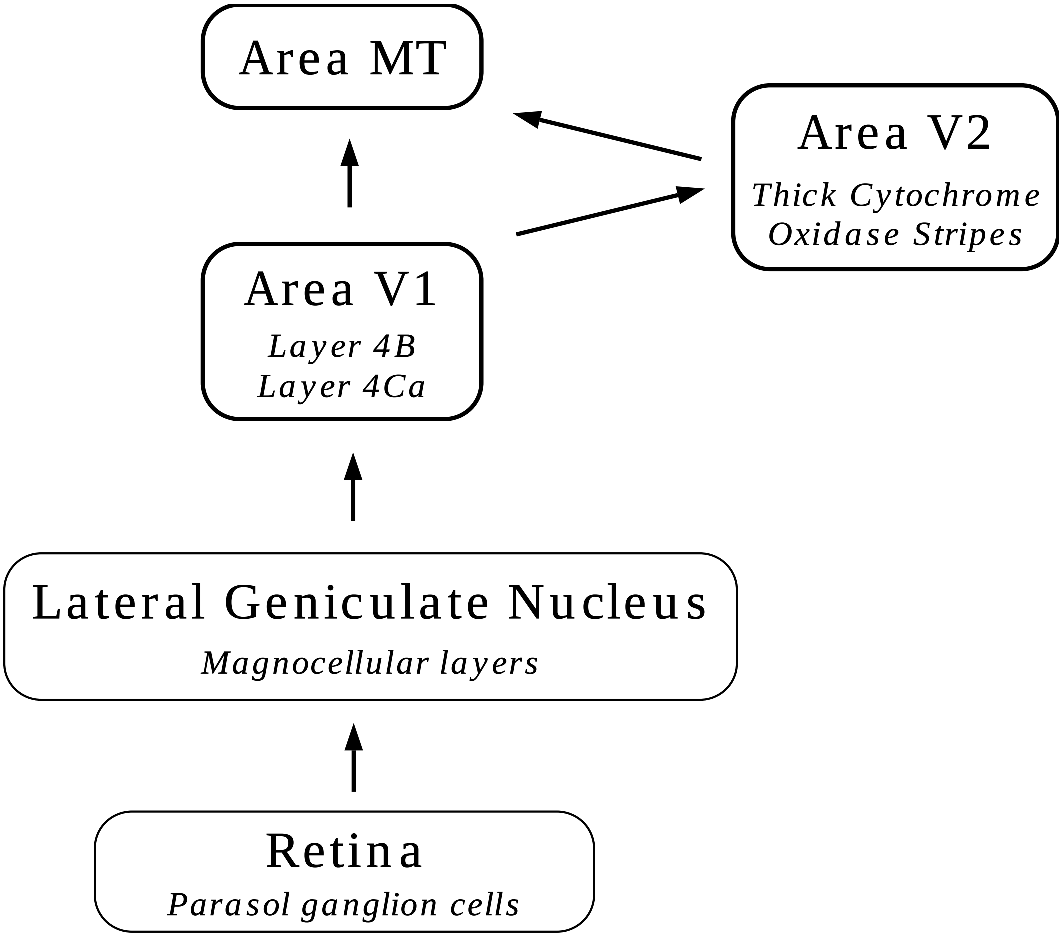

Figure 1.19: *Anatomy of the motion pathway.* The signal from parasol ganglion cells follow a pathway a discrete pathway into the brain. Their signals can be traced through the parvocellular layers of the lateral geniculate nucleus and within area V1 to area MT. Neurons in area MT respond more strongly to parasol than midget ganglion cell signals.

In other visual areas prior to MT, direction selective neurons represent only a fraction of the population. For example, about one-quarter of the neurons in area V1 are direction selective, and these neurons fall within a subset of the cortical layers in V1 (Hawken et al., 1988) . The proportion of direction-selective neurons appears to be even lower in area V2. Area MT appears to be the first area in which the vast majority of neurons, distributed throughout the anatomical area, are direction selective.

