Spatiotemporal energy models for the perception of motion

Edward H. Adelson and James R. Bergen

David Sarnoff Research Center, RCA, Princeton, New Jersey 08540

Received July 9, 1984; accepted October 12, 1984

A motion sequence may be represented as a single pattern in *x*-*y*-*t* space; a velocity of motion corresponds to a three-dimensional orientation in this space. Motion information can be extracted by a system that responds to the oriented spatiotemporal energy. We discuss a class of models for human motion mechanisms in which the first stage consists of linear filters that are oriented in space-time and tuned in spatial frequency. The outputs of quadrature pairs of such filters are squared and summed to give a measure of motion energy. These responses are then fed into an opponent stage. Energy models can be built from elements that are consistent with known physiology and psychophysics, and they permit a qualitative understanding of a variety of motion phenomena.

1. INTRODUCTION

When we watch a movie, we see a sequence of images in which objects appear at a sequence of positions. Although each frame represents a frozen instant of time, the movie gives us a convincing impression of motion. Somehow the visual system interprets the succession of still images so as to arrive at a perception of a continuously moving scene.

This phenomenon represents one form of apparent motion. How is it that we see apparent motion? One possibility is that our visual system matches up corresponding points in succeeding frames and calculates an inferred velocity based on the distance traveled over the frame interval. Much research on apparent motion has taken the establishment of this correspondence to be the fundamental problem to be solved.[1-3] We argue that this correspondence problem can often be bypassed altogether; we take up this argument after discussing various approaches to the problem of motion analysis.

Figure la shows a vertical bar, which is presented at a sequence of discrete positions at a sequence of discrete times. In a typical feature-matching model, the visual system is said to (1) find salient features in successive frames; (2) establish a correspondence between them; (3) determine Δx, the distance traveled, and Δt the time between frames; and, finally, (4) compute the velocity as $\Delta x/\Delta t$. In this example, the features to be matched might be the edges of the bar.

In a typical global matching model, the visual system would perform a match over some large region of the image, in essence performing a template match by sliding the image from one frame to match the image optimally in the next frame. Most cross-correlation models (see, e.g., Lappin and Bell[4]) are examples of the global matching approach. Once again, Δx and Δt can be determined, and the velocity can be inferred.

Matching models are designed to make predictions about stimuli presented as sequences of frames (e.g., movies). Not all stimuli fall naturally into such a description. In an ordinary television, for example, the electron beam illuminates adjacent points in a rapid sequence, sweeping out the even lines of the raster pattern on one field and then returning to fill in the odd lines on the next field (two fields constitute a frame). Should the matching be taken between frames or between fields? For that matter, why should it not be taken between the successively illuminated points themselves? (Note that the motion of the raster itself which is normally invisible, *will* become visible if the raster is quite slow.)

Although the answer is not immediately obvious, it is clear that we need to consider the well-known persistence of visual responses—i.e., the temporal filtering imposed by early visual mechanisms—in order to make sense of even the simplest phenomena of apparent motion. The rapidly illuminated points on a television screen are blended together in time, effectively making all the lines of a frame (including both fields) visually present at one time. One approach to motion modeling, therefore, is to build in a temporal-filtering stage that preprocesses the visual input before it is passed along to the matching system. The resulting model treats the stimulus in both a continuous and a discrete fashion. Filtering is a continuous operation and leads to a continuously varying output, whereas matching is discrete, taking place between images sampled at two particular moments in time. Having been forced to introduce filtering into the model, we would like to make full use of its properties. In fact, filtering can be used to extract the motion information itself, thus rendering the discrete matching stage superfluous.

There are other reasons for shying away from matching models as they are commonly presented. They can usually make predictions about simple stimuli such as a moving bar, but they may run into trouble when presented with a sequence such as is shown in Fig. 1b. Here, a sequence of vertical random noise patterns is presented. When this sequence is viewed, complex motions are seen, varying from point to point in the image. Different velocities are seen at different positions, and these velocities change rapidly. A feature-matching model has difficulty making predictions because of the familiar problems: What constitutes a feature? What should be matched to what? Most feature-based models are not well enough defined to offer predictions about a stimulus such as that of Fig. 1b. Yet motion is seen, and we would like to believe that this motion percept is generated by the same lawful processes that generate the percept of the moving bar.

Can a global matching model, such as a cross-correlation model, do better? Again, it is hard to know what such a model will predict. Most global matching models have been formulated only to deal with the visibility of single global motions and thus cannot be easily applied to the situation in which many motions are seen at different points in the field.

0740-3232/85/020284-16$02.00

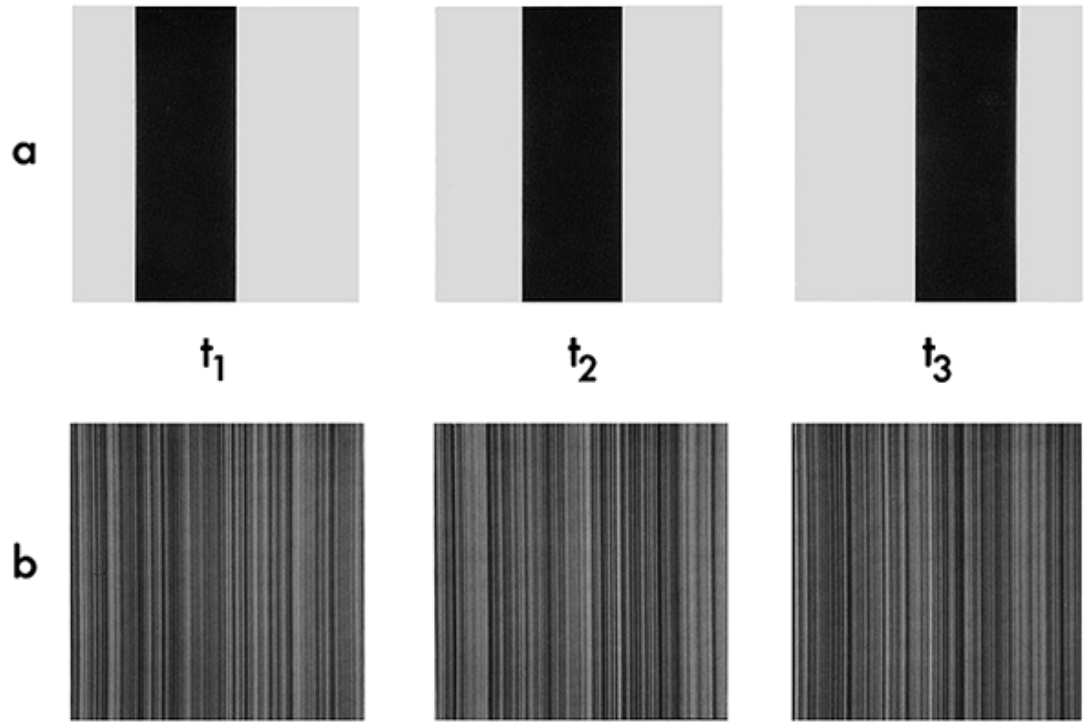

Fig. 1. a, A sequence of images presented at times t_1, t_2, and t_3 showing a bar moving to the right. b, A sequence of vertical random noise patterns, also shown at three successive instants of time. Motion is seen in each case. The motion percept is simple in a and complex in b, but a motion model should be able to handle both cases.

A number of approaches have recently been developed that can be used with complex inputs such as the dynamic noise of Fig. lb. Marr and Ullman[5] describe a method for extracting the motion of zero crossings in the outputs of linear filters by comparing the sign of the filter output to the sign of its temporal derivative at the zero crossing. A rather different approach has been described by van Santen and Sperling[6] in an elaboration of Reichardt's[7] model in which a local correlation (i.e., multiplication) is performed across space and time. In van Santen and Sperling's model, filters tuned for spatial frequency serve as the inputs to the correlator stages. Van Santen and Sperling provide a formal analysis of the model's properties, describe a set of linking assumptions, and show that the model makes correct predictions about a large variety of simple motion displays. A third approach has been described by Watson and Ahumada[8]: Motion information is extracted with simple linear filters without a multiplicative stage, the filters are tuned for spatial and temporal frequency as well as velocity, and directional selectivity is achieved by setting up the appropriate phase relationships between an underlying pair of filters. It is notable that this approach achieves directional selectivity without any nonlinearities (although some sort of nonlinearity must, of course, be present at some point for motion detection to occur). Ross and Burr[9] have also proposed that the visual system extracts motion information with directionally tuned linear filters. Morgan[10] has applied linear-filtering concepts to stroboscopic displays, and Adelson[11] has discussed how a number of motion illusions can be understood in terms of mechanisms that respond to the motion energy within particular spatiotemporal-frequency bands.

Although it is not immediately apparent, there are significant formal connections between the linear-filtering approach and the correlational approach of a Reichardt-style model, as has been previously noted.[6,12] The topic is taken up in Appendix A; at this point, we simply comment that both types of model can be considered to respond to motion energy within a given spatiotemporal-frequency band (a property that will be discussed at greater length below).

Our interest in this paper is not so much to discuss a particular model as to discuss a general class of models and not so much to discuss this class as to discuss a general approach to the problem of motion detection. We will consider models closely related to the ones just mentioned—models that are based on a simple low-level analysis of visual information, starting with the outputs of linear filters. This kind of processing is well understood and can be readily applied to any stumulus input. Moreover, it is just the kind of processing that is considered to occur early in the visual pathway, based on a large variety of psychophysical and physiological experiments.[13-16]

Low-Level Processing in Motion Perception

A low-level approach seems particularly appropriate when one is dealing with motion phenomena that occur with a rapid sequence of presentations. Many investigators have found that these rapid presentations lead to motion percepts that are determined by rather simple low-level properties of the stimuli.

Braddick[17] provided evidence for two distinct kinds of motion mechanisms in apparent motion. He called them long-range and short-range mechanisms. The short-range process operates over rather short spatial distances and short time intervals and involves low-level kinds of visual informa-

tion. The long-range mechanism can operate over large spatial separations and longish time intervals and may involve somewhat higher-level forms of visual information.

Hochberg and Brooks[18] also found evidence for two processes in motion perception. They presented a sequence of images containing collections of simple shapes, such as circles, triangles, and squares. Each shape could take one of two motion paths: it could take a short path but change identity (e.g., a triangle could take a short path by turning into a square), or it could take a longer path and retain its identity. At lower presentation rates, the identity of the objects became important and a triangle would remain a triangle even if it meant taking a longer path. But with rapid presentations, the shorter path length won out, even though it meant abandoning stable object identity.

Sperling[19] found that rapid, multiple-presentation motion stimuli gave much more compelling motion than did the slower two-view stimuli of classic apparent-motion experiments. Evi-dence for a fast, low-level process in motion perception has also been presented by various others.[2,20,21] The models that we develop below are designed to deal with the rapid-presentation situation and are based on the simplest, lowest-level processes that we can use. We will try to avoid the concept of matching altogether.

