Fitting Models to Distributed Representations of Vision

Sourabh A. Niyogi
Department of Electrical Engineering and Computer Science
MIT Media Laboratory, 20 Ames Street, Cambridge, MA 02139

Abstract

Many representations in early vision can be constructed by performing orientation analysis along several sampling dimensions. Texture is often oriented in space, motion is oriented in space-time, and stereo is oriented in space-disparity. In these modalities, we can construct distributed representations with oriented energy measures used in models of biological vision. Surface models of orientation, velocity, and disparity can easily be fit to distributed representations of texture, motion, and stereo by combining tools of orientation analysis and regularization. We describe base representation construction and model fitting processes in these domains.

1 Distributed Representations

Conventional approaches to early vision have focussed on recovering explicit models of local structure and subsequently fitting models to local structure. When the assumptions made in modeling local structure hold, the local structure is successfully recovered. When model assumptions are violated, however, local structure is incorrectly recovered in unstructured ways. For example, conventional motion estimation algorithms assume that there is a single motion, that the intensity is constant in space time, and that the flow field is smooth. These assumptions are violated, for example, at motion boundaries. At motion boundaries, two motions are present, intensity is not conserved, and the flow field is discontinuous. Consequently, conventional approaches of recovering flow fields fail at occlusion boundaries. Various techniques have been proposed to counter these assumptions: regularizing with line processes [Poggio *et* a/., 1985], generalizing local models [Shizawa and Mase, 1991], using global parametric models [Bergen *et* al., 1992], and using robust estimation [Black and Anandan, 1993].

Alternatively, we can use *distributed* representations of vision, where a population of responses implicitly represents local structure. This strategy is used in visual cortex, where neurons do not represent parameters explicitly, but instead are *tuned* to stimulus parameters of spatiotemporal frequency, horizontal disparity, velocity, global motion patterns, and so forth. Several have produced models of tuning properties of cells in the visual cortex, in parallel to conventional approaches of modeling and recovering local structure in computer vision: spatially oriented filters model orientation selectivity of simple cells [Bergen and Landy, 1991]; spatiotemporally oriented filters model directional selectivity [Adelson and Bergen, 1985]; stereo spatial filters out of phase model disparity tuning [Qian, 1994; Sanger, 1988]; and combining the outputs spatiotemporally oriented filters models velocity tuning [Grzywacz and Yuille, 1990; Heeger, 1987; Simoncelli, 1993].

The construction of distributed representations in different modalities becomes remarkably similar once observing that most local structure in visual stimuli is oriented along several sampling dimensions [Adelson and Bergen, 1991]:

- *Space.* Spatial textures and edges are oriented in space, where x = *(x,y).*
- *Space-time.* Motion is orientation in space-time, where x = (x:,t) or x = *(x,y,t)* in one and two dimensional motion (see [Adelson and Bergen, 1985]).
- *Space-disparity.* Stereo is orientation in space-disparity, where x = (x,V_x) or x = (z,y, V_x) for one and two-dimensional stereo analysis. Typically, two samples of V_x form a stereo pair.

Figure 1 illustrates oriented structures in each of these dimensions. In the above modalities, constructing distributed representations of orientation is possible by filtering a stimulus /(x) with multiple oriented filters tuned to different orientations in space, space-time, or space-disparity. Each spatiotemporally oriented filter signals the presence of a given orientation; by having banks of oriented filters tuned to different orientations, a population of responses can be obtained to represent local structure. Figure 2(a,b) shows an *{x,t)* stimulus with occluding and transparent motions, and a slice of the base representation $E_a(x,t, 6)$ we can construct from the stimulus using oriented filters. Where there is transparency, a bimodal distribution of energy is present at each point x, and where there is occlusion, a bimodal distribution of energy is present to both sides of x. While it is unclear how properties such as color and shading could be encoded in distributed representations, we can take a unified approach for the domains of texture, motion and stereo.

Rather than fitting models to *explicit* representations of orientation, such as range data or optical flow, we fit models to *implicit, distributed* representations of vision. Distributed representations are advantageous because they: (1) represent complex naturally occurring stimuli, such as transparency and occlusion; (2) represent uncertainty implicitly; (3) degrade gracefully with noise; (4) do not *commit* to an (often wrong) answer, unlike explicit representations; (5) are also used in biological visual systems.

The problem remains, however, of how to utilize distributed representations of vision to recover information about global surfaces. Most *of* the past approaches used distributed representations to estimate local structure, in the domains of texture [Kass and Witkin, 1987], stereo [Sanger, 1988; Qian, 1994] and motion [Fleet and Jepson, 1990; Grzywacz and Yuille, 1990; Heeger, 1987]. In regions containing little or no texture, filter responses are negligible, so approaches which attempt to use distributed representations to estimate local structure will fail. Thus the *surface interpolation* problem remains.