The neurons in area MT are principally driven by signals originating in the magnocellular pathway (see Figure 1.19). Recall from Chapters **??** and **??** that the magnocellular pathway terminates in layer 4Ca of primary visual cortex The output from 4Ca passes to layer 4B, and the output from 4B is communicated either directly to area MT, or first through regions within area V2 and then to area MT. The majority of direction selective neurons in area V1 fall within the same layers of V1 that communicate with area MT. Hence, the direction selective neurons in area V1 appear to send their output mainly to area MT[12].

There is one further piece of evidence concerning the significance of motion and area MT. The direction-selectivity of neurons within area MT is laid out in an organized fashion. Nearby neurons tend to be selective for motion in the same direction (Albright, 1984). This is analogous to the retinotopic organization evident in cortical areas V1 and V2 (see Chapter **??**). Taken together, the evidence argues that area MT plays an important role in motion perception (Merigan and Maunsell, 1993).

As we measure from the periphery to visual cortex, we find that the receptive field properties of the neurons within the motion pathway respond to increasingly sophisticated stimulus properties. The first major transformation is direction selectivity, which appears within neurons in layer 4B of area V1. Direction selectivity is a new feature of the receptive field, a feature which is not present in the earlier parts of the pathway.

Earlier in this chapter we saw that it is possible to estimate motion flow fields using neurons with receptive fields that are oriented in space-time (Figure 1.7). DeAngelis et al. (1993; see also Mclean and Palmer, 1994) measured the space-time receptive fields of neurons in cat visual cortex, and they found that some direction selective neurons have linear space-time oriented receptive fields. Some of their measurements are illustrated in Figure 1.20. The sequence of images in that figure shows the two-dimensional spatial receptive field of a neuron measured at different moments in time following the stimulus. Below the volume of measurements is the space-time receptive field for one-dimensional stimulation, shown in the (t, x) representation. This receptive field is also shown at the right where it is easy to see that the receptive field is oriented in the space-time plot.

Movshon et al. (1985) discovered a new feature of the receptive fields in some MT neurons that represents an additional property of motion analysis. They call the neurons that have this new receptive field property *pattern-selective* neurons to distinguish them from simple direction-selective neurons, such as we find in area V1, that they call *component-selective*. They identified these two neuronal classes in area MT using a simple mixture experiment.

First they measured the direction-selective tuning curves of neurons in area MT using one-dimensional sinusoidal grating patterns. Figure 1.21 shows the tuning curves of two MT neurons. In these polar plots, the neuron's response to a stimulus is plotted in the same direction from the origin as the stimulus motion. A point's

[12] The parvocellular pathway does make some contribution to MT responses. This has been shown by blocking magnocellular pathway responses and observing responses to signals in the parvocellular pathway. But, by biological standards the separation is rather impressive (Maunsell et al., 1990).