2. REPRESENTING MOTION IN *X-Y-T* SPACE

Moving stimuli may be pictured as occupying a three-dimensional space, in which x and y are the two spatial dimensions and t is the temporal dimension. Consider a vertical bar moving continuously to the right, as shown in Fig. 2a. The three-dimensional spatiotemporal diagram is shown in Fig. 2b; the moving bar becomes a slanted slab in this space. If the continuous motion is sampled at discrete times, the result is Fig. 2c, which shows a movie of a moving bar.

In Fig. 3, only the x-t slice of the space is shown (we can ignore the y dimension since a vertical bar is unchanging along the y direction). The moving bar in Fig. 3a becomes a slanted strip. The slant reflects the velocity of the motion. Figure 3b shows the result of sampling the continuous motion. In practice, when one presented the movie corresponding to Fig. 3b, one would leave each frame on for a period of time before replacing it with the next one. Figure 3c shows the spatiotemporal plot of a movie in which each frame lasts almost through the full interval between frames. (In most actual movie projection, a single frame is broken up into several shorter flashes in order to minimize the perception of 24-Hz flicker; for simplicity, we do not consider the case of multiple shuttering here.)

We know that the sampled motion of Fig. 3c will look similar to the continuous motion of Fig. 3a. Indeed, if the sampling is sufficiently frequent in time the two stimuli will look identical. Pearson[22] has discussed how this may be understood by applying the standard notions of sampling and aliasing to the case of three-dimensional sampling in space and time and considering the spatiotemporal-filtering properties of the human visual system. The argument, in brief, is this: A continuously moving image has a three-dimensional Fourier spectrum in f_x-f_y-f_t. A sampled version of the display has a different spectrum. The differences between the spectra of the continuous and sampled scenes may be called sampling artifacts (when these artifacts intrude on the spectrum of the original signal they are known as aliasing components). It is these components that allow an observer to distinguish between a continuous and a sampled display. The task of a display engineer is therefore to ensure that the artifactual components that are due to sampling are of such low contrast that they are invisible to the human observer. To achieve this goal, it is necessary not to remove the artifactual components altogether but merely to prevent them from reaching threshold visibility. This can be done by appropriately prefiltering, sampling, and postfiltering the moving images.

It is not always easy to assess the visibility of sampling artifacts; one must take into account subthreshold summation between the artifactual components as well as masking by true-image components. However, Watson *et al.*[23] have described a set of conditions under which one may be confident that the artifacts will not be visible. For sufficiently high spatial and temporal frequencies, human contrast sensitivity is zero; that is, components lying outside a certain spatiotemporal-frequency limit (which Watson *et al.*[23] call the window of visibility) cannot be seen regardless of their contrast. If the sampling is suffciently fine to keep all the spectral energy of the sampling artifacts outside this window, then the artifacts must be invisible.

Morgan[10] has applied frequency-based analyses to the problem of motion interpolation and has described two different approaches. In the first approach, the analysis begins with the extraction of a position signal, i.e., a single number that varies over time. Low-pass filtering is then applied to this signal. Thus the first stage of motion analysis is highly nonlinear (position extraction), and linear filtering follows it. In Morgan's second approach, the filtering is applied directly to the stimulus itself; position is extracted after the filtering has occurred. The present discussion (like that of Pearson and that of Watson *et al.*) *is* more closely connected to the second approach than to the first. But one should note that position as such need not be extracted in the computation of motion, as will become clear in what follows.

When temporal sampling is too coarse—as in an old movie—motion tends to look jerky. But motion is still seen. That is, to convey the impression of motion, it is not necessary that a sampled stimulus be indistinguishable from a continuous one. A spatiotemporal-frequency analysis helps one to understand this as well, because a continuous and a sampled stimulus share a great deal of spatiotemporal energy, even if they do not share it all. We can expect the two stimuli to look similar insofar as there are visual mechanisms that respond to the shared energy.

It is sometimes helpful to perform the analysis in the orig-

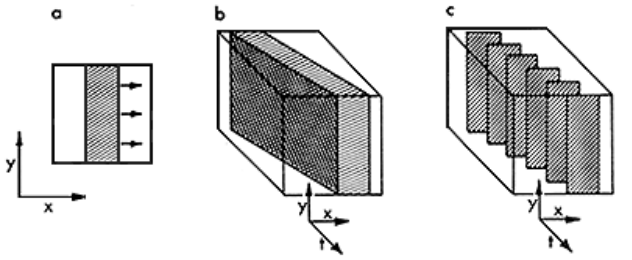

Fig. 2. a, A picture of a vertical bar moving to the right. b, A spatiotemporal picture of the same stimulus. Time forms the third dimension. c, A spatiotemporal picture of a moving bar sampled in time (i.e., a movie).

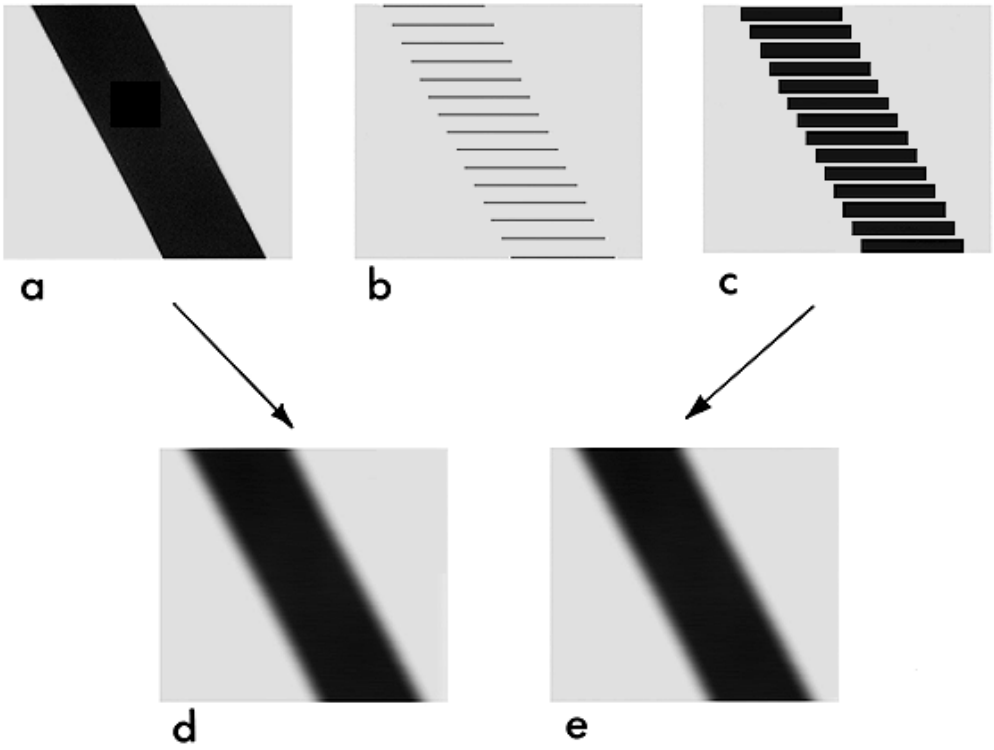

Fig. 3. a, An (x, t) plot of a bar moving to the right over time. Time proceeds downward. The vertical dimension is not shown. b, An (x, t) plot of the same bar, sampled in time. c, The sampled motion as displayed in a movie in which each frame remains on until the next one appears. d, Continuous motion after spatiotemporal blurring. e, Sampled motion after spatiotemporal blurring. The middle- and low-frequency information is almost the same for the two stimuli.

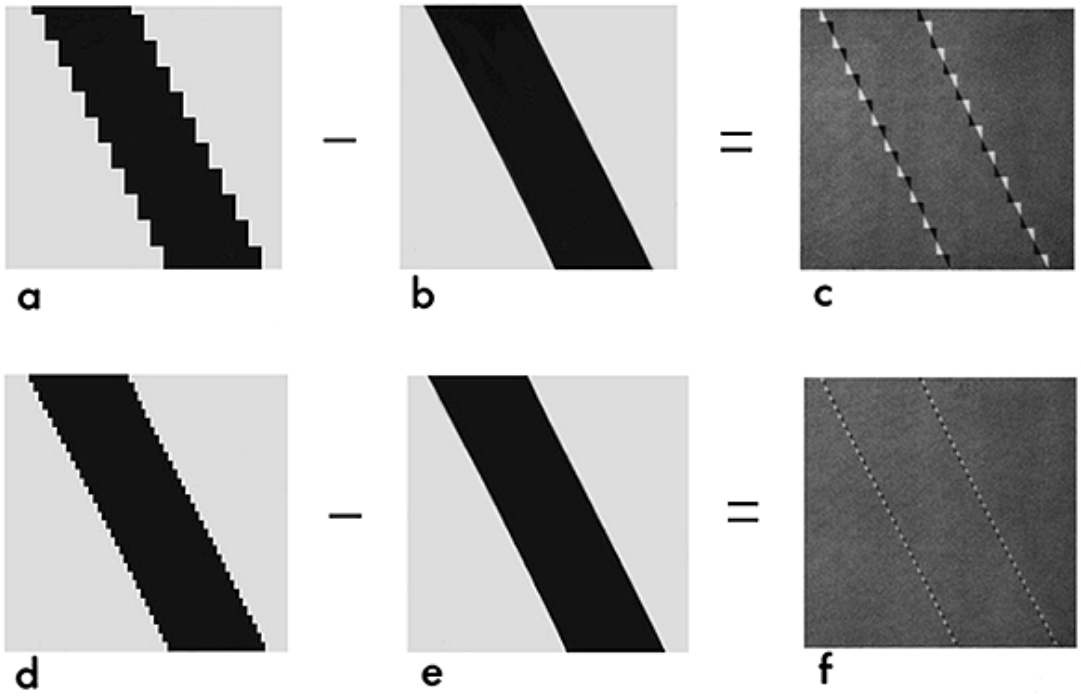

Fig. 4. (x, t) plots of moving bars. a, A movie of a bar moving to the right. b, A bar moving to the right continuously. c, The difference (sampling artifacts) between the sampled and continuous motions. d, A movie sampled at a high frame rate. e, Continuous motion. f, The difference between the finely sampled and continuous motion. When the sampling rate is high, the sampling artifacts become difficult or impossible to see.

inal space-time domain, rather than in the frequency domain. Figure 4 makes explicit the difference between the sampled and continuous versions of the moving bar. If we simply subtract the continuous pattern (Fig. 4b) from the sampled one (Fig. 4a), we can derive a new spatiotemporal plot of the sampling artifacts, as illustrated in Fig. 4c. Since the difference can be positive or negative, we have displayed it on a gray pedestal, so that gray corresponds to zero, white to positive, and black to negative. Observe that the sampled-motion stimulus of Fig. 4a can be considered to be the sum of the real motion of Fig. 4b and the artifacts of Fig. 4c. That is, we can think of the sampled motion as being continuous motion with sampling noise added to it.

If the motion is sampled more frequently in time, the approximation to continuous motion is improved, as shown in Fig. 4d. In this case, the artifacts (Fig. 4f) have rather little energy in the range of frequencies that we can see. If sampling is made frequent enough, there will plainly come a point at which the artifactual components have so little energy in the visible spatial- and temporal-frequency range that they will become invisible, since the fine spatiotemporal structure of the artifacts will be blurred to invisibility by the spatial and temporal response of the eye. At this point, the continuous and the sampled stimuli will be perfectly indistinguishable.