The surface interpolation problem in *explicit* representations is typically solved by taking sparse representations of explicit local measurements, and smoothly interpolating between them. Canonical approaches to surface interpolation (e.g. [Terzopoulos, 1988]) use both a fidelity criteria, describing the fit of the surface to sparse measurements, and a smoothness criteria so as to constrain regions without measurements. Typically, weighted combinations of the square of the first and second derivatives of surfaces are used as a measure of smoothness. By combining tools of orientation analysis with simple modifications of standard regularization processes, we show that it is straightforward to fit models to distributed representations.

Rather than recovering explicit local structure, and *then* fitting surface models to local structure, we fit models to distributed representations of vision directly. The generic approach is contrasted with our approach in Figure 3(a,b). We describe base representation construction and our adaptation of regularization in the domains of space, space-time, and space-disparity, and demonstrate preliminary success on simple imagery.

2 Fitting Models

Our task is to fit models to time varying visual stimuli $I(\mathbf{x}) = I(x, y, t, V_x)$ with sheet models $\mathbf{u}(u, t)$; a list of the domains we consider is shown in Figure 4. For one-dimensional problems, $\mathrm{u} = u$ tessellates spatial variable x in $\mathbf{x}$ when $\mathbf{x} = (x, t)$ or $\mathbf{x} = (x, V_x)$. For two-dimensional problems, $\mathrm{u} = (u, v)$ tessellates the spatial variables (x, y) in $\mathbf{x}$ when $\mathbf{x} = (x, y)$, $\mathbf{x} = (x, y, t)$, or $\mathbf{x} = (x, y, V_x)$.

For sheet model fitting problems, the surface is modeled with an array of nodes, where at each node, the spatial coordinates and an orientation Θ is stored in the combined model $\mathbf{u}(\mathrm{u}, t)$. The goal is to fit a surface $\Theta(\mathrm{u}, t)$ to a base representation $E_\sigma(\mathbf{x}, \Theta)$ indicating the presence of the orientation Θ at point $\mathbf{x}$. For two-dimensional texture, the orientation $\Theta = \theta$ is lo-

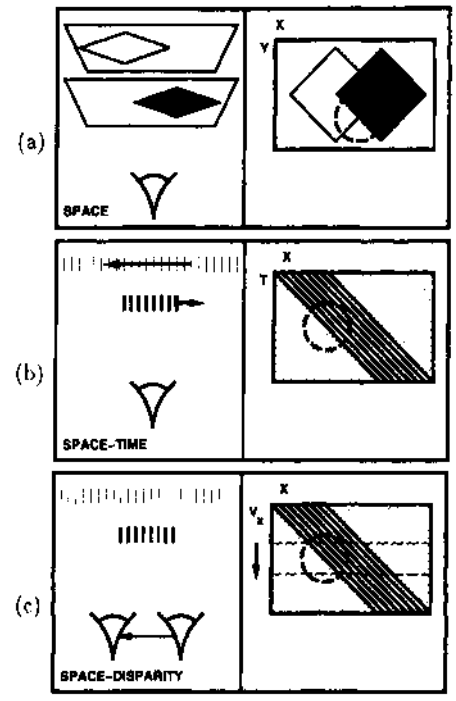


Figure 1: *Orientation analysis in three modalities.* (a) Edges and texture are oriented in space (x, y); (b) Motion is orientation in space-time (x, t); (c) Stereo is orientation in space-disparity (x, V_x).

Figure 2: *Example stimulus.* (a) Stimulus $I(x, t)$; (b) Base representation $E_\sigma(x, v_x, t)|_{t=t_0}$. From [Simoncelli, 1993], with permission.

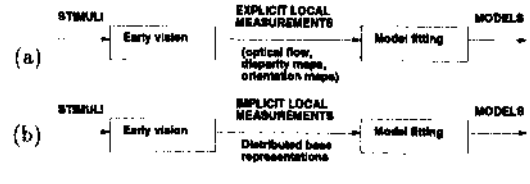


Figure 3: *Architectures,* (a) Standard computer vision approaches to fit models to explicit representations of local structure, which is prone to failure where explicit representations of local structure cannot be accurately recovered, (b) We advocate fitting models to distributed representations of local structure.

cal spatial orientation; for one-dimensional and two-dimensional stereo, the orientation $\Theta = V_x$ is horizontal disparity; for one-dimensional motion, the orientation $\Theta = v_x$ is horizontal motion; for two-dimensional motion, the orientation $\Theta = (v_x, v_y)$ is a two dimensional vector.

The model u(u,t) indexes into the base representation $E_\sigma(\mathbf{x}, \Theta)$, so that a total "energy" of a surface can be obtained. We seek to fit a model where the population response in the base representation $E(\mathbf{x}, \Theta)$ across the surface is maximized, *and* where the magnitudes of the derivatives in the orientation surface are minimal, as in standard regularization processes (c.f. [Terzopoulos, 1988]). Since multiple surfaces can be fit, no answer is correct; the only satisfactory answer will require the observer to shift attention between portions of the base representation.