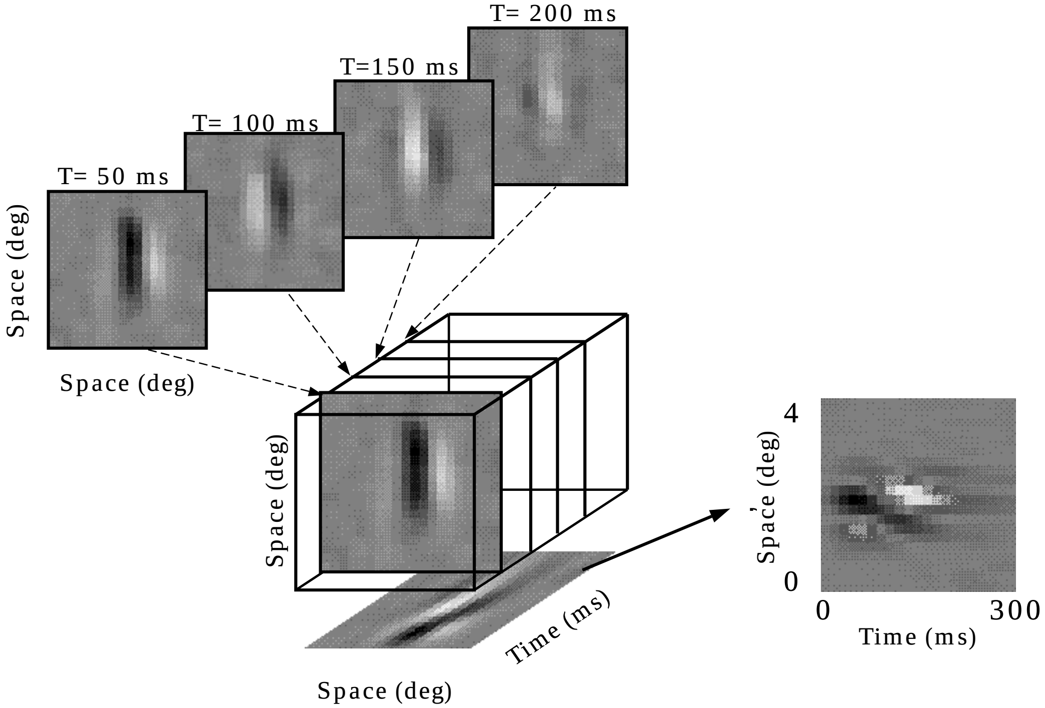

Figure 1.20: *Space-time oriented receptive field in cat cortex.* The images on the upper left show the spatial receptive field measured at different moments in time following stimulation. Taken together, these measurements form a space-time volume representation of the neural receptive field. The space-time receptive field for one-dimensional spatial stimulation is shown at the bottom of the volume and again on the right. In this (t, x) representation, the receptive field is oriented, implying that the neuron has a larger amplitude response to stimuli moving in some directions than others (After: DeAngelis, Ohzawa and Freeman, 1993).