Again, it is not necessary that the sampled stimulus look identical to the continuous one in order for the motion to look similar. A motion mechanism that responds to low spatial and temporal frequencies will give the same responses to the two stimuli, even if mechamisms sensitive to higher frequencies give different responses.

So far, we have discussed the conditions under which different moving stimuli may be expected to give similar impressions of motion. But we have not discussed how motion information, in itself, might be extracted; this constitutes our next problem.

3. MOTION AS ORIENTATION

Motion can be perceived in continuous or sampled displays, when there is energy of the appropriate spatiotemporal orientation. This is illustrated in Fig. 5, which shows spatiotemporal diagrams of a bar: a, moving quickly to the left; b, moving slowly to the left; c, stationary; d, moving slowly to the right; and e, moving quickly to the right. The velocity is inverse with the slope.

The problem of detecting motion, then, is the problem of detecting spatiotemporal orientation. How can this be done? We already know a way of detecting orientation in ordinary spatial displays, namely, through the use of oriented receptive fields like those described by Hubel and Wiesel[24] and sometimes referred to as bar detectors and edge detectors. Simple cells in visual cortex are now known to act more or less as linear filters: Their receptive-field profiles represent a weighting function, with both positive and negative weights, which may be taken as the spatial impulse response of a linear system.[14]

If we could construct a cell with a *spatiotemporal* impulse response that was analogous to a simple cell's *spatial* impulse responses, we would have the situation shown at the bottom of Fig. 5 (cf. Ross and Burr[9]). The cell's spatiotemporal impulse response is oriented in space and time. In Fig 5f, it responds well to an edge moving continuously to the right. In

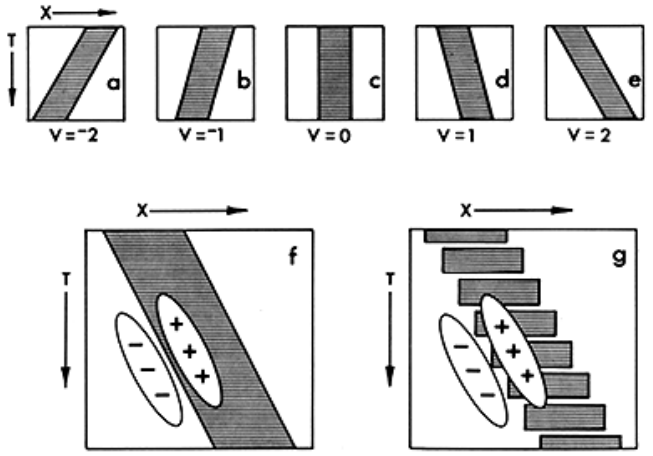

Fig 5. a-e (x, t) plots of bars moving to the left or to the right at various speeds. f, Motion is like orientation in (x, t), and a spatiotemporally oriented receptive field can be used to detect it. g, The same oriented receptive field can respond to sampled motion just as it responds to continuous motion.

Fig. 5g, it responds well to a sampled version of the same stimulus. As far as this hypothetical cell is concerned, both stimuli have substantial rightward-motion energy.

The models that we will develop will be based on idealized mechanisms; in discussing these mechanisms we will use the terms "unit" and "channel." A unit corresponds roughly to a cell or to a small set of cells working in concert to extract a simple property at one position in the visual field. A channel consists of an array of similar units distributed across the visual field.

In principle, there is no reason why an oriented unit could not be constructed directly. The unit would gather inputs from an array of photoreceptors covering the spatial extent of its receptive field, and it would sum their outputs over time with the appropriate temporal impulse responses. In practice, however, such a unit would be difficult to construct because it would require a different temporal impulse response correctly tailored to each spatial position in the receptive field.

The problem, then, is to construct a unit that responds to spatiotemporal orientation (i.e., motion) and yet that is built out of simple neural mechanisms. In Section 4, we will discuss how such a unit can be built by combining impulse responses that are space-time separable by using an approach similar to that of Watson and Ahumada.[8] For those readers who are not en-tirely comfortable with these notions, we begin by reviewing space-time separability as well as spatiotemporal impulse responses.

4. SPATIOTEMPORAL IMPULSE RESPONSES

Many cells in the visual system respond (to a good approximation) by performing a weighted integration of the effect of light falling on their receptive field; the receptive-field profile, with its positive and negative lobes, defines the weighting function, or spatial impulse response. Across the top of Fig. 6 is an idealized spatial impulse response from such a cell. Since any spatial pattern can be thought of as a sum of points of light of various intensities packed together side by side, one can easily predict the response of a linear unit to an arbitary

input pattern by summing its responses to the varying local intensities, point by point.

A temporal impulse response is shown running down the left side of Fig. 6. One normally thinks of the temporal impulse response as representing the time course of a unit's response following a impulse input. However, one may also think of the impulse response as a temporal weighting function, which describes how inputs in the past are summed to produce the response at the present moment (the time axis must be reversed).

If the spatial and temporal impulse responses are combined in the simplest manner, the result is the separable spatiotemporal impulse response shown schematically in the center of the figure. If the spatial impulse response is $H_s(x)$, a function of x, and the temporal impulse response is $H_t(t)$, a function of t, then the spatiotemporal impulse response is $H_{st}(x, t) = H_s(x)\ 5\ H_t(t)$. In this case, there are six lobes, alternately positive and negative, forming a checkerboardlike pattern. This pattern describes how inputs at various positions and times are to be summed to give the current output.

Spatiotemporally separable impulse responses are easy to build. If a unit gathers inputs from a set of spatially distributed positions, weights them by a spatial impulse response, and then sends the output through a temporal filter, the resulting spatiotemporal impulse response will be separable. Or if the outputs of a large number of receptors are are temporally filtered in the same way, and the filtered outputs are combined with a spatial weighting function, then again the net response will be separable. Separability is frequently observed in the early stages of cortical visual processing.[25,26]

Figure 7 illustrates how the spatiotemporal impulse response may be used to analyze the way a unit will respond to a stimulus. The stimulus here is a light dark edge that is initially stationary, then moves to the right, and then becomes stationary again. The stimulus may be considered to lie on a continuous strip, which is drawn upward over time, as shown in Fig. 7a. A picture of the unit's impulse response is overlaid on the stimulus, to show how inputs at all points and times are

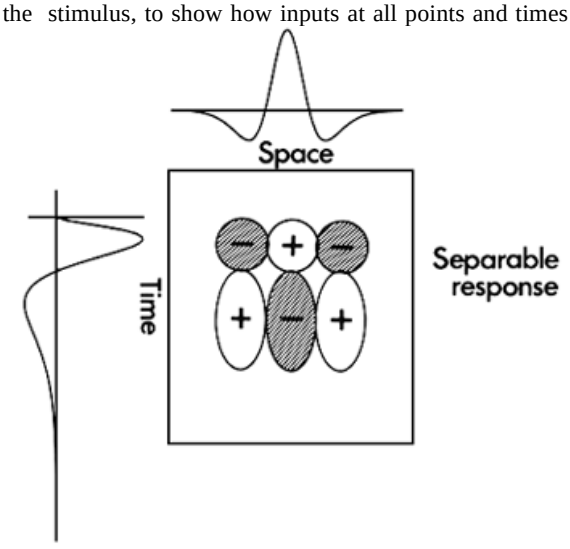

Fig. 6. A spatiotemporally separable impulse response. The spatial and temporal impulse responses are shown along the margins. Their product is shown schematically in the center. The spatiotemporal impulse re-sponse is a weighting function that sums inputs at various positions and times to determine the present output.

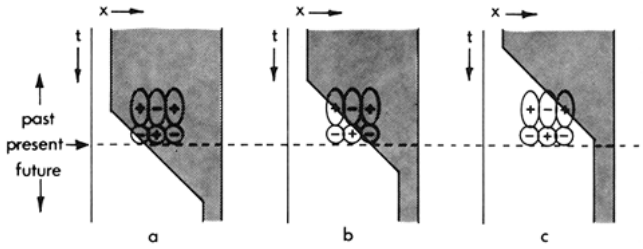

Fig. 7. One may think of a spatiotemporal impulse response as being fixed, while the spatiotemporal stimulus slides beneath it as if pulled along on a strip. At any moment, the integral of the pointwise product of the two functions determines the output; i.e., the two functions are convolved in time (the impulse response is time reversed here; otherwise the operation would be a correlation). This particular unit will respond strongly when the motion lies within its receptive field but will not respond to blank areas or areas without motion.

weighted to give the unit's current output. The present (dashed line) refers to the spatiad pattern that is being shown at the instant that the unit's response is being measured. Inputs from the past lie above this line; inputs that are yet to come lie below it.

At time t_1, the unit is just beginning to "see" the stimulus; as time proceeds (t_2 and t_3), the response will oscillate positive and negative, depending on how the lobes of the spatiotemporal receptive field line up with the spatiotemporal stimulus pattern.

A given unit's output, over time, represents the temporal convolution of the unit's impulse response with the spatiotemporal-input pattern. An array of similar units positioned at various positions in space can be thought of as performing a convolution in both space and time. The resulting channel acts as a filter, which selectively passes some of the spatiotem-poral energy of the stimulus.

To illustrate the use of convolution, consider the spatiotemporal stimulus of Fig. 8a. A light dark edge is first stationary, then moves to the right, then to the left, then right again, and then stops. Convolving this input with the separable impulse response of Fig. 8b (which is magnified for clarity) results in the output of Fig. 8c. White indicates a strong positive response, black a strong negative response, and gray indicates zero. The response is strong when the edge is moving and is absent when the edge is stationary (at the start, at the end, and at the extremes of the trajectory when the edge reverses direction).

