M.I.T. Media Laboratory Vision and Modeling Group Technical Report No. 244, October 1993

# Separation of Transparent Motion into Layers using Velocity-Tuned Mechanisms

**Trevor Darrell and Eero Simoncelli†**

Vision and Modeling Group
The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street
Cambridge MA, 02139

†Grasp Laboratory
University of Pennsylvania
Philadelphia PA 19104

**ABSTRACT**

This paper presents a model for the perception of transparently combined moving images. We advocate a framework consisting of a local motion mechanism which can operate in the presence of transparency, and a global mechanism that integrates information across space. We present a new method for the local motion testing mechanism, using "donut" velocity selective mechanisms formed from the weighted combination of spatio-temporal energy units. This method has the advantage over traditional methods that it does not fail when there are multiple motions in the sequence. The global layer selection mechanism attempts to account for the local velocity distributions with a small set of global functions. Using donut mechanisms permits a simplified layer selection optimization, in which inhibition between layers is determined by the product of their predicted velocity distributions. With this scheme, we demonstrate the decomposition of image sequences containing additively combined multiple moving objects into a set of layers corresponding to each object.

## 1 Introduction

The human visual system can easily distinguish multiple motions that are transparently combined in an image sequence. However, traditional computational models of the motion perception process have largely been confused by scenes with multiple motions. In this paper we wish to provide a model for the perception of displays in which moving

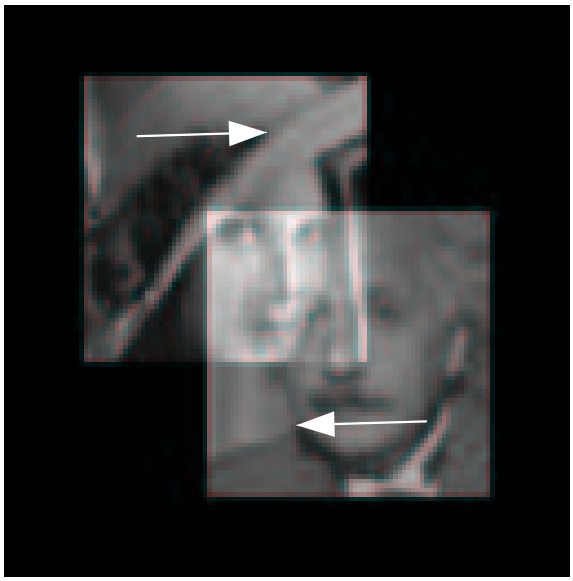

Figure 1: Frame from an image sequence with two transparently combined moving subimages. Each subimage is moving in a different direction, as indicated by the superimposed arrows.

images overlap and are transparently combined.

A synthetic example of a transparent motion display can be constructed by spatially shifting and adding together several images, as shown in Figure 1. Real world examples of these transparent viewing conditions are common; for example people looking out a car window often see both the outside world and a clear reflection of objects inside the car. As any driver under inclement weather conditions has observed, the human visual system can perform accurate navigation functions even when a large percentage of the image signal is obscured by a corrupting, but coherently organized, noise process such as rain, sleet or snow.

Transparent segregation can be performed on a moving stimuli when a static display of a single image from the stimuli does not support such segregation. Mulligan [14] has shown that human observers could easily segregate two coherently moving patterns of spatial noise that were unseparable in static presentation. With both the structured and unstructured (spatial noise) stimuli, there is usually the perception of a small number of surfaces corresponding to what was used in the actual physical or synthetic construction of the stimulus. We desire a model which provides for a similar qualitative result, grouping information into a small number of chunks based only on the motion information present in the visual signal.