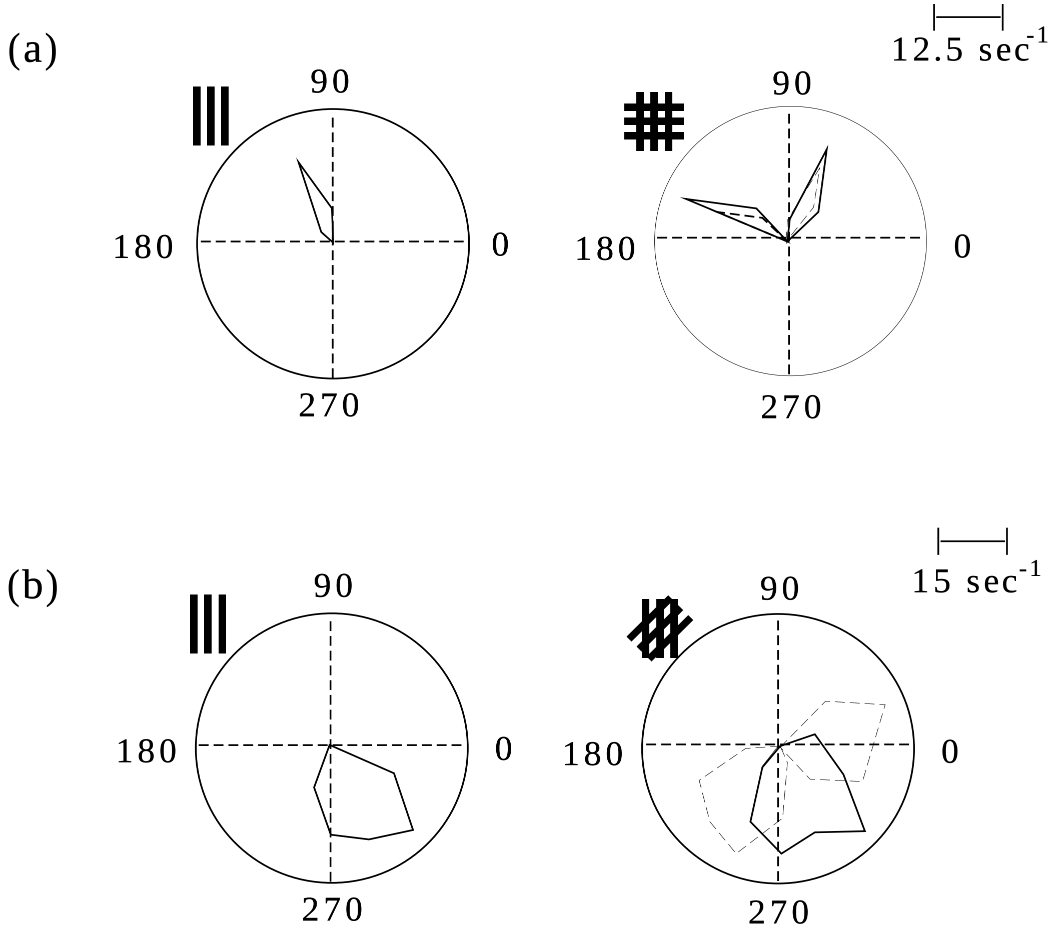

Figure 1.21: *Direction selectivity in area MT.* (a) The direction tuning of a component-selective neuron in area MT. The neuron responds to a grating moving up and to the left, but not to a plaid moving in the same direction. Instead, the neuron's responses to the plaid are predicted by its direction selectivity to the components of the pattern. The predicted responses, based on the components, are shown as dashed lines. This response pattern is typical of direction selective cells in area V1 and about half of the cells in area MT. (b) The direction tuning of a pattern-selective neuron in area MT. This neuron responded well to single gratings moving down and to the right. The cell also responded well to a plaid, whose components were separated by 135 deg, moving down and to the right. Neither component of the plaid alone evokes a response from this neuron. Hence, this neuron responds to the direction of motion of the pattern, not to the direction of motion of the components.

distance from the origin indicates the size of the neuron's response. The inner circle on the plot indicates the neuron's spontaneous firing rate. The direction-selective tuning curves of these neurons are similar to the tuning curves of neurons in area V1.

Having measured the tuning curve, Movshon and his colleagues asked the basic linear systems question: can we use the tuning curve to predict the neuron's response to other patterns? To answer this question, they used new patterns formed by adding together individual grating patterns.

Consider the response of a *component-selective* MT neuron, shown on the top of Figure 1.21. This neuron responded well only to a narrow range of directions of a sinusoidal pattern, upward and to the left. Movshon and his colleagues measured the neuron's response to a plaid consisting of components separated in orientation by 90 degrees. This MT neuron responds well to the plaid stimulus whenever one of the plaid components moves upward and to the left. But, the neuron did not respond well when the pattern as a whole was moving upward and to the left since in that case neither of the plaid components is moving up and to the left.

Recall that when people view a moving plaid, it appears to move approximately in the direction predicted by the intersection of constraints (see Figure 1.11). The component-selective neuron's activity does not correlate with the perceived direction of motion. The component-selective neuron responds well only when the individual components appear to be moving upward and to the left. It does not respond well when the plaid appears to move in this direction.

The response of a pattern-selective neuron, shown on the bottom of the Figure, does correlate with the perceived direction of motion. The pattern-selective neuron's response is large when a single sinusoidal grating moves down and to the right. The neuron also responded well to a 135 degree plaid pattern moving down and to the right. Remember, when the plaid is moving down and to the right, the plaid components are moving in directions that are outside of the response range of this neuron. If we isolate the components of the 135 motion plaid moving down and to the right and present them to the neuron, neither will evoke a response. Yet, when we present them together they result in a powerful response from the neuron. The nonlinear neuron response correlates with the perceived direction of motion of the stimulus.

In their survey of area MT, Movshon and his colleagues found that approximately 25% of the neurons in area MT were pattern-selective, half the neurons were classified as component-selective and the rest could not be classified. Having understood the signal transformation, we are now poised to understand the circuitry that implements the transformation. We would like to understand how pattern-selectivity is implemented, just as we now understand how direction-selectivity can be implemented.