The separable unit would be a good candidate for a motion detector, except for one flaw: it cannot tell left from right. It responds equally well to motion that is spatiotemporally oriented either to the left or to the right because it has no orientation of its own. A hypothetical unit that does have spatiotemporal orientation, on the other hand, will do the trick that we need, as shown in Fig. 8e (again, magnified for clarity). In this case, the unit's impulse response is an oriented Gabor function, which is to say, a patch of drifting sine wave under a spatiotemporal Gaussian window. (This is a convenient function to work with, but other oriented functions would serve just as well.) As is illustrated by the convolution of Fig. 8f, this unit gives strong responses to the rightward motion but little or no response to the leftward motion. Thus it is truly selective for direction of motion. Of course, like the

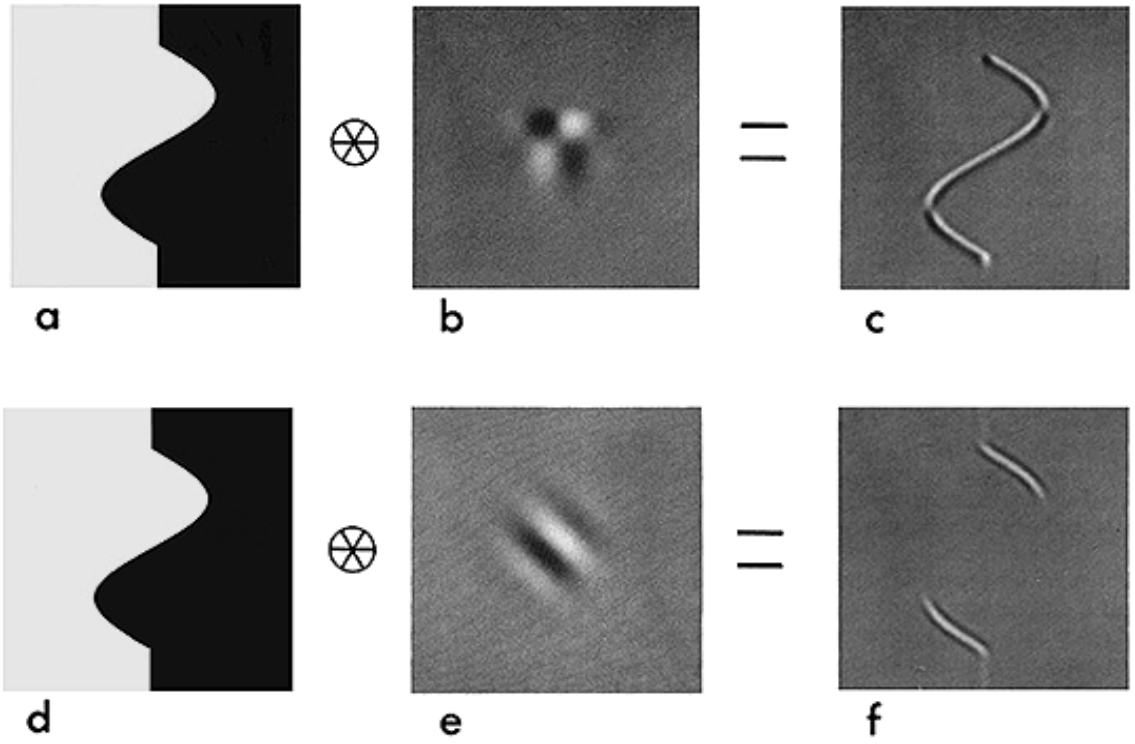

Fig. 8. a, An (x, t) plot of an edge that is stationary, then moves sinusoidally, and then is stationary again. b, A separable spatiotemporal impulse response magnified four times for clarity. c, The convolution of a and b, i,e., the output of a separable channel. There is no selectivity for direction of motion. d The stimulus again. e, A spatiotemporally oriented Gabor function, magnified four times. f, The convolution of d and e. The output is strongly selective for rightward motion.

edge detector of Fig. 5f, this is not an easy unit to build in a physiological system, but there are ways of approximating it that are physiologically plausible.

5. EXTRACTING SPATIOTEMPORAL ENERGY

Spatiotemporally oriented filters are quite useful in analyzing motion, but they pose some difficulties as they stand. They are phase sensitive, which is to say that their response to a moving pattern depends on how the pattern happens to line up with their receptive field at each moment. Thus, as an example, a moving sine-wave grating will elicit a response that itself oscillates sinusoidally over time. At a given moment, the unit's output may be positive, negative, or zero, so that the instantaneous output does not directly signal the motion. We may prefer a response that takes on a constant value for a constant motion; this corresponds to our experience of a constant motion and corresponds to the behavior of many direction-selective complex cells.

The phase problem also shows itself when a moving bar is passed in front of a linear motion-selective unit. The unit's output oscillates positive and negative during the traverse, so again the instantaneous response cannot be used as a simple measure of the motion. Moreover, the sign of the response will depend on the sign of the stimulus contrast, so that a black bar and a white bar moving in the same way will give inverted responses. Later processes would be needed to interpret the oscillating responses, in order to extract a motion measure that was independent of the polarity and momentary phase of the stimulus. (On the other hand, Watson amd Ahumada[27] have discussed how the oscillations might be used to advantage in computing velocity.)

A phase-independent motion detector can be built as shown in Fig. 9. We begin with two units that act as linear spatiotemporal filters on the input. For mathematical convenience we consider the ideal case of oriented Gabor funtions; the one at the left-hand side of Fig. 9a has cosine (even) phase, whereas that at the right-hand side has sine (odd) phase. The phase problem is still with us for each of these two units (as it must be for any linear units).

However, by squaring and summing the two units' outputs, we can extract a measure of local motion energy. (This procedure takes advantage of the fact that $\sin^2 + \cos^2 = 1$. We use the term "energy" rather than "power" to emphsize the fact that time and space are all part of a single continuum and are treated in the same way.) The two Gabor functions are sine and cosine functions weighted by the same Gaussian window, and they allow us to extract energy within a spatiotemporal-frequency band. The resulting response will always be positive, and it will grow and fall smoothly in the region of the moving edge. The energy response will also be the same for a moving white-black edge as it is for a black-white edge moving in the same direction; thus it will be sensitive to the direction of motion but insensitive to the sign of the stimulus contrast. Finally, the energy response will be constant as a sine-wave grating is moved across the freld. Thus constant rightward motion of the grating will give an unmodulated positive response, in accord with the behavior of many complex cells and in accord with our percept of the motion as being smooth and unchanging.

It might be advantageous to have a square root or other compressive nonlinearity following the sum-of-squares stage, in order to keep the outputs within a reasonable range (cf. the research of Pantle and Sekuler[28] on the motion aftereffect). Such a monotone transformation would not affect the basic properties of the motion-extraction process but could have an effect on the performance observed in various tasks, such as the accuracy with which changes in speed and contrast could be judged.

Energy was extracted in Fig. 9a by using the standard trick of squaring the outputs of two filtes that are 90 deg out of phase, i.e., that form a quadrature pair. In the case of Gabor functions, this was done by simply using the sine and cosine versions of the same filter (which are effectively in quadrature). We now consider how similar results can be achieved with more-realistic filters.

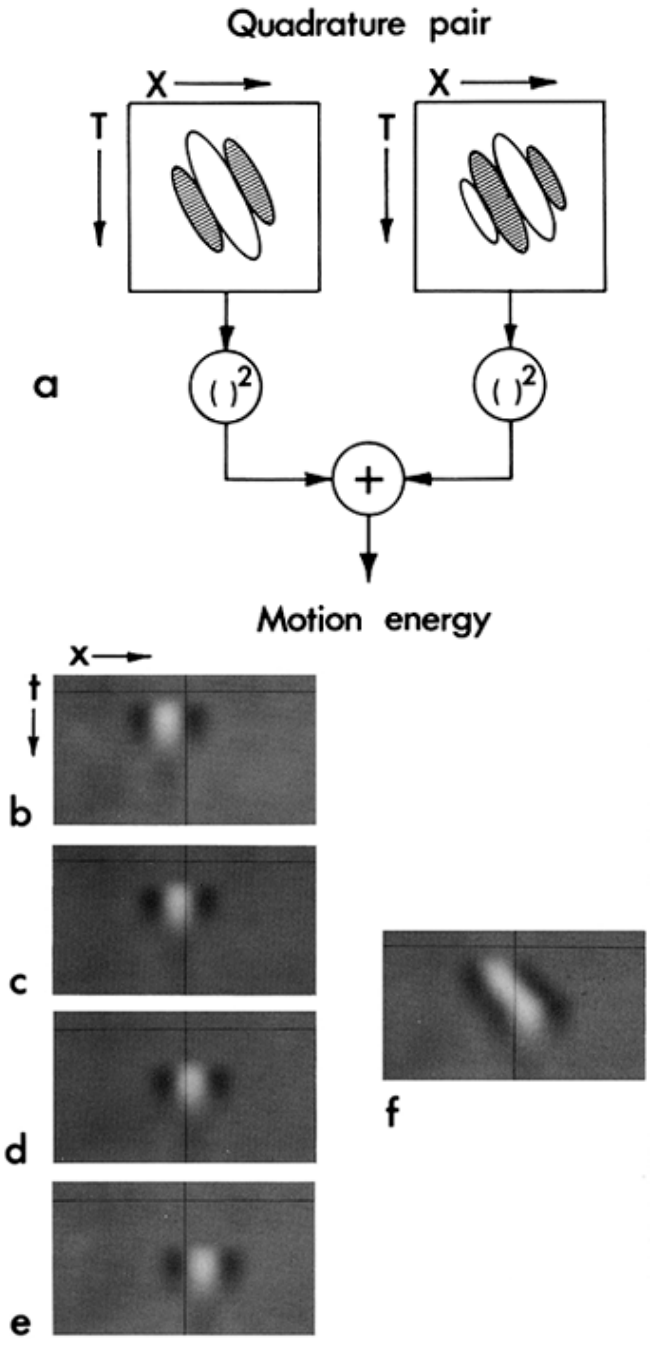

Fig. 9. a, Two linear filters, whose responses are 90 deg out of phase, form a quadrature pair. If their responses are squared and summed, the resulting signal gives a phase-independent measure of local motion energy (within a given spatial-frequency band). The filters shown here resemble spatiotemporally oriented Gabor functions. To approximate such functions, a number of separable filters b-e, which are shifted in-phase and time, can be summed to form f.

6. PHYSIOLOGICALLY PLAUSIBLE FILTERS

Watson and Ahumada[8] have described how spatiotemporally oriented filters can be constructed by adding together the outputs of two separable filters with appropriate spatiotemporal characteristics. The principle can be extended to include a wide variety of filter combinations; the main thing is to create spatiotemporal orientation. Figures 9b-9f show how one can create a spatiotemporally oriented filter by summing the outputs of four separable filters, which are identical except for a shift in receptive-field center and a temporal delay. (The spatial impulse responses are Gabor functions, and the temporal impulse responses are multistage low-pass filters with a small amount of inhibition.) An approximate quadrature partner for this filter can be constructed by using an odd-asymmetric spatial Gabor function; or (for a cruder approximation) one can simply shift the filter spatially by about 90 deg of phase.

A single separable filter can never be directionally selective, and the minimum that one can get away with is a sum of two separable filters. Unless these filters are carefully designed, the resulting tuning will fall short of the ideal.[8]

There is a particularly elegant way of using separable pairs to construct quadrature pairs tuned for both directions, as is shown in Fig. 10. We start with two spatial impulse responses (Fig. 10a) and two temporal impulse responses (Fig. 10b). In this case, the spatial impulse responses have been chosen as second and third derivatives of Gaussians, and the temporal impulse responses are based on linear filters of the form

$$f(t) = (kt)^n \exp(-ktt)[1/n! - (kt)^2/(n + 2)!], \quad (1)$$

where n takes the values of 3 and 5. There is nothing magical about these particular functions, but they serve as plausible approximations to filters inferred psychophysically.[30]

Now there are four possible ways to combine the two spatial and two temporal filters into separable spatiotemporal filters; let us make all four. These are shown across the top of Fig. 10c. By taking appropriate sums and differences, we can construct the four oriented filters shown across the bottom of Fig. 10c. Two are oriented for leftward motion and two for rightward motion. The two members of each pair are approximately 90 deg out of phase with each other. When their outputs are squared and summed, the resulting signal provides a fairly good measure of the motion energy falling in the range of frequencies for which this detector system is tuned.