In general, a model for the perception of transparent motion must address two questions: 1) what local motion measurements are made? and 2) how are these local measurements used to group coherently moving regions of the scene? Several current computational approaches provide interesting insights into these issues. The algorithm of Shizawa and Mase [17, 18] directly computes two velocity vectors for each location in the image, but does not address the problem of perceptual grouping of coherently moving regions of

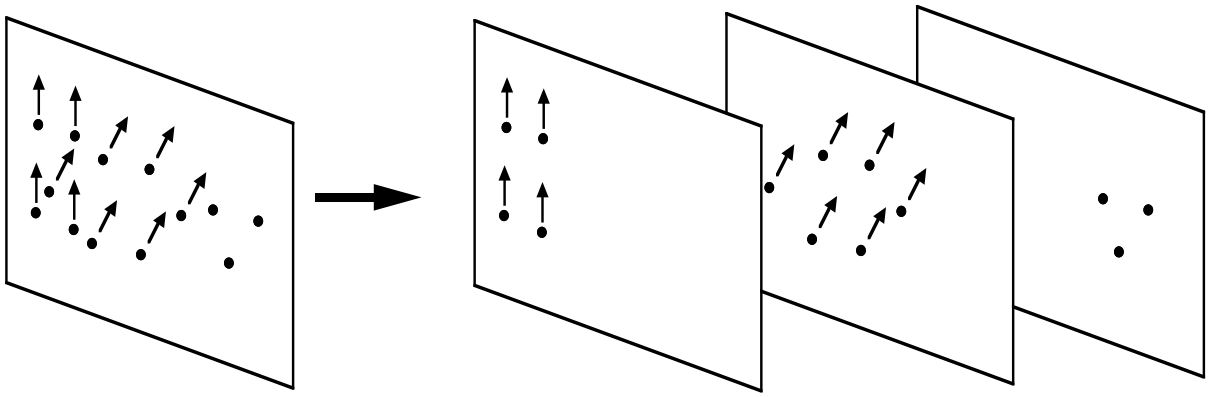

Figure 2: Separation of random-dot image motion into layers. Each group of coherently moving points are grouped together on the same layer.

the scene. The algorithms of Bergen et. al. [3] and Irani and Peleg [12] compute global affine optical flow fields, but use local measurements that are only capable of determining a single velocity estimate at each point.

We have adopted a model that combines the best of both approaches, using a global constraint mechanism with arbitrary spatial grouping, and local measurements that are robust in the presence of transparent motion. Our grouping mechanism is based on the idea of a "layered representation" proposed by Adelson [2, 1], in which each object or process in a scene is represented by a data structure describing aspects of the image attributable to that object. In general, layers often mimic the physical structure of the scene, and can have an ordinal or metric depth ordering, or just be "distinct" surfaces or objects, as in the case of intertwined sheets. Our version of layers adopts the latter notion, and does not enforce any depth ordering on the objects. For example, in the case of a random dot stimulus illustrated in Figure 2, our framework would group the coherently moving dots into separate layers, irrespective of whatever depth ordering was present.

# 2 Hypothesize and Test Approach to Layer Recovery

Recovering a layer-based image representation requires estimating the number of objects or surfaces in the scene, their support (projection) in the image plane, and parameters which describe the their observed properties. Several approaches have been introduced to solve this task. One method, presented by Irani [12], is to assume a spatially dominant "background" whose parameters can be estimated based on the entire image data since the outlier contamination from the foreground will be relatively small. The background estimate can be further refined using an iterative robust estimation technique, re-estimating the parameters based only on the points with low residual error. The

motion of a foreground object can be computed by computing an estimate based only on the complement of the background support. With multiple objects, these recursive estimate-and-segment approaches fail when objects exist at the same scale. In this case the percentage of outliers in the signal relative to the estimation of either object will far exceed the breakdown point of the robust estimation method [13].)

Wang and Adelson [21] have developed a layer recovery algorithm that fits motion parameters to optical flow estimates over small initial regions of the image, and then iteratively merges regions using a $k$-means parameter clustering algorithm. This method will not have difficulty when there are multiple objects of the same relative size, since for each object there will likely be a small initial subregion with few, if any, outliers due to occlusion by other objects. However this method will have difficulty when stable and accurate estimates of the true motion parameters can not be computed from small initial regions. This is true for certain global motion conditions, such as rotation about the vertical camera axis, and for objects that are completely transparent within a scene. Additionally, the reliance on an optical flow stage of computation makes it difficult to apply their framework to transparent stimuli.

We adopt a hypothesize-and-test approach in which there are no specific restrictions on the source of initial hypotheses, which does not require the computation of optical flow, and is directly applicable to the case of transparency. There are three aspects to a hypothesis-and-test method that need to be specified: 1) a method of generating the initial set of hypothetical motion fields, 2) a mechanism for testing the validity of each hypothesis as a description of the data, and 3) a selection mechanism that chooses the subset of hypotheses which are the best overall description of the data.