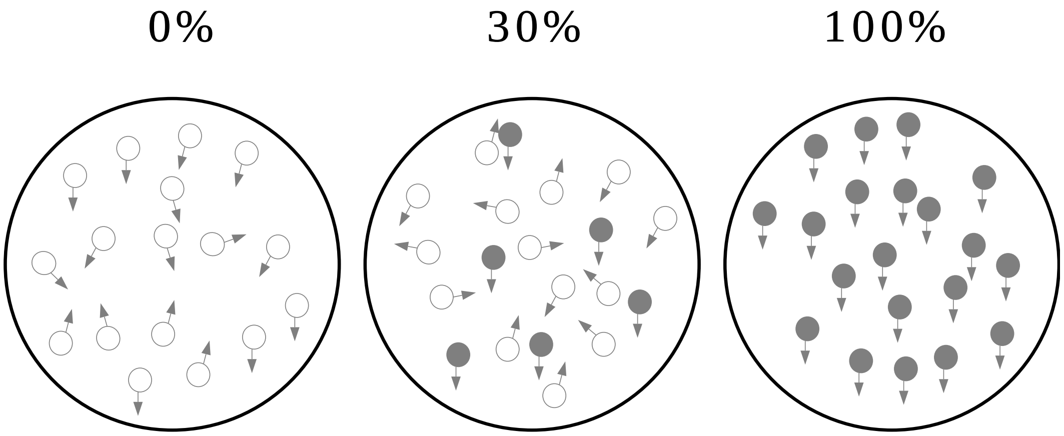

Figure 1.22: *A random dot kinematogram used to measure motion sensitivity.* Each dot is flashed briefly at random positions on the screen. When the correlation is zero, a dot is equally likely to move in any direction in the next frame. The experimenter can introduce motion into the stimulus gradually by increasing the correlation between dots presented in successive frames. Because the correlated dots (shown as filled circles) all move in the same direction, the fraction of correlated dots controls the net motion signal in the display. In the actual display, the correlated and uncorrelated dots in a frame appear the same; the filled and open dots are used in the figure only to explain the principle (Source: Newsome and Pare, 1988).

Motion Perception and Brain Activity

Lesions of area MT. Lesion studies provide further evidence that area MT plays a role in motion perception. A lesion in area MT causes performance deficits on various motion tasks with no corresponding loss in visual acuity, color perception, or stereoscopic depth perception (Newsome and Pare, 1988; Siegel and Anderson, 1986; Schiller, 1993).

Newsome and Pare (1988) found profound deficits when the animal was forced to discriminate motion using the random dot kinematogram shown in Figure 1.22. In this type of kinematogram, each dot is flashed briefly at random positions on the screen. The correlation of the dot positions from frame to frame is the independent variable. When the correlation is zero, a dot is equally likely to appear anywhere on the screen in the next frame. At this correlation level, there will be some local motion signals, by chance, but the average motion will be zero. The experimenter can introduce net motion in the stimulus by correlating some of the dot position in adjacent frames. When the correlation is positive, some fraction of the dots, the *correlated dots*, reappear displaced by a fixed amount in one direction. Hence, the correlated dots introduce a net motion direction into the display.

Ordinarily, the monkey is asked to discriminate whether the net direction of dot motions is in one of two directions. When few of the dot motions are correlated, performance is near chance (50 percent correct). When many of the dot motions are correlated, performance is nearly perfect (100 percent correct). The investigators measure threshold by varying the number of correlated dots required for the monkey to judge the correct direction of motion on 75% of the trials.

There are two MT areas, one on each side of the brain. Each area receives input from the opposite visual hemifield. After lesioning area MT on one side of the brain, Newsome and Pare found that threshold for detecting motion increased by a factor of roughly 4 for stimuli in the relevant hemifield. Threshold for stimuli in the other hemifield remained at pre-operative performance levels, as does the animal's performance on non-motion tasks, such as orientation discrimination (Newsome and Wurtz, 1988).

In Newsome and Pare's experiment, the motion deficit is transient; performance returns to pre-operative levels within a week or two following the lesion. In other motion tasks, such as speed discrimination, the lesion-induced deficit can be permanent (Schiller, 1993). The transient loss of function in certain tasks affords the opportunity to study neural plasticity. Presumably, following removal of MT, however, some of the functions of the lost area are taken over by existing areas.

There is evidence of functional reorganization of many cortical functions For example, Gilbert et al. (1992) have shown that after a retinal lesion that created a blindspot in the animal's visual field, the receptive fields of neurons within area V1 reorganize fairly quickly. Neurons whose receptive fields were driven by retinal signals originating in the lesioned area begin to respond to the signals from surrounding retinal areas. This plasticity seems to be a fundamental and special capability of the cortex, since the reorganization was not present in the subcortical lateral geniculate nucleus.