Figure 11 shows the spatiotemporal energy spectrum of a motion unit of the sort just described, sensitive to leftward motion. The system extracts energy in the two blobs that lie along a diagonal through the origin; spectral energy along this diagonal corresponds to motion at a given velocity.

The spectrum is not quite so clean as that which could be achieved by summing filters with more-ideal properties or by summing a greater number of separable filters. But the filter will do much the same job in extracting motion energy within its spatiotemporal-frequency band.

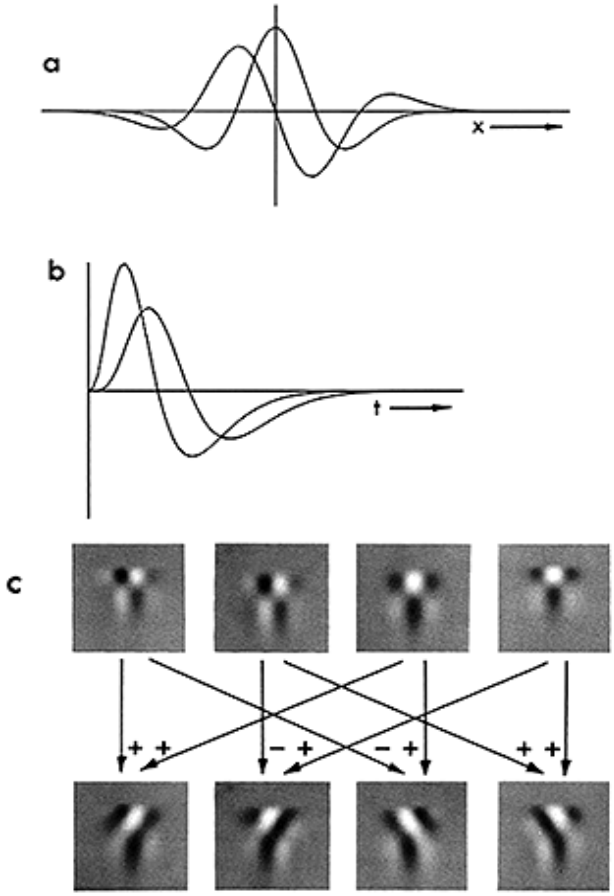

Fig. 10. A method for constructing spatiotemporally oriented impulse responses from pairs of separable ones, following Watson and Ahumada.[8] Two spatial and two temporal impulse responses are shown in a and b. The four spatiotemporal impulse responses shown across the top of c are the products of two spatial and two temporal impulse responses. The ones across the bottom are sums and differences of those above. The result is a pair of leftward- and a pair of rightward-selective filters. Members of a pair are approximately in quadrature.

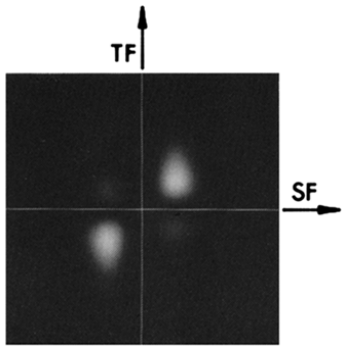

Fig. 11. The spatiotemporal energy spectrum of a direction-selective filter built as the sum of two separable filters.

We can think of a set of spatiotemporally oriented filters as parceling up the spatiotemporal-frequency space into a set of overlapping bands. Figure 12 shows the spectra of three such filters, tuned to rightward motion, leftward motion, and stationary energy (low or zero velocity) and all tuned to the same spatial frequency. Presumably, filters like these cover the entire region of the spectrum that can be seen (the window of visibility of Watson *et al.*). As indicated by the dashed line, this region is bounded by an envelope that is shaped like a blunted diamond in linear coordinates The diamond shape reflects the fact that there is a nearly arithmetic trade-off between spatial and temporal frequency, so that as the spatial frequency is increased, the temporal frequency must be decreased a like amount for the stimulus to remain visible (see, for example, the data of Robson[30] or Kelly[31]). (The diamond shape of the envelope is familiar to display engineers; it allows for the efficient sampling and display of television images through 2:1 interlace. With interlace, one can cut the transmission bandwidth almost in half, while keeping most of the degradations and artifacts outside the diamond. Further exploitation of this shape has been proposed in connection with high-definition television systems[32].) Several investigators have provided evidence on the tuning of mechanisms sensitive to different regions of the spectrum,[33-35] and it appears that a battery of tuned mechamisms parcel the spectrum up. The exact nature of the parceling is not yet clear.

7. MOTION OPPONENCY

The motion detectors that we have described will respond independently to rightward and leftward motion. The existence of such independent channels has been experimentally supported by Levinson and Sekuler[36] and Watson *et al.*[37] (see also Kelly[31] and Stromeyer *et al.*[38]). At the same time, there is reason to believe that motion detection is inherently opponent. First, it is not generally possible to see leftward and rightward motion at the same place and time within the same frequency band: Two sine-wave gratings traveling in opposite directions lead to a perception of a grating flickering in counterphase, as if the rightward and leftward motions had canceled each other out. Second, adaptation phenomena such as the motion aftereffect suggest that motion perception in-

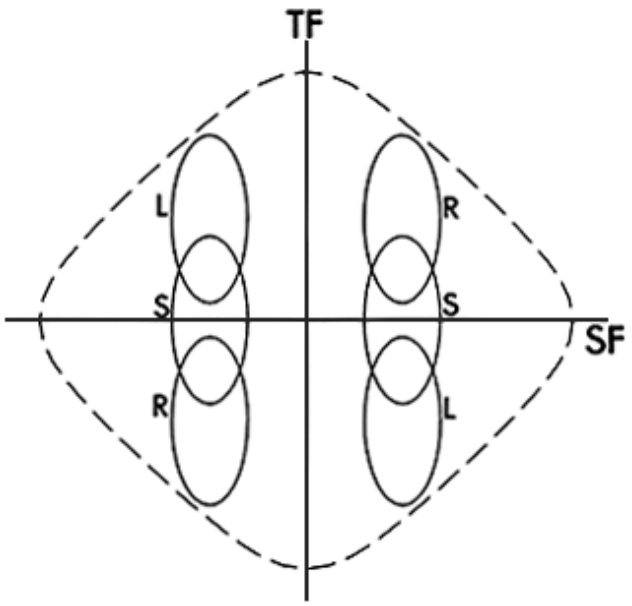

Fig. 12. A spatiotemporal-frequency plot showing the sensitivities of rightward (R), leftward (L) and stationary (S) units. Similar units are presumed to cover the entire visible region of the spatiotemporal spectrum (the bounds of which are shown by the dashed line).

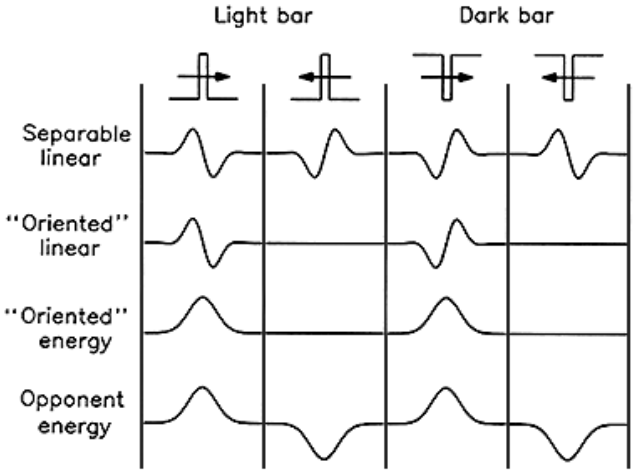

Fig. 13. The time courses of the responses of various stages to a light or a dark bar, moving to the left or to the right. Responses are shown for idealized filters to make the qualitative differences clear. The separable linear stage responds to both polarities and both directions of motion. The spatiotemporally oriented linear stage responds only to rightward motion; the response oscillates and depends on the polarity of the bar. The spatiotemporally oriented energy stage responds to rightward motion only and gives the same positive response regardless of bar polarity. The opponent-energy stage gives a positive response to rightward motion and a negative response to leftward motion, regardless of bar polarity.

volves the balance between opposing leftward- and rightward-motion signals. And third, Stromeyer *et al*.[38] have found that leftward- and rightward-moving gratings can effectively cancel each other's detectability when presented against suprathreshold masks.

All this suggests that the two motion channels may be hooked together in an opponent fashion. A simple opponent-motion channel can be constructed by taking the arithmetic difference between the leftward and the rightward responses. This channel gives a positive output when there is rightward motion, a negative output for leftward motion, and no output for stationary patterns or for counterphase flicker.

We have now completed the basic definition of an opponent-motion energy unit. Figure 13 reviews the stages by which the unit is constructed. To illustrate the properties of each stage, we show the response (over time) of a single unit as a bar is moved across its receptive field. (The response plot is analogous to the record that a physiologist would derive in testing a cell with a moving bar.) The bar can move left or right, and it can be light or dark. Each row shows how a given stage responds to the various moving stimuli.

The first row shows the response of a separable filter. Both rightward and leftward motions lead to strong responses. The next row shows the case of a spatiotemporally oriented (i.e. direction-selective) filter. Now rightward motion gives a good response, whereas leftward motion gives none. (The response shown is for an ideal filter; in a practical filter, there could be a weak response in the reverse direction.) This stage is phase dependent; the response oscillates as the bar makes its traverse, and the response to a light bar has the opposite sign of the response to a dark bar. The third row shows the response of a rightward-moving energy detector, built by squaring and summing the outputs of two spatiotemporally oriented filters in quadrature (and then taking the square root, in the case shown). The response is no longer phase dependent. A right-ward motion leads to a nonoscillating positive response, regardless of whether the bar is light or dark. A leftward motion gives no response. Finally, the last row shows the response of an opponent energy detector, which takes the difference between rightward and leftward energy responses. A rightward motion produces a positive response, whereas leftward motion produces a negative response. The sign of the response reflects the direction of the motion and is independent of the polarity of the bar.

8. EXTRACTING VELOCITY

Although stimulus polarity will not effect the response of a motion energy detector, stimulus contrast will. A given detector will give a weak response if the stimulus is of low contrast or if the stimulus energy happens to fall outside the detector's region of sensitivity. This means that velocity is confounded with contrast (along with spatial and temporal frequency).

If velocity itself is to be extracted, then contrast must somehow be discounted. Presumably the perception of velocity is based not on the response of a single channel but rather on the relative responses of two or more channels (cf. Thompson[39]). Figure 14 suggests a scheme in which velocity is derived by comparing the outputs of several chamnels within the same spatial-frequency band. The three Gaussian-like curves in Fig. 14 represent the sensitivities of a leftward-sensitive, a static, and a rightward-sensitive channel. A single moving grating (indicated by the bold arrow) will stimulate the channels in ratios that are determined by the relative sensitivities of the three channels to the grating's spatial and temporal frequency. If the grating's contrast is changed, the absolute value of the responses will change, but the ratios between them will remain fixed, as long as each channel's response grows in proportion to the contrast of the input. Deviations from proportional growth will cause apparent velocity to change with contrast.[39] (The conditions for invariance are actually broader than this; homogeneity, rather than proportionality, will suffice. If the outputs all pass through a common power function, then their ratios will still remain fixed as contrast is varied.)