Models are fit over multiple scales by constructing a scale-space:

$$I_\sigma(\mathbf{x}) = G_\sigma(\mathbf{x}) * I(\mathbf{x})$$

where $G_\sigma(\mathbf{x})$ represents a Gaussian convolution kernel of width σ. Scale-space is approximated with a Gaussian pyramid at discrete scales of σ. An *N*-dimensional interpolatable base distributed representation $E_\sigma(\mathbf{x}, \Theta)$ is constructed from I0(x) through orientation analysis. Steerable [Freeman and Adelson, 1991; Koenderink, 1992; Perona, 1992] filters allow for continuous interpolation to any Θ. The stimulus at a given scale is used to compute n steerable "basis images" $J^i_\sigma(\mathbf{x})$ via convolution with a set of basis filters g_i,(x):

$$J^i_\sigma(\mathbf{x}) = g_i(\mathbf{x}) * I_\sigma(\mathbf{x})$$

$$E_\sigma(\mathbf{x}, \Theta) = f(J^1_\sigma(\mathbf{x}), \ldots, J^n_\sigma(\mathbf{x}), \Theta)$$

where Θ represents spatial orientation, spatiotemporal orientation, velocity, or disparity, depending on the sampling dimensions being considered. We will describe the details of base representation construction in subsequent sections.

Models can be fit these base representations by modifying standard regularization procedures *(c.f.* [Terzopoulos, 1988]) to use this base representation to dynamically estimate the state of a model u(u, *t):*

$$\dot{\mathbf{u}} = \mathbf{K}\mathbf{u} + \mathbf{f}$$

where assumptions of smoothness in the nth order spatial derivatives of orientation are embedded within the stiffness matrix K, and base representations $E_\sigma(\mathbf{x}, \Theta)$ are used to apply observation "forces" f(u,t) to the surface by integrating forces through discrete summation over several scales:

$$\mathbf{f}(\mathrm{u}, t) \approx \mathbf{H}^T \sum_\sigma \left. \frac{\partial E_\sigma(\mathbf{x}, \Theta)}{\partial \Theta} \right|_{\dot{\mathbf{u}}(\mathrm{u},t)}$$

Here, H indexes the orientation surface component $\Theta(\mathbf{u}, \mathbf{t})$ of the model ii(u,t). We can easily modify the above to introduce additional "forces" to shift attention from certain portions of base representation to others, as in the deformable contours [Kass *et a*/., 1987].

Thus we employ the following procedure to fit models u to stimuli J(x):

Stimulus $I(\mathbf{x})$, Model u(u)
Space (2-d) $I_\sigma(x, y)$
Model $\mathbf{u}(u,t) = [x(u,v,t), y(u,v,t), \theta(u,v,t)]^T$
Space-time (1-d) $I_\sigma(x, t)$
Model $\mathbf{u}(u,t) = [x(u,t), v_x(u,t)]^T$
Space-disparity (1-d) $I_\sigma(x, V_x)$
Model $\mathbf{u}(u,t) = [x(u,t), V_x(u,t)]^T$
Space-time (2-d) $I_\sigma(x, y, t)$
Model $\mathbf{u}(u,v,t) = [x(u,v,t), y(u,v,t), v_x(u,v,t), v_y(u,v,t)]^T$
Space-disparity (2-d) $I_\sigma(x, y, V_x)$
Model $\mathbf{u}(u,v,t) = [x(u,v,t), y(u,v,t), V_x(u,v,t)]^T$

Figure 4: *Fitting models to distributed representations of vision.*

1. Approximate scale-space $I_\sigma(\mathbf{x})$ through construction of a Gaussian pyramid of I(x).
2. Over all scales sigma, construct base representation $E_\sigma(\mathbf{x}, \Theta)$ by convolving $I_\sigma(\mathbf{x})$ with *n* basis filtersg i(x) to compute $J^i_\sigma(\mathbf{x})$, where *i* = 1, . . . , *n.*
3. Initialize model u.
4. Continually subject model u to data forces *f(u,t)* derived from base representation $E_\sigma(\mathbf{x}, \Theta)$

Below we describe base representation construction, and methods to apply observation forces derived from base representations in the domains of texture, motion, and stereo. We illustrate model fitting processes on canonical examples in our preliminary experiments here.