Probably, this ability to reorganize the visual pathways is a fundamental component of the visual system. We know, for example, that visual development depends upon receiving certain types of visual stimulation (e.g. Freeman et al., 1972; Shatz 1992; Stryker and Harris, 1986). The recovery of the animals in the Newsome and Pare study, as well as the rapid reorganization of receptive fields in Gilbert et al. described above, suggest that this reorganization may be a pervasive feature of the visual representation in adult animals as well.

Behavior and Neural Activity The visual pathways are constantly inferring the properties of the objects we perceive. These algorithms are essential to vision, but mainly, they are hidden from our conscious awareness. We have spent most of our time trying to understand these algorithms, and how they are implemented in the visual pathway.

There is a second important and intriguing question about the cortical representation of information: this is the question of our *conscious experience*. At some point, the visual inference is complete; the motor pathways must act, and perhaps our conscious awareness must be informed about the inference. What is the nature of the representation that corresponds to the final visual inference? Which neural responses code this representation?

There is a growing collection of papers that probe the relationship between behavior and neural responses. An important part of the ability to perform such studies has been the development of techniques to measure the neural activity of alert, behaving monkeys. To obtain such measurements, the experimenter implants a small tube into the animal's skull. During experimental sessions the experimenter inserts a microelectrode through the tube to record neural activity. The electrode insertion is not painful so there is no need to anesthetize the animal. In this way, the experimenter can measure neural activity while monkeys are engaged in performing simple perceptual tasks. These experiments provide an opportunity to compare behavior and neural activity within a single, alert and behaving animal (Britten et al., 1992; Parker and Hawken, 1985).

The relationship between performance and neural activity has been studied for the detection of contrast patterns, orientation discrimination, and motion direction discrimination (e.g. Barlow et al, 1987; Britten et al. 1992; Hawken, et al. 1990; Parker and Hawken, 1985; Tolhurst, 1983; Vogels, 1990). In the motion experiment reported by Britten et al. (1992), for example, the experimenter first isolated a neuron in area MT and determined the neuron's receptive field and best motion direction. The monkey was then shown a random dot kinematogram moving in one of two directions within the receptive field of the neuron. The animal made a forced-choice decision concerning the perceived direction of motion, and at the same time the experimenters recorded the activity of the neuron.

The response of individual neurons did not predict the animals response on any single trial. But, using a simple statistical model Britten et al. discovered that, on average, the response of a single MT neuron discriminated the motion direction as well as the whole animal can discriminate the motion direction.

Considerably more information is encoded by a single MT neuron about motion than is encoded by a single V1 neuron about pattern. For example, the sensitivity of individual V1 neurons to sinusoidal contrasting gratings is substantially lower than the animal's sensitivity. The similarity between behavioral and neural sensitivity on the motion task supports the view that area MT is specialized to represent motion perception. The finding also raises some interesting questions about how information is pooled within the nervous system to make behavioral decisions. In area MT alone there are many hundreds of neurons with equivalent sensitivity to this stimulus. If the responses of these neurons are largely independent, then pooling their outputs would improve performance substantially. Yet, the animal's performance is not much better than we would expect if the animal were simply