The velocity situation may be likened to that in color vision, in which overlapping cone spectral responses can give color information that is invariant with changes in brightness.

At high velocities or low contrasts, the denominator in the ratio can become quite small, and so the velocity estimate will

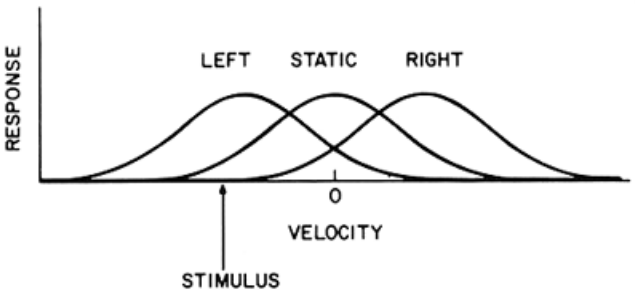

Fig 14. The overlapping response curves of three motion units plotted as a function of velocity. Any single unit's response is a function of both the velocity and the contrast of a stimulus. However, the relative responses of the various units can be used to compute a measure of velocity that is invariant with contrast.

blow up or become unreliable. The visual system must have some means of tagging the velocity estimate with a confidence measure; the simplest approach would be to use the output of the static channel as it stands. High velocities or low contrasts would then lead to low-confidence measures for the velocity of the pattern. When this information was combined with motion information from other channels, its low confidence would cause it to receive a relatively low weight in determining the final motion percept.

9. APPLICATIONS OF SPATIOTEMPORAL ENERGY MODELS

We have described a type of spatiotemporal energy model for an individual motion channel, that is, a channel tuned to a particular band of spatial frequencies. A complete motion percept will be the result of the combined responses of many motion channels, and one must know how these many responses are combined in order to make complete predictions about the perception of moving stimuli. Since we cannot yet offer a thoroughly elaborated model, we will restrict ourselves to considering the responses of individual channels. In this section, we show that the spatiotemporal-energy channels do have many of the basic properties needed for building models of human motion perception.

The pictures that follow are computer simulations of the responses of channels built from the filters of Fig. 10.

A. Continuous Motion

The first requirement of a motion-detecting system, of course, is that it should be able to respond appropriately to ordinary continuous motion. Figure 15a shows the stimulus that was used in Fig. 8a. Figures 15b and 15c show the energy extracted by motion channels sensitive to rightward and leftward motion. Figure 15d shows the difference between the rightward and the leftward responses, i.e., the output of an opponent-motion channel. As we would hope, its response is positive (light) for rightward motion, negative (dark) for leftward motion, and zero (gray) for stationary or blank regions. The output of a stationary channel is shown in Fig. 15e. A measure of velocity (not shown here) can be derived by comparing the outputs of the stationary and the motion channels. Thus the system has the basic qualities that we need.

B. Sampled Motion

Figure 15f shows the same moving edge as in Fig. 15a, but now it is presented as a movie with a moderate frame rate. Figures 15g-15j show the outputs of rightward, leftward, opponent motion and stationary channels for this sampled input. The dominant response is the same as it was for continuous motion. Note, however, that the motion responses are not entirely smooth but fluctuate in synchrony with the frame rate. The static channel shows a similar fluctuation and is stimulated in the midst of the motion. This is consistent with the jerky appearance that movies can have when the frame rate is moderately low.

Thus the spatiotemporal-motion extraction reveals the essential properties of the motion percept that we would like a model to explain. Leftward and rightward motions give rise to leftward and rightward responses, and this occurs by the same mechanism whether the motion is continuous or sampled. If the sampling rate is too slow, the motion will not appear perfectly continuous: rather, a rapid variation will be superimposed upon the dominant motion.

C. Reverse Phi

If a pattern of random black and white bars is moved to the right in steps, it appears, not surprisingly, to be moving to the right (albeit jerkily if the step rate is low). If, on the other hand, the polarity of the bars is changed on each step, so that black bars become white and the white bars become black, then the perceived motion may be reversed: it will now look as if the pattern were moving to the left.[2]

Anstis and Rogers[40] have discussed a model for this effect based on spatial filtering, and Anstis[41] has pointed out that the lower spatial frequencies really are moving backward. The phenomenon can be better understood if one plots the space-time diagrams of the normal and reverse-phi stimuli. Figure 16a shows the case of normal sampled motion; Fig. 16b shows the reverse-phi case. It is clear from glancing at the patterns that the normal case has a great deal of rightward-motion ener-gy, whereas the reverse-phi case has a great deal of leftward en-ergy (in spite of the fact that the reverse-phi pattern was gen-erated by moving the contrast-reversing pattern to the right).

Figures 16c and 16d show the outputs of a motion detector for the two patterns. Not surprisingly, Fig. 16c is light, indicating rightward motion (positive responses are light), whereas Fig. 16d is dark, indicating leftward motion. Note that the response in Fig. 16d is actually rather complex: different amounts of leftward motion are signaled in different regions. These variations in response are sensible if one looks back to the stimulus in Fig. 16b: Different regions should give motion responses of different strengths. Also note that the motion regions themselves move along to the right, even though the regions contain leftward energy. These response properties are roughly consistent with what one often sees when looking at a reverse-phi stimulus. Again, the full motion percept of the reverse-phi stimulus will be the combined result of the activity of many channels with different frequency responses, so that the output of a single channel cannot be used to give a full prediction of the appearamce of the motion. But the spatiotemporal-energy approach does handle the basic phenomenon of direction reversal quite easily.

D. Fluted-Square-Wave Illusion

If a square wave jumps to the right in steps that are 90 deg of its period (i.e., one quarter of a cycle), then it is seen to be doing just that. If, on the other hand, the fundamental component is removed from the square wave, then the resulting wave form (which is like a fluted square wave) appears to be jumping to the left.[11] This phenomenon is reasonable when one considers that the strongest component remaining is the third harmonic and that it moves by 270 deg, or -90 deg, of its own period when the square wave jumps.

Figures 17a and 17b show the spatiotemporal stimuli produced by the ordinary and fluted square waves as they make their rightward jumps. In both cases, the actual motion of the stimuli is rightward, but the fluted square wave appears to be jumping to the left. Figures 17c and 17d show the output of a motion channel, which is sensitive to spatial frequencies in the range containing most of the stimulus en-

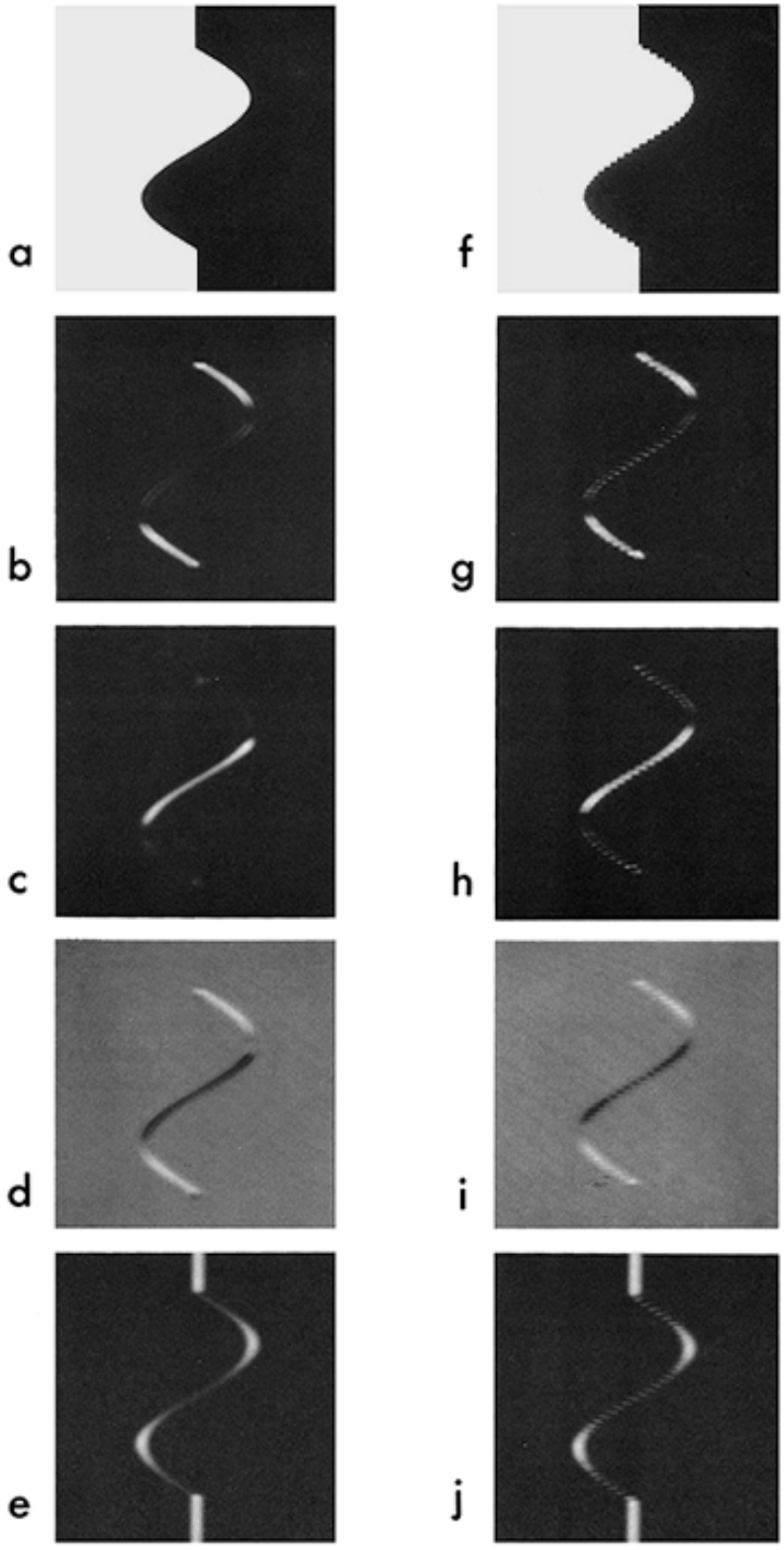

Fig. 15. a, An (x, t) plot of a stimulus consisting of a moving light-dark edge. b, The response of a rightward-energy unit. c, A leftward-energy unit. d, An opponent-energy unit. f, An (x, t) plot of a movie of the stimulus shown in a. g-j, The responses of units selective for rightward energy, opponent energy, and static energy, respectively.

Fig. 16. a, An (x, t) plot of a random bar pattern, moving to the right in steps. b, The reverse-phi version: The pattern moving to the right and the bars reverse polarity on each step. c, The response of an opponent-energy channel to normal motion. Tbe response is mainly positive, signaling rightward motion. d, The response of the channel to the reverse-phi display. Now the response is mainly negative, signaling leftward motion.