Our method assumes an *a priori* model of global structure. That model can be strictly parametric, for example an affine model of optical flow, or only implicitly parametric, such as a smoothness assumption on flow, or a rigid body motion constraint. Irrespective of the model used, hypotheses are generated either by sampling the parameter space, or by fitting initial guesses to samples of the data (e.g. small initial windows, as used in [7, 21]). Since the latter approach is impractical with transparent displays, in the work presented in this paper we initialize hypotheses by sampling the parameter space. We assume that a sufficiently dense sampling of the parameter space is available so that we can be assured that at least one hypothesis will match the actual flow fields in the scene. While this assumption will restrict the applicability of our algorithm, we have found that there are many application domains for which it is reasonable.

In the next section, we will develop in detail the mechanism for computing the support or evidence a particular velocity hypothesis has in a particular image sequence. Following that we will review the selection mechanism that selects an set of candidate layers based on finding a minimal length encoding of the scene, and present the extension to handle the case of transparency.

# 3 Local Velocity Testing

A hypothesized flow field will give rise to local predictions of velocity at every point in the image, which can be tested. The problem of analyzing the local velocity information present in an image signal is an area of active research in computer vision and visual perception. If, in fact, a local patch of the scene contains only a single motion and no illumination effects, then the well-known gradient-based approach to velocity computation is an appropriate and sufficient model.

## 3.1 The gradient model of velocity computation

In the gradient-based approach, first-order local image derivatives are used to form a constraint equation. (See [15] for a summary of the gradient-based approach.) According to the gradient constraint, the spatial and temporal image derivatives at a point obey the following equation:

$$v_x(x,y)\frac{\partial I(x,y,t)}{\partial x} \; + \; v_y(x,y)\frac{\partial I(x,y,t)}{\partial y} \; = \; -\frac{\partial I(x,y,t)}{\partial t}. \tag{1}$$

where $I(x,y,t)$ is an image sequence and $\vec{v}(x,y)$ is a velocity field. Using an operator notation,

$$D(\vec{v}) = \begin{bmatrix} v_x \\ v_y \\ 1 \end{bmatrix} \cdot \nabla = (v_x\frac{\partial}{\partial x} \; + \; v_y\frac{\partial}{\partial y} \; + \; \frac{\partial}{\partial t}), \tag{2}$$

we can express eq. (1) as

$$D(\vec{v}(x,y))I(x,y,t) \; = \; 0. \tag{3}$$

Taken in the Fourier domain, the gradient constraint implies that all energy associated with a velocity $\vec{v}$ lies on a plane in space-time [22]. From eqs. (2,3) we can see $D(\vec{v})$ is a linear operator, and that it must have zero response for spatio-temporal frequencies on the constraint plane. We can thus call $D(\vec{v})$ a "nulling" filter, as it removes any energy along the constraint plane from the image sequence. Figure 3 shows a plot of the spectral response of this filter for a 1D velocity; in this case, all energy is constrained to lie along a line in space-time.

## 3.2 Computing velocity support with the gradient model

Our goal in this section, as stated above, is to see if a velocity is present in a local region of an image sequence. We define the support for a velocity at a point as the conditional probability that a velocity is present given the observed image derivatives. As shown by Simoncelli [19], we can introduce a noise model into the gradient constraint equation (in the 1D case):

$$D(\vec{v}(x,y) + n_1(x,y))I(x,y,t) \; = \; n_2(x,y) \quad n_i = N(0,\sigma_i), \tag{4}$$

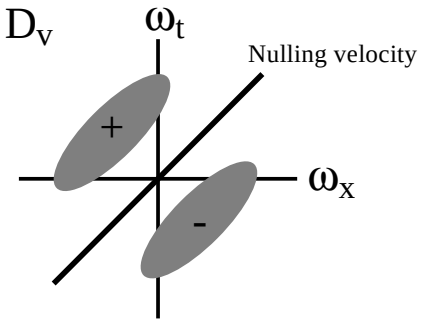


Figure 3: Idealized depiction of the spectrum of a directional derivative operator. The operator "nulls" energy along the line specified by the constraint velocity $v$, and passes any other energy. Thus the operator will have zero output, $D(v)I = 0$, if all of the energy lies on the constraint line.

where the noise term $n_1$ describes errors in the planarity assumption, and $n_2$ accounts for any remaining errors in the temporal derivative measurement.