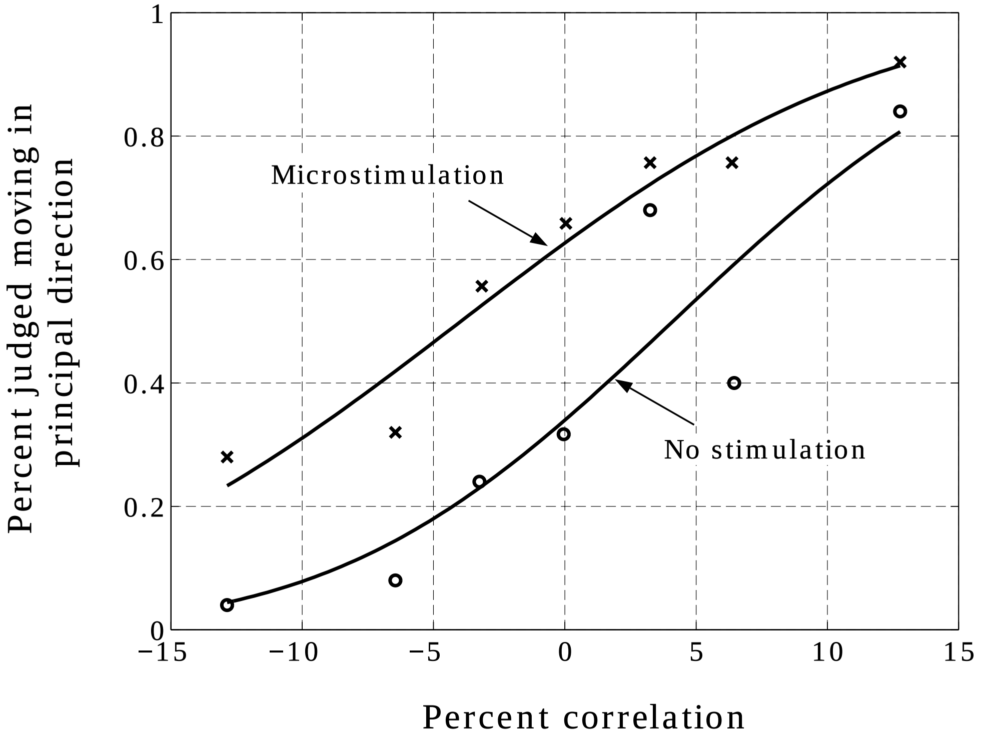

Figure 1.23: *The effect of electrical stimulation in area MT on motion perception.* An alert behaving monkey judged the direction of motion of a random dot kinematogram. Judgments made without electrical stimulation are shown by the open symbols; judgments made in the presence of small amounts of electrical stimulation within area MT are shown by the filled symbols. In this experiment, the microstimulation had the same effect as increasing fraction of correlated dots by 10 percent (Source: Salzman and Newsome, 1992).

using the output of a single neuron. Perhaps this is so because the neural responses are correlated (Zohary et al., 1994).

Microstimulation Studies of Motion and MT Generally, observations based on correlations are a weaker form of evidence than observations based on direct experimental manipulations. Newsome and his collaborators extended their analysis beyond correlational by manipulating the neural responses during behavioral trials (Salzman and Newsome, 1992).

As in the correlational experiments, the investigators first isolated a neuron in area MT. Within area MT, nearby neurons tend to have similar direction selectivity (Albright, 1984). The experimenters used a test stimulus whose direction corresponded to the best direction of the isolated neuron; the presumption is that this direction defines the best direction of the receptive field for most of the nearby neurons.

Again, the monkey made a forced-choice decision between kinematogram motion in the best direction of the neuron or in the opposite direction. On one half of the trials, randomly selected, the investigator injected a small amount of current into the brain, stimulating the neurons near the electrode. The microstimulation changed the monkey's performance, as if the current strengthened the motion signal in the direction of the local neurons' best direction sensitivity. The open and filled symbols in the Figure 1.23 show the monkey's performance on trials with and without the current injection, respectively. The monkey was more likely to say the stimulus moved in the direction preferred by the neurons in the presence of microstimulation than in its absence. In this particular experimental condition, the microstimulation was equivalent to increasing the percentage of dots moving in the test direction by ten percent.

There are two reasons why this experiment is very significant. First, the method involves a direct experimental manipulation of the animal's behavior, rather than an inference based on a correlation. The investigator actively controls the state of the neurons and observes a change in the behavior. Second, the method reminds us of the hope of someday designing visual prosthetic devices. By understanding the perceptual consequences of visual stimulation, we may be able to design visual prosthetic devices that generate predictable and controlled visual sensations.

1.6 Conclusion

Many important aspects of motion perception can be understood and predicted based on computations using only the local space-time variations of image intensities. Many of the computational elements of motion calculations, such as space-time oriented linear filters and velocity constraint lines, have a natural counterpart in the receptive fields of neurons within the motion pathway.