Fig. 17. a, An (x, t) plotof a square wave's motion. b, A (x, t) plot of a fluted square wave's motion. c, The response of a medium spatialfrequency opponent-motion channel when stimulated by the square wave. Rightward motion (bright) is signaled. d, The response of the same channel when stimulated by the fluted square wave. Leftward motion (dark) is signaled.

ergy (between the fundamental and third harmonic). In the case of a normal square wave, rightward motion is signaled, as indicated by the bright field of Fig. 17c. In the case of the fluted square wave, leftward motion is signaled, as shown by the dark field of Fig. 17d. Motion channels in other frequency bands will give different responses (some rightward and some leftward), but they will be of lower amplitude. The spatiotemporal-energy analysis, although not offering a full account of the effect, is qualitatively consistent with it.

The fluted-square-wave effect brings up another problem in motion perception: How is the extracted motion perceptually assigned to the forms that are seen in the display? When the fundamental is removed, one sees the fluted grating pattern, and one sees the motion, and one erroneously perceives the entire grating as moving with that motion. The motion percept is correct in the sense that there is real leftward energy in the stimulus, and the form percept is correct in that at any instant the pattern consists of a square wave minus its fundamental; however, it is simply not true that the entire pattern is moving to the left. The only simple percept that can correctly account for all the physical stimulation is the percept of a rightward motion, but this is rarely seen. So motion assignment in this case leads to a percept that contradicts information that is readily available in the stimulus.

10. SUMMARY AND CONCLUSIONS

We have discussed a class of motion models that arise from a simple spatiotemporal conceptualization of motion. A moving pattern may be considered to reside in a three-dimensional space, where the dimensions are x, y, and t. In this space, a moving stimulus is one that is sheared in time, so that its representation is slanted. The problem of detecting motion is then entirely analogous to the problem of detecting orientation in space; the orientation exists in space-time rather than just in space.

Filters with appropriately "oriented" impulse responses (or units with appropriately "oriented" receptive fields) will selectively respond to motion in particular directions. Such filters can be constructed by using simple building blocks such as the separable mechanisms already thought to be present in the visual system.

To extract spatiotemporal energy, filters can be chosen as quadrature pairs and their outputs squared and summed. Thus one can derive a phase-independent-motion energy response by combining the outputs of two linear filters, each sensitive to motion in the same direction but with sensitivities 90 deg out of phase. A compressive nonlinearity (such as a square root) may follow this stage.

Leftward and rightward energy detection can be combined to produce an opponent energy detector. An (R-L) detector gives a positive response to rightward motion and a negative response to leftward motion. A steady motion of an edge or bar leads to a nonoscillating response, the sign of which depends on the direction of the motion and not on the polarity of the stimulus.

The resulting system has many desirable properties (properties that would be useful in any motion-detecting system and that seem to be a part of the human motion-detecting system). The system gives a motion response that is localized in space, time, and spatial frequency; thus a unit's output can be taken as evidence about the direction of motion within a given frequency band at a given location at a given moment in time. The model can be used as a framework in which to understand many basic phenomena in motion perception, including the perception of continuous motion; the perception of so-called apparent motion seen in sampled

displays (e.g., movies); and the perception of various motion illusions, such as the fluted square wave and reverse phi.

Energy-based models lead to a way of thinking about motion that is rather different from some other approaches. Energy models do not solve, but rather bypass, the correspondence problem. Moving stimuli contain motion energy, whether they are displayed continuously or stroboscopically; thus apparent motion (at least in conditions of rapid presentation) can be thought of as a natural and necessary result of extracting motion energy rather than as an illusion actively constructed by a matching mechanism.

Neither does one have to think of a motion detector as computing a change of position over time. No edges are identified, no peaks are localized, and no landmarks are tagged in the extraction of motion energy. Instead, spatiotemporal orientation can be considered to be a local property of spatiotemporal stimuli, and it can be extracted with the same kind of simple mechamisms that are used for extracting spatial orientation.

It is also noteworthy that energy models are closely related to van Santen and Sperling's type of Reichardt model and in some cases are formally identical (see Appendix A). The two kinds of model are thus computing essentially the same thing in different ways. The models suggest complementary ways of thinking about the same issues in rnotion perception.

Energy models do have their limits. They do not seem appropriate for the conditions that Braddick and others have identified with a long-range mechanism; it may well be that more-traditional matching concepts are needed to understand these conditions. And energy models cannot deal with the motion of the energyless beat patterns that arise when two moving gratings (of different frequency or orientation) are summed.[42] But the models do allow one to make sense of some basic phenomena in low-level motion perception. And the spatiotemporal-energy approach provides conceptual tools that may be useful in analyzing a variety of problems in motion perception.

APPENDIX A: FORMAL RELATIONSHIPS BETWEEN ENERGY MODELS AND REICHARDT-TYPE MODELS

One of the classic approaches to motion modeling was introduced by Reichardt[7] and has been recently extended and applied to human motion perception by van Santen and Sperling[6] In a Reichardt-type model, responses from two spatial locations are multiplied together, a lag having been introduced into one of the response pathways before the multiplying stage. This has the effect of correlating the two outputs with a delay. In van Santen and Sperling's model, the input stages include spatial-frequency-tuned receptive fields (such as Gabor functions); pairs that differ in phase or position by about 90 deg are used in building a motion-detecting unit. Similar separable filter pairs can be combined *linearly* to produce direction-selective filters, as described by Watson and Ahumada.[9] Van Santen and Sperling[6] noted that a leftward- and a rightward-sensitive Watzon-Ahumada filter could be used to build a Reichardt-equivalent model if the outputs of the two filters were squared, their difference were taken, and the output were averaged over the entire period of the display. Adelson and Bergen[12] described energy models such as those outlined in this paper and noted that a Reichardt equivalence could be established with four filters (to give quadrature). The use of quadrature requires more filters but avoids the need for time averaging.

Consider the version of a Reichardt model that is shown in Fig. 18(a). (This version is somewhat different from that used by van Santen and Sperling, but it can be conveniently compared with the energy model in Fig. 18(b). We do not assume that the output is a single time average taken over the full display but instead assume that it is a continuously time-varying signal.) A continuous-image sequence $I(x, t)$ is fed into two spatial filters $f_1(x)$ and $f_2(x)$ representing receptive fields that are displaced in position or in phase. The outputs pass through two different temporal filters $h_1(t)$ and $h_2(t)$; one filter delays or low passes the signal more than the other. The four separable combinations of flters lead to the four outputs A(t, $A'(t)$, $B(t)$, and $B'(t)$. The signa $A(t)$, for example, is given by

$$A(t) = h_1(t) * [I(x, t) \cdot f_1(x)],$$

where $*$ indicates convolution in time and $\cdot$ indicates the spatial dot product.

Pairs of these separable responses are multiplied, giving the outputs AB' and BA'; the difference is then taken to produce the final output AB'-BA'. We label these stages as half-phase opponent energy and full opponent energy for reasons that will soon become clear.

Now consider Fig. 18(b), which shows an example of an energy model of the sort described in this paper. Once again, the input signal $I(x, t)$ passes through the spatial and temporal filters to produce the four separable responses A, A', B, and B'. Sums and differences are taken to produce the spatiotemporally oriented linear responses (responses that are selective for direction of motion). Response pairs for leftward and for rightward motion are combined by summing their squares, leading the oriented-energy responses. Opponent energy is then computed with a difference operation. Working out the simple algebra, we find that the final output is just $4(AB'-A'B)$. And this is the same as the output of the Reichardt-type model, except for the scale factor.

Thus a Reichardt model can be thought of as computing the opponent-energy response (or, of course, the energy model can be thought of as computing the Reichardt correlation). Note that, in a Reichardt model, the computations are inherently opponent, and there are no individual responses to leftward and to rightward motion. The terms AB' and BA' do not represent leftward and rightward energy; rather, each is a motion-opponent signal that represents the difference between one spatial phase of the rightward-motion signal and the other the spatial phase of the leftward-motion signal.

Note that arbitrary spatial and temporal impulse responses may be used to form the separable filters in the above discussion. Thus one can build an equivalent energy model for almost any Reichardt-type model, including the original insect-eye models that used Gaussian (low-pass) spatial weighting functions rather than Gabor-like (bandpass) weighting functions.

When an energy model is built with motion-selective filters that are the sum of three or more separable filters (as in Fig. 9f), then the simple equivalence no longer holds. But, even when an energy model has no simple Reichardt equivalent, its behavior may be similar to that of a Reichardt-type model.