Assuming there is only a single motion in the neighborhood, the conditional probability of a velocity $v$ given the image gradient $\nabla I$ is given by

$$-\log P(\vec{v}|\nabla I) \propto \frac{||D(\vec{v})I||^2}{\sigma_1||\nabla I||^2 + \sigma_2} + \frac{||v||^2}{\sigma_v}, \tag{5}$$

where we assume a prior distribution of velocities of the form $P(\vec{v}) \propto N(0, \sigma_v)$. In this case our support function is simply the conditional probability:

$$s(x, y, \vec{v}) \;=\; P(\vec{v} \mid \nabla I(x, y, t)). \tag{6}$$

## 3.3 Gradient model fails on Multiple Motions

As described above, the gradient model assumes there is only a single motion within the spatial extent of the derivative filters used in the computation. Thus if there there are multiple motions within a local region, as there are in transparent displays, support computation using the gradient model will fail.

For example, the Figure 4 shows the support maps computed using the gradient constraint for the two global translations present in the moving face stimuli described in Figure 1. Regions of the stimuli which have only a single motion are well supported, but regions which have multiple motions are not supported by either hypothesis (nor any other single motion hypothesis).

This can be explained by examining the behavior of this mechanism in the spatio-temporal frequency domain. The additive model of multiple motions tells us that if two patches with different motions are added, they will yield different planes in spatio-temporal energy space. More formally, the principle of superposition states that if the

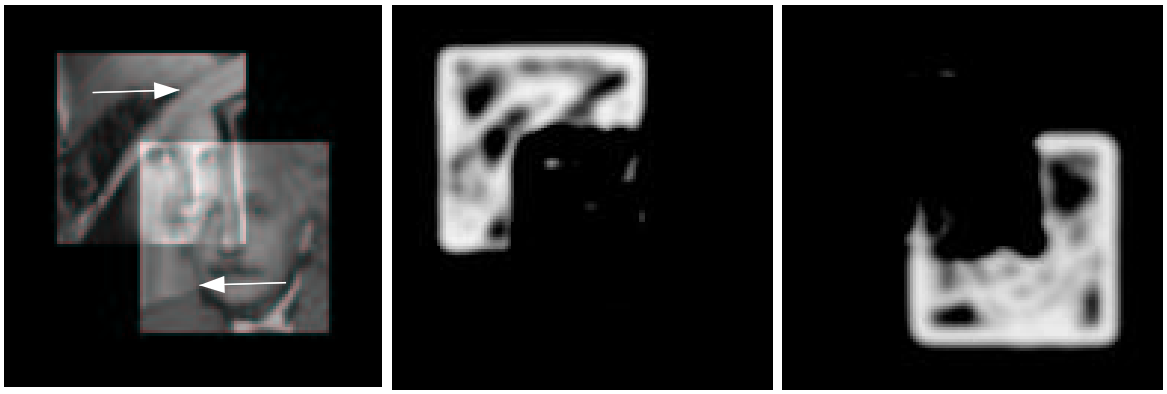

Figure 4: Support maps for actual motions in image sequence presented in Figure 1. Support was computed using the single-motion gradient constraint; regions with transparency (multiple motions) fail to be supported.

image sequences are added, the corresponding constraints are multiplied. A gradient-based mechanism typically squares energy off the plane, and so treats a second motion as extreme noise. It will not yield zero response if a second motion is present.

### 3.4 Additive models of motion transparency

When there is motion transparency in a scene, more than one motion may be present at each point in the image. To compute the support of a velocity field in the presence of transparency, we need to rely on local measurements that are robust in the presence of multiple motions at a given point. To derive these measurements, we need a local model of multiple (transparent) motion.