But the results of many behavioral experiments suggest that the surface and object representations also provide a source of useful information for the computation of motion and depth. Observers see moving surfaces sliding transparently across one another, they see motions of texture elements defined by implicit edges, they infer three-dimensional shapes and depth from the limited information in sets of random dots. It seems to me that we must understand how to incorporate surfaces in our computational theories, and we must understand how surfaces are represented in the neural pathways, to arrive at our next level of understanding of motion perception. The coupling of computational, behavioral and neural measurements has served us well this far, and I suspect that trying to incorporate surface representations using all of these methods will continue to be our best chance of understanding motion.

Taken together, it is evident that our inferences of motion and depth are not an isolated visual computations, but rather they are part of a web of visual judgments. We perceive motion in a way that depends on contrast, color, and other more abstract image features such as surface, edge and transparency. Integrating this information requires some sophisticated neural processing, and we are just at the beginning of studying this process both behaviorally and neurophysiologically. In the next chapter, we will review some of the more interesting but complex aspects of how we integrate different types of visual cues in order to make sense of the retinal image.

Exercises

1. Answer the following questions about the response of space-time oriented linear filters.

   (a) Qualitatively, what will the response of a space-time oriented linear filter be to a drifting sinusoidal pattern?

   (b) Suppose you know the response of a space-time oriented linear filter to two stimuli. Will you be able to predict the filter's response to the superposition of the stimuli?

2. Different display technologies use different methods for temporal sampling of motion sequences. Several of these strategies are described in a footnote within this chapter.

   (a) Represent the different sampling schemes used by television and by movies on a space-time diagram.

   (b) Represent these temporal sampling methods on a spatial frequency versus temporal frequency diagram, (f_t, f_x).

3. A squarewave grating consists of the weighted sum of a several sinusoidal components. It is possible to create an interesting motion illusion, called the *fluted squarewave*, by sampling the motion of the squarewave and removing the fundamental component (Adelson and Bergen, 1985; Georgeson and Harris, 1990).

   (a) Make an (f_t, f_x) drawing showing the locations of the Fourier components a stationary squarewave grating. (See Chapter **??** if you don't know what the Fourier components of a squarewave are.)

   (b) Make an (f_t, f_x) drawing showing the locations of the Fourier components a squarewave grating drifting continuously to the right.

   (c) Make a (f_t, f_x) drawing showing the locations of the Fourier components of a *sampled* squarewave grating drifting to the right. Choose the temporal sampling interval so that the energy in the temporal replicas are aligned into a uniform grid.

   (d) Now, remove the components present in the original sinusoid from the drawing containing the sampled squarewave. Examine the pattern of Fourier components that remain and predict the perceived motion of this pattern.

4. There is a classic visual illusion called the *Pulfrich effect* that illustrates the interconnection between motion, depth and timing. Suppose you observe a

pendulum that is swinging back and forth in a plane parallel to the front of your face. Now, cover one with with a dark filter, such as a half log unit neutral density filter that you can purchase at a camera store. After a few moments, the pendulum will appear to be swinging in outside of the plane, following an elliptical arc.

(a) The Pulfrich effect is explained by assuming that (a) under normal viewing temporal synchrony is used to identify corresponding signals from the two eyes, and (b) the signal from the eye whose light intensity is attenuated by the neutral density filter arrives later to the brain than the signal from the unattenuated eye. Make a diagram showing why this temporal lag can explain the perceived elliptical path of the swinging pendulum.

(b) Read the articles by Thompson (1993) and Carney et al. (1989) for recent analyses of this phenomenon in terms of the visual pathways.

(c) Read the article by Hofeldt and Hoefle (1993) for an application of the Pulfrich effect to predicting the performance of baseball players.

5. The following questions are meant to be thought-provoking, not to have simple answers. Write brief answers concerning how you might develop a research strategy to answer these questions.

   (a) Why should perceived velocity depend on the color or contrast of the moving stimulus? What implications does the existence of this relationship have for the way human motion detectors are organized? Is your hypothesis testable by experiment?

   (b) Very young infants probably cannot control their eye movements very well. What implication does this have for their ability to perceive motion? Answer using different assumptions concerning the quality of the efference copy signal and the retinal signals concerning motion flow fields.

   (c) Microstimulation experiments have been very useful in analyzing the relationship between area MT and motion perception. Design a microstimulation experiment to evaluate the relationship between a visual area and color perception.