2.1 Two-dimensional space

For two-dimensional spatial texture, x = *(x, y),* and 6 = θ, where θ represents the local orientation of the texture. A single 2-d sheet model is superimposed on the data:

$$\mathbf{u}(u,v,t) = [x(u,v,t), y(u,v,t), \theta(u,v,t)]^T$$

Oriented energy [Bergen and Landy, 1991] measures are used to construct the base representation $E_\sigma(x, y, \theta)$ via:

$$E_\sigma(x,y,\theta) = (I(x,y) * G_\theta(x,y))^2 + (I(x,y) * H_\theta(x,y))^2$$

$$= (\sum_{i=1}^{i_0} J^i_\sigma(x,y) f_i(\theta))^2 + (\sum_{i=i_0+1}^{n} J^i_\sigma(x,y) f_i(\theta))^2$$

where the spatial stimulus $I_\sigma(x, y)$ at multiple scales is used to compute a set of steerable "basis" images $J^i_\sigma(x, y)$ with:

$$J^i_\sigma(x,y) = g_i(x,y) * I_\sigma(x,y)$$

Filters *gi(x, y)* form the steerable separable basis for the quadrature pair of filt $G_\theta(x, y)$i and $H_\theta(x, y)$, which are interpolated to any orientation θ with interpolation functions $f_i(\theta)$; refer to [Freeman and Adelson, 199l] for filter taps and interpolation functions for all of our experiments; we use a second derivative of Gaussian as a basis in our examples except where noted. An example stimulus, basis images, and the base representation are shown in Figure 5.

Forces are applied to the model via the gradient of $E_\sigma(x, y, \theta)$ integrated over all scales:

$$\mathbf{f}(u,v,t) = \mathbf{H}^T \sum_\sigma \left. \frac{\partial E_\sigma(x,y,\theta)}{\partial \theta} \right|_{\dot{\mathbf{u}}(u,v,t)}$$

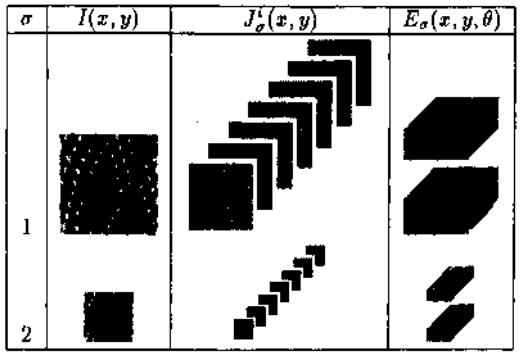


Figure 5: $I(x, y)$ *Base representation construction.*

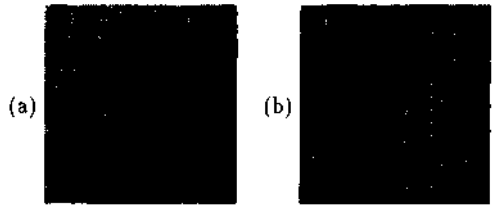


Figure 6: *Two-dimensional texture.* (a) A simple oriented texture; (b) Orientation map depiction of the fitted model.

here H = [0I0II] indexes into the surface 6(x,y) component of n(u,/). To avoid singularities due to the discontinuous nature of angles, a vector n — [cos(0),sin(0)]^T is used to represent the state of the model, and is continuously normalized to one (n0\ — 1). We illustrate the analysis of oriented texture patterns containing simple deformations in texture orientation. Figure 6(a) shows one of several sample oriented patterns we have experimented with. Figure 6(b) shows an orientation map recovered using our approach. With multiple orientations in the stimulus, either multiple models must be fit, or an attentional mechanism to shift the surface is necessary. With detected orientation stopping (c.f. [Heitger, 1992]), smoothing processes should be halted.

2.2 One-dimensional space-time

For one-dimensional space-time, x = (x,t) and 6 = v_x, where v_x represents local 1-d motion. A time-varying 1-d sheet model is imposed on the data:

$$\mathbf{u}(u,t) = [x(u,t), v_x(u,t)]^T$$

For spatiotemporal stimuli /(x,t), motion is orientation in space-time [Adelson and Bergen, 1985]; the motion v_x is related to the spatiotemporal orientation 6 via $v_x = \sigma \tan(\theta)$.

Just as in (x, y), oriented energy measures are used for base representation construction in (x,t):

$$E_\sigma(x,t,\theta) = (I_\sigma(x,t) * G_\theta(x,t))^2 + (I_\sigma(x,t) * H_\theta(x,t))^2$$
$$= (\sum_{i=1}^{i_0} J^i_\sigma(x,t) f_i(\theta))^2 + (\sum_{i=i_0+1}^{n} J^i_\sigma(x,t) f_i(\theta))^2$$

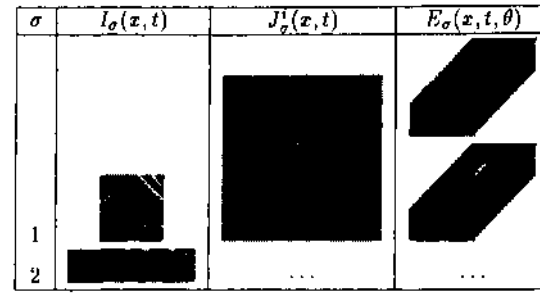


Figure 7: $I(x,t)$ *Base representation construction.*

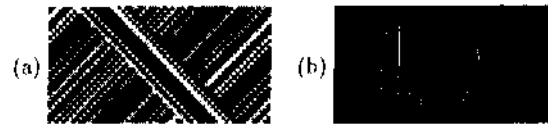


Figure 8: One-dimensional motion, (a) Stimulus I(x,t) - a one-dimensional surface is moving to the right in front of a background moving to the left; (b) Spatiotemporal orientation recovered for the stimulus for a given instant in time. Leftward vs. rightward motion is successfully extracted.