For image sequences that have been additively combined, we can simply chain the gradient constraint operators of the constituent velocities:

$$D(\vec{v}(x,y))D(\vec{v'}(x,y))I(x,y,t) \;=\; 0 \tag{7}$$

Constructing this dual-motion constraint operator, we have

$$D_2(\vec{v},\vec{v'}) = D(\vec{v})D(\vec{v'}) = (v_x\frac{\partial}{\partial x} \;+\; v_y\frac{\partial}{\partial y} \;+\; \frac{\partial}{\partial t})(v'_x\frac{\partial}{\partial x} \;+\; v'_y\frac{\partial}{\partial y} \;+\; \frac{\partial}{\partial t}) \tag{8}$$

$$= (v_xv'_x)\frac{\partial^2}{\partial x^2} \;+\; (v_yv'_y)\frac{\partial^2}{\partial y^2} \;+\; (v_xv'_y + v_yv'_x)\frac{\partial^2}{\partial xy} +$$

$$(v_x + v'_x)\frac{\partial^2}{\partial xt} \;+\; (v_y + v'_y)\frac{\partial^2}{\partial yt} \;+\; \frac{\partial^2}{\partial t^2}$$

and in this notation the dual motion constraint is simply

$$D_2(\vec{v}(x,y),\vec{v'}(x,y))I(x,y,t) \;=\; 0 \tag{9}$$

In the spatio-temporal domain, this corresponds to the dual constraint that all of the energy in the signal lie on *either* of the planes specified by the two velocities. Thus

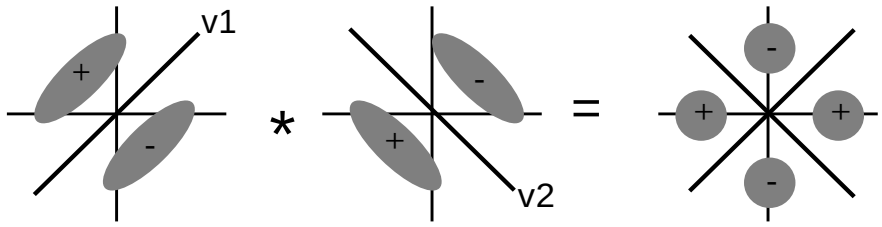


Figure 5: Idealized spectrum of a second-order derivative operator, constructed by composing two first-order directional derivatives. The operator "nulls" energy along the two lines specified by the constraint velocities $\vec{v}$ and $\vec{v'}$. Thus the operator will have zero output, $D_2(\vec{v}, \vec{v'}) = 0$, if all of the image energy lies on the constrained lines.

this operator "nulls" two constraint planes (or lines, in the 1D case). Figure 5 shows a spectral plot of $D_2$ for two velocities.

Shizawa and Mase [17] introduced this formulation and proceeded to develop methods for the analytic estimation of two velocities at each point given the local image derivative information. Our focus, however, is on the computation of support, i.e. the likelihood of a velocity given the image information.

If one of the motions in a local patch were known then we could use the constraint operator as a "nulling prefilter" to remove the energy corresponding to that motion. However, we typically don't know any of the motions in a sequence *a priori*. One approach, explored in [6, 9], is to enumerate a bank of prefilters using a set of sampled velocities $\{\vec{u_0}, \vec{u_1}, ..., \vec{u_n}\}$ and test a set of constraints for each $v$, $\{D_2(\vec{v}, \vec{u_i}) \mid 0 \leq i \leq n\}$, to see if any are zero. This approach is insensitive to a single other motion that may be present in the scene, but has the disadvantage that it is difficult to sample the velocity space finely enough to "null" all possible motions.

Instead of using a "nulling filter" paradigm, computing an operator which is zero when the motion is present, one can use a dual notion, and compute an operator which is maximal when the motion is present. This "maximizing filter" paradigm is more appropriate when computing velocity information with an unknown number of motions. We develop this approach further in the following section.

## 3.5 Velocity selective approach to support testing

To compute velocity support in the presence of transparency, we use the output of narrowly tuned, velocity selective motion mechanism to test for the presence of a particular velocity. Simoncelli has developed a model of a velocity selective mechanism which is in turn constructed out of a set of spatio-temporal-frequency energy mechanisms (STEMs) [20]. The Fourier response of this "donut" mechanism is aligned along a plane which intersects the origin, with preferential response along a fixed radius from the origin. The mechanism is interpolated from a set of STEMs and can be computed efficiently and in a biologically plausible manner.