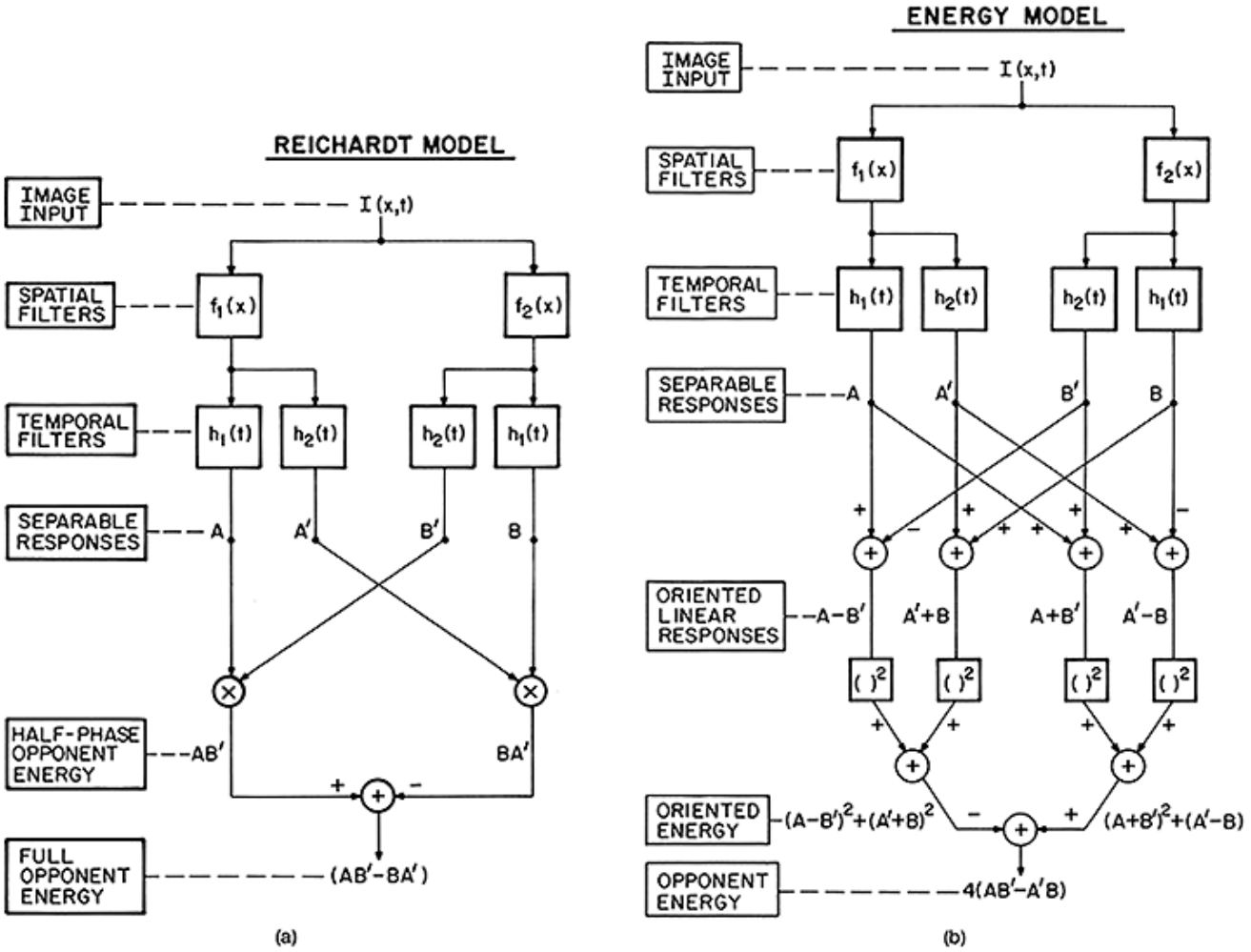

Fig. 18. (a) A version of the Reichardt model that is formally equivalent to a version of an energy model. The visual input $I(x, t)$ passes through the two spatial impulse responses $f_1(x)$ and $f_2(x)$. Following van Santen and Sperling,[6] these functions can be bandpass, differing in phase or in position. Each output passes through the two temporal functions $h_1(t)$ and $h_2(t)$ where $h_2(t)$ is more low passed or more delayed than $h_1(t)$. The four separable responses are labeled A, A', B, and B'. The products AB' and BA' are generated, and their difference constitutes the final output. (b) An equivalent spatiotemporal energy model. The same spatial and temporal filters are used. Sums and differences generate directionally selective filters. Sums of squares of quadrature pairs give motion energy for each direction. The difference between the rightward and leftward signals gives the final output. This turns out to be identical with the output of the Reichardt model. The equivalence holds only for energy models that are based on sums of separable filter pairs.

Distinctions between the models may then be possible only on the basis of physiological and psychophysical experiments that examine motion responses in detail. For example, the independent detection of rightward and leftward motion at threshold[36,37] is fairly easy to accommodate in a spatiotemporal energy model but is not readily accommodated in a Reichardt model. Experiments of this kind exploit nonlinearities such as thresholds, causing the equivalence of the two models to break down. The order in which things are computed does influence the output when thresholds come into play.

But it is more appropriate to stress the strong similarities between these models rather than their differences. In most suprathreshold situations, a spatiotemporal energy model of the sort described here will be experimentally indistinguishable from a model of the sort described by van Santen and Sperling. This is a remarkable fact: Two approaches to motion modeling, motivated by different philosophies, converge on models that are almost identical from a functional point of view. Thus in many situations either model can be used, the choice being determined by conceptual and mathematical convenience. A Reichardt-type model is built of fewer stages than is an equivalent energy model of the sort that we have described and in this sense is simpler; it also appeals to intuitions about matching over time. By the same token, the spatiotemporal-energy approach, which derives its motion selectivity through tuned linear filters, fits in directly with the familiar mathematics of linear systems theory and thus may be easier to apply in many situations. The energy approach also encourages one to develop intuitions about motion as orientation when stimuli are represented in x-y-t space. These intuitions can be quite helpful in thinking about motion.

ACKNOWLEDGMENT

We wish to thank J. A. Movshon, J. P. van Santen, and G. Sperling for helpful discussions.

Note added in proof: It has come to our attention that Fahle and Poggio[43] have previously discussed how sampled motion (and its perception) may be analyzed in the spatiotemporal-frequency domain and have described the construction of spatiotemporally oriented filters as sums of separable pairs.

REFERENCES

1. S. Ullman, *TheInterpretation of Visual Motion* (MIT U. Press, Cambridge, Mass., 1979).
2. S. M. Anstis, "The perception of apparent movement," Phil. Trans. R. Soc. London Ser. B **290**,153-168 (1980).
3. S. M. Anstis, "Apparent Movement," in *Handbook of Sensory Physiology, Vol. VIII, Perception,* R. Held, H. W. Leibowitz, and H.-L. Teuber, eds. (Springer-Verlag, New York, 1977).
4. J. S. Lappin and H. H. Bell, "Perceptual differentiation of sequential visual patterns," Percept. Psychophys. **12**, 129-134
6. D. Marr and S. Ullman, "Direction selectivity and its use in early visual processing," Proc. R. Soc. London Ser. B **211**, 151-180 (1981).
6. J. P. H. van Santen and G. Sperling, "Temporal covariance model of human motion perception," J. Opt. Soc. Am. A **1**, 451-473 (1984).
7. W. Reichardt, "Autocorrelation, a principle for the evaluation of sensory information by the central nervous system," in *Sensory Communication,* W. A. Rosenblith, ed. (Wiley, New York, 1961).
8. A. B. Watson and A. J. Ahumada, Jr., "A look at motion in the frequency domain," NASA Tech. Memo. TM-84352 (1983).
9. J. Ross and D. Burr, "The psychophysics of motion," in *Proceedings of the Workshop* of *Vision, Brain, and Cooperative Computation,* M. A. Arbib and A. R. Hanson eds. (U. Massachusetts Press, Amherst, Mass., 1983); *Vision, Brain, and Cooperative Computation* (Bradford, Amherst, Mass., to be published)
10. M. J. Morgan, "Perception of continuity in stroboscopic motion: a temporal frequency analysis." Vision Res. **19**, 491-500 (1979); "Analogue models of motion perception," Phil. Trans. R. Soc. London Ser. B **290**,117-135 (1980).
11. E. H. Adelson "Some new illusions and some old ones, analyzed in terms of their Fourier components," Invest. Opthalmol. Vis. Sci. Suppl. **22**, 144 (1982).
12 E. H. Adelson and J. R. Bergen, "Spatio-temporal energy models for the Perception of Motion," J. Opt. Soc. Am. **73**, 1861 (1983).
13. C. Enroth-Cugell and J. G. Robson, "The contrast sensitivity of retinal ganglion cells of the cat," J. Physiol. London **187**, 517-552 (1966).
14. J. A. Movshon, I. D. Thompson, and D. J. Tolhurst, "Spatial summation in the receptive fields of simple cells in the cat's striate cortex." J. Physiol (London) **283**, 79-99 (1978).
15. G. W. Campbell and J. G. Robson, "Application of Fourier analysis in the visibility of gratings," J. Physiol. (London) **197**, 551-566 (1968).
16. H. R. Wilson, and J. R. Bergen, "A four mechanism model for threshold spatial vision," Vision Res. **19**,19-33 (1979).
17. O. Braddick, "A short-range process in apparent motion," Vision Res. **14**, 519-529, (1974); "Low-level and high-level processes in apparent motion," Phil. Trans. R. Soc. London B **290**,137-151 (1980).
18. J. Hochberg, and V. Brooks, "The perception of motion pictures," in *Handboak of Perceptian,* E. C. Carterette and M. Friedmen, eds. (Academic, New York, 1978), Vol. 10.
19. G. Sperling, "Movement perception in computer-driven visual dis plays," Behav. Res. Methods Instrum. **8**, 144-151 (1976).
20. P. Burt and G. Sperling, "Time, distance, and feature trade-offs in visual apparent motion," Psych. Rev. **88**, 171-195 (1981).
21. A. J. Pantle and L. Picciano, "A multi-stable movement display: Evidence for two separate motion systems in humans," Science **193**, 500-502 (1976).
22. D. E. Pearson *Transmisssion and Display* of *Pictoral information* (Wiley, New York, 1975).
23. A. B. Watson, A. Ahumada ,Jr. and J. E. Farrell, "The window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays," NASA Tech. Paper TP-2211 (1983).
24. D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurons in the cat's striate cortex," J. Physiol. (London) **148**, 574-591 (1959).
25. D. H. Tolhunst and J. A. Movshon, "Spatial and temporal contrast sensitivity of striate cortical neurons," Nature **257**, 674 675 (1975).
26. J. A. Movshon, I. D. Thompson, and D. J. Tolhurst, "Spatial and temporal contrast sensitivity of neurons in areas 17 and 18 of the cat's visual contex," J. Physiol. (London) **283**, 101-120 (1978).
27. A. B. Watson and A. J. Ahumada, Jr., "A model of how humans sense image motion," Invest. Opthalmol. Vis. Sci. Suppl. **25**, 14 (1984).
28. A. Pantle and R. Sekuler, "Contrast response of human visual mechanisms sensitive to orientation and motion." Vision Res. **9**, 397-406 (1969).
29. J. R. Bergen and H. R. Wilson, "Prediction of flicker sensitivities from temporal three pulse data," Vision Res. (to be published).
30. J. G. Robson, "Spatial and temporal contrast sensitivity function of the visual system," J. Opt. Soc. Am. **56**, 1141-1142 (1966).
31. D. H. Kelly, "Motion and vision, II. Stabilized spatio-temporal threshold surface" J. Opt. Soc. Am. **69**, 1340-1349 (1979).
32. T. J. Long "Why not compatible high-definition television?" IBA Tech Rev. **21**, 4-12 (1983), T. S. Robson, "Extended-definition television service," Proc. IEEE **129**, 485-489 (1982).
33. A. B. Watson and J. G. Robson, "Discrimination at threshold: labelled detectors in human vision," Vision Res. **21**, 1115-1122 (1981).
34. P. Thompson, "The coding of velocity of movement in the human visual system," Vision Res. **24**, 41-45 (1984).
35. D. J. Tolhurst, "Sustained and transient channels in human vision," Vision Res. **15**, 1151-1155 (1975).
36. E. Levinson and R. Sekuler "The independence of channels in human vision selective for direction of movement," J. Physiol. London **250**, 347-366 (1975).
37. A. B. Watson, P. G. Thompson, B. J. Murphy, and J. Nachmias, "Summation and discrimination of gratings moving in opposite directions," Vision Res. **20**, 341-347 (1980).
38. C. F Stromeyer III, R. E. Kronauer, J. C. Madsen, and S. A. Klein, "Opponent mechanisms in human vision," J. Opt. Soc. Am. A **1**, 876-884 (1984).
39. P. Thompson, "Perceived rate of movement depends on contrast," Vision Res. **22**, 877-380 (1982).
40. S. M. Anstis and B. J. Rogers, "Illusory reversal of visual depth and movement during changes of contrast," Vision Res. **15**, 957-961 (1975).
41. S. M. Anstis, Department of Psychology York University, Toronto, Ontario, Canada (personal communication, 1981).
42. E. H. Adelson and J. A. Movshon "Phenomenal coherence of moving gratings," Nature **200**, 523-525 (1982).
43. M. Fahle and T. Poggio, "Visual hyperacuity: spatio-temporal interpolation in human vision," Proc. R. Soc. London Ser. B **213**, 451-477 (1981).