where $J^i_\sigma(x,t)$ are steerable basis images:

$$J^i_\sigma(x,t) = g_i(x,t) * I_\sigma(x,t)$$

An example stimulus, basis images, and its base representation are shown in Figure 7.

Forces f (u, t) are applied to the model via the gradient of $E_\sigma(x,t,\theta)$ integrated over all scales:

$$\mathbf{f}(u,t) = \mathbf{H}^T \sum_\sigma \frac{\partial E_\sigma(x,t,\theta)}{\partial \theta} \frac{\partial \theta}{\partial v_x}\Big|_{\hat{\mathbf{u}}(u,t)}$$

Here, H = [0|I] indexes into the surface v_x(x,t) component of the model u(u,/). Figure 8(a) shows a simple stimulus where one surface is moving rightward occluding a background moving leftward. Figure 8(b,c) shows the sheet model fit via our approach for the stimulus for a given instant in time.

2.3 Two-dimensional space-time

In two dimensional space-time, $\mathbf{x} = (x, y, t)$ and $\Theta = (v_x, v_y)$, where $\vec{v} = [v_x, v_y]^T$ represents local 2-d motion. A sheet model is imposed on the data:

$$\mathbf{u}(u,v,t) = [x(u,v,t), y(u,v,t), v_x(u,v,t), v_y(u,v,t)]^T$$

Now simple orientation analysis is inadequate; a two-stage calculation is necessary [Heeger, 1987]. The first stage involves filtering an image sequence with spatiotemporally oriented filters. The second stage pools the squared responses of these filters to construct units tuned to velocity. Stimuli translating with velocity $\vec{v} = [v_x, v_y]^T$, when viewed in the frequency domain, contain energy along the plane $w_x v_x + w_y v_y + w_t = 0$.

Constructing $E_\sigma(x, y, t, v_x, v_y)$ is possible through linear filtering of the stimulus $I(x, y, t)$ with banks of spatiotemporal filters centered on this plane [Heeger, 1987; Simoncelli, 1993].

We utilize the method developed in [Simoncelli, 1993] to extract energy along these planes in the frequency domain, using tools of steerable filtering. We summarize the method below. Convolution of an image sequence $I(\mathbf{x})$ with a set of basis filters $g_i(\mathbf{x})$ allows for steerable filtering to any portion of the frequency domain $\mathbf{w}_j = \pm[w_1, w_2, w_3]^T$, where $|\mathbf{w}| = 1$, with interpolation functions $f_n(\mathbf{w})$:

$$F_{\mathbf{w}_j}(x, y, t) = \sum_i f_i(\mathbf{w}_j) J_\sigma^i(x, y, t)$$

$$J_\sigma^i(x, y, t) = g_i(x, y, t) * I_\sigma(x, y, t)$$

Steering functions are summarized in Figure 9(a) for Nth order derivative of Gaussians ($N = \{1, 2\}$). Filter coefficients used in these (x, y, t) experiments can be found in [Niyogi, 1995b]. To tile the plane $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$, energy is extracted with $N + 1$ weightings centered at locations $\mathbf{w}_j$, as summarized in Figure 9(a). Figure 9(b) depicts energy extraction for a first derivative of a 3-d Gaussian (N=1); two separate weightings are required to tile the plane. Choosing a different scale of Gaussian extracts a different sized ring centered around the plane. By filtering the image sequence $I(\mathbf{x})$ with filters with frequency selectivities centered at locations $\mathbf{w}_j$, and subsequently applying squaring operations, we extract energy $E_{\vec{v}_n}(\mathbf{x})$ along a ring on the plane via:

$$E_\sigma(x, y, t, v_x, v_y) = G_0(\mathbf{x}) * \sum_{j=1}^{N+1} \left(F_{\mathbf{w}_j}(\mathbf{x})\right)^2$$

High frequency components introduced by squaring are diminished through subsequent convolution with a 3-d Gaussian $G_0(\mathbf{x})$. An example stimulus, two of the basis image sequences used to construct the base representation, and the base representation sampled at one particular velocity is shown in Figure 10.

Time-varying forces $\mathbf{f}(u, v, t)$ are applied to the model via:

$$\mathbf{f}(u, v, t) = \mathbf{H}^T \sum_\sigma \left[\begin{array}{c} \frac{\partial E_\sigma}{\partial v_x}(x, y, v_x, v_y, t) \\ \frac{\partial E_\sigma}{\partial v_y}(x, y, v_x, v_y, t) \end{array} \right]^T \Bigg|_{\dot{\mathbf{u}}(u,v,t)}$$

We illustrate this approach on a simple example. Figure 11(a) shows an (x, *y,t)* cube representation of a simple stimulus where two moving surfaces occlude one another over a static background. Figure 11(b) shows recovered velocity flow field for the stimulus; the motion is extracted for the rightward and leftward moving bar correctly, although oversmoothing is present at the boundaries as in conventional approaches. We have designed methods to detect motion boundaries defined by kinetic occlusion (c.f. [Niyogi, 1995a]) using distributed representations of motion; detected motion boundaries should constrain smoothing processes.

(a) $\mathbf{w}_a = \frac{1}{\sqrt{v_x^2+v_y^2}}[-v_y, v_x, 0]^T, \mathbf{w}_b = \mathbf{n}_{\vec{v}_n} \times \mathbf{w}_a$

N	Steering function $F_{\mathbf{w}_i}(\mathbf{x}), \mathbf{w}_i (i = 1 \ldots N+1)$
1	$F_{\mathbf{w}_i}(\mathbf{x}) = I_x(\mathbf{x})w_1 + I_y(\mathbf{x})w_2 + I_t(\mathbf{x})w_3$ $\mathbf{w}_1 = \mathbf{w}_a, \quad \mathbf{w}_2 = \mathbf{w}_b$
2	$F_{\mathbf{w}_i}(\mathbf{x}) = I_{xx}(\mathbf{x})w_1^2 + 2I_{xy}(\mathbf{x})w_1w_2 + 2I_{xt}(\mathbf{x})w_1w_3 + I_{yy}(\mathbf{x})w_2^2 + 2I_{yt}(\mathbf{x})w_2w_3 + I_{tt}(\mathbf{x})w_3^2$ $\mathbf{w}_1 = \mathbf{w}_a, \mathbf{w}_2 = \frac{1}{2}\mathbf{w}_a + \frac{\sqrt{3}}{2}\mathbf{w}_b, \mathbf{w}_3 = \frac{\sqrt{3}}{2}\mathbf{w}_a + \frac{1}{2}\mathbf{w}_b$

(b) 

Figure 9: $I(x, y, t)$ *motion energy extraction.*

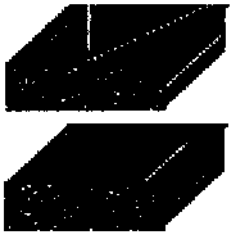
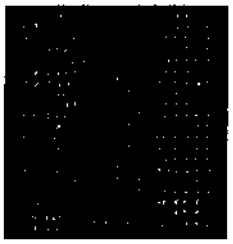
(a)  (b) 

Figure 11: *Two-dimensional motion.* (a) Stimulus *I(x, y,t)* - two bars moving across the field of view move in front of a static background; (b) Recovered model (velocity flow field) for a particular instant in time.

2.4 Two-dimensional space-disparity

In two-dimensional space-disparity, $\mathbf{x} = (x, y, V_x)$ and $\Theta = V_x$, where where V_x represents horizontal disparity. A stereo pair $I_\sigma(x, y, V_x)$ is sampled at two positions of V_x to yield two signals $I_\sigma^l(x, y)$ and $I_\sigma^r(x, y)$. We construct a two-dimensional $I(x, y, V_x)$ window from the stereo pair with:

$$I_\sigma(x, y, V_x) = \begin{cases} I_\sigma^l(x, y) & V_x < 0 \\ \frac{1}{2}(I_\sigma^l(x, y) + I_\sigma^r(x, y)) & V_x = 0 \\ I_\sigma^r(x, y) & V_x > 0 \end{cases}$$

At each scale σ, the disparity V_x is related to the local orientation θ in the (x, V_x) direction extracted from this window by $V_x = \sigma \tan(\theta)$. A 2-d surface model is imposed on the data:

$$\mathbf{u}(u, v, t) = [x(u, v, t), y(u, v, t), V_x(u, v, t)]^T$$

Oriented energy measures are used to construct the base representation $E_\sigma(x, y, \theta)$:

$$E_\sigma(x, y, \theta) = (I_\sigma(\mathbf{x}) * G_\theta(\mathbf{x}))^2 + (I(\mathbf{x}) * H_\theta(\mathbf{x}))^2\big|_{V_x=0}$$

$$= (\sum_i J_\sigma^i(x, y) f_i(\theta))^2 + (\sum_i J_\sigma^i(x, y) f_i(\theta))^2 \Big|_{V_x=0}$$

where the space-disparity stimulus $I_\sigma(x, y, V_x)$ at multiple scales σ is used to compute a set of steerable "basis" images $J_\sigma^i(x, y)$:

$$J_\sigma^i(x, y) = g_i(x, y, V_x) * I_\sigma(x, y, V_x)|_{V_x=0}$$

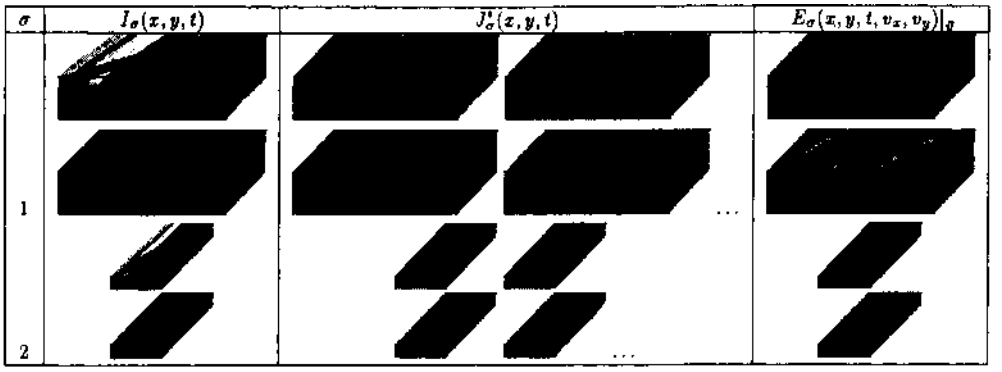


Figure 10: $I(x,y,t)$ *Base representation construction.*

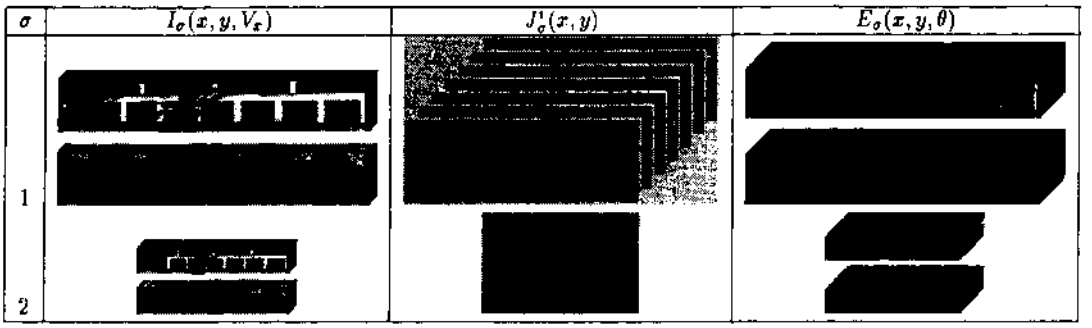


Figure 12: $I(x,y,V_x)$ *Base representation construction.*

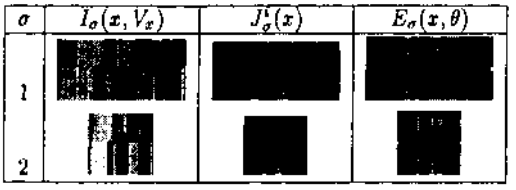


Figure 13: $I(x,V_x)$ *Base representation construction.*

Figure 14: *Two-dimensional stereo,* (a) A stereo pair $I_l(x,y)$ and $I_r(x,y)$\ observer is fixated on a plane with two bars in front of and behind this plane; (b) Recovered disparity between the left and right views.

Now, $gi(x,y,V_x)$ — $6(y)$ * $gi(x,V_x)$ forms an approximation to $G_0(x)$ and $H_0(x)$, but only the center slice (V_x =. 0) of the basis is used to represent the space-disparity orientation (x, V_x) at each point in space (x, y). An example stimulus, basis signals used to simulate the base representation, and a constructed base representation are shown in Figure 12. A slice at a given height *y* of the above signals is shown in Figure 13; fitting models to distributed representations of one-dimensional space-disparity signals would operate on these signals.

Forces $f(u,v,t)$ are applied to the model via the gradient of $E_a(x,y,0)$ integrated over all scales *a:*

$$\mathbf{f}(u,v,t) = \mathbf{H}^T \sum_\sigma \frac{\partial E_\sigma(x,y,\theta)}{\partial \theta} \frac{\partial \theta}{\partial V_x}\bigg|_{\hat{\mathbf{u}}(u,v,t)}$$

Figure 14(a) shows a simple stimulus where the observer is fixated on a plane, with a bar on the left and right closer and further away from the observer than the plane. Figure 14(b) shows the sheet model fit for stimulus.

3 Discussion

In addition to sheet models, which are directly attached to image data, other models can be fit. Surface models, attached only to a portion of the visual field, would be required to recover the parameters of an object. We have not dealt with the problem of model initialization. To do so, we require a model of attention; extensions to surface models are straightforward once a model of attention is adopted (see, for example, [Olshausen o*t* a/., 1993]). Models of controlling attention shifts between surfaces in these distributed spatial representations must be designed.

Seeking enhanced base representations improves results greatly. For example, some degree of normalization

[Heeger, 1992] in the base representation may reduce the dependence on local contrast, although the convenience of steerability is lost. Some operations, such as occlusion detection [Heitger, 1992; Niyogi, 1995a], may be based on the distributed representations and/or the surfaces fit to them; using these representations to constrain model-fitting processes is desired. Linking our model-fitting processes to solutions to surface segmentation mechanisms using distributed representations [Heitger and von der Heydt, 1993; Finkel and Sajda, 1992] is clearly desirable.

The majority of computer vision efforts in early vision have been directed towards estimating explicit representations of local structure, and subsequently fitting models to these representations. These explicit representations are prone to failure where model assumptions break down. Implicit representations do not have these problems. Problems in modeling local structure, object model recovery, handling occlusion, surface segregation, attention, etc. can and should be attempted using distributed representations of vision. A parallel effort using distributed representations of vision is likely to be more fruitful than relentlessly improving methods of recovering explicit local structure. We have presented a first step at fitting models to distributed representations of vision here.

References

[Adelson and Bergen, 1985] E. II. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *JOSA A,* 2:284 299, 1985.

[Adelson and Bergen, 1991] E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing*. MIT Press, 1991.

[Bergen and Landy, 1991] J.R. Bergen and M. Landy. Computational models of visual texture segregation. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing.* MIT Press, 1991.

[Bergen *et al,* 1992] J. R. Bergen, P. J. Burt, K. Hanna, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE PAMI,* 14:886-896, 1992.

[Black and Anandan, 1993] M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *ICCV,* pages 231-236, Berlin, Germany, 1993.

[Finkel and Sajda, 1992] L. H. Finkel and P. Sajda. Object discrimination based on depth-from-occlusion. *Neural Computation,* 4:901-921, 1992.

[Fleet and Jepson, 1990] D. Fleet and A. Jepson. Measurement of image velocity by phase information. *International Journal of Computer Vision,* 5(1):77, 1990.

[Freeman and Adelson, 1991] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE PAMI,* 13(9):891-906, 1991.

[Grzywacz and Yuille, 1990] N. M. Grzywacz and A. L. Yuille. A model for the estimate of local image velocity by cells in the visual cortex. *Proceedings Royal Society London B,* 239:129-61, 1990.

[Heeger, 1987] D. Heeger. Model for the extraction of image flow. *JOSA A,* 4:1455-1471, 1987.

[Heeger, 1992] D. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience,* 9:181-197, 1992.

[Heitger and von der Heydt, 1993] F. Heitger and R. von der Heydt. A computational model of neural contour processing: Figure-ground segregation and illusory contours. In *ICCV,* pages 32-40, Berlin, Germany, 1993.

[Heitger, 1992] F. et al. Heitger. Simulation of neural contour mechanisms: From simple to end-stopped cells. *Vision Research,* 32(5):963 981, 1992.

[Kass and Witkin, 1987] M. Kass and A. Witkin. Analyzing oriented patterns. *Computer Vision, Graphics and Image Processing,* 37:362 385, 1987.

[Kass *et al,* 1987] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision,* 1(1):1, 1987.

[Koenderink, 1992] J. Koenderink. Generic neighborhood operators. *IEEE PAMI,* 14(6):597- 605, 1992.

[Niyogi, 1995a] S. Niyogi. Detecting kinetic occlusion. In *International Conference on Computer Vision,* 1995.

[Niyogi, 1995b] S. Niyogi. Kinetic occlusion. Technical Report 319, MIT Media Lab, 1995.

[Olshausen *et al,* 1993] B. Olsh ausen, C. Anderson, and D. C. van Essen. A neurobiological model of visual attention and invariant pattern recognition based of dynamic routing of information. *Journal of Neuroscience,* 13(11):4700, 1993.

[Perona, 1992] P. Perona. Steerable-scalable kernels for edge detection and junction analysis. In *Proceedings of the European Conference on Computer Vision,* pages 3-18, 1992.

[Poggio *et al,* 1985] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature,* 317(26):314-319, 1985.

[Qian, 1994] N. Qian. Computing stereo disparity and motion with known binocular cell properties. *Neural Computation,* 6(3):390, 1994.

[Sanger, 1988] T. D. Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics,* 59:405-418, 1988.

[Shizawa and Mase, 1991] M. Shizawa and K. Mase. A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis. In *IEEE CVPR*, 1991.

[Simoncelli, 1993] E. Simoncelli. *Distributed Representations of Motion.* PhD thesis, MIT EECS, January 1993.

[Terzopoulos, 1988] D. Terzopoulos. The computation of visible surface representations. *IEEE Pattern Analysis and Machine Intelligence,* 10(4):417-438, 1988.