The donut mechanism is defined through a generalized error function based on a set

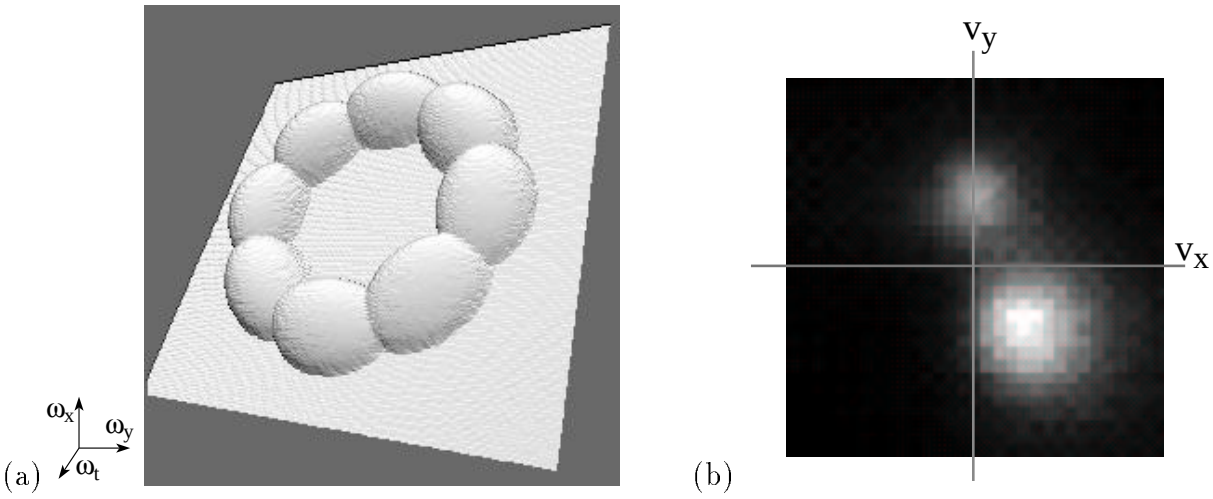


Figure 6: (a) Construction of "donut" in Fourier domain (b) Population response of many donuts tuned to different velocities on image sequence with two motions.

of $N$-th order directional derivatives. The error function is given by

$$\mathcal{P}_N(\vec{v}) \;=\; \sum_{\vec{\omega}} \sum_{i=0}^{N} [\hat{u}_i(\vec{v}) \cdot \hat{\omega}]^{2N} |F(\vec{\omega})|^2 \;. \tag{10}$$

where the derivative measurements are given by

$$\hat{u}(\vec{v}) \;=\; \cos\left(\frac{2\pi i}{N+1}\right) \hat{u_a}(\vec{v}) \;+\; \sin\left(\frac{2\pi i}{N+1}\right) \hat{u_b}(\vec{v}) \quad 0 \;\leq i \;\leq\; N \;. \tag{11}$$

with

$$\hat{u_a}(\vec{v}) \;=\; \hat{u}(\vec{v}) \times \hat{e}_x$$
$$\hat{u_b}(\vec{v}) \;=\; \hat{u}(\vec{v}) \times \hat{u_a}(\vec{v})$$

where $\hat{e}_x$ is a unit vector in the direction of the $x$-axis and $\hat{u}(\vec{v}) = (v_x, v_y, 1)^T / \sqrt{|v|^2 + 1}$. The construction of this donut mechanism is shown in Figure 6. For full details of the mechanism and its derivation see [20]. In the examples presented here, we have used donut mechanisms based on third derivative measurements.

This donut operator is selective in the Fourier domain for a particular velocity, and will not be confused by any other transparently combined velocities. We thus can compute the support for a particular velocity by computing the normalized donut response for the donut corresponding to the velocity to be tested.

$$s(x, y, \vec{v}) \;=\; \frac{\mathcal{P}_N(x, y, \vec{v})}{\sum_{x,y,\vec{v}} \mathcal{P}_N(x, y, \vec{v})} \tag{12}$$

# 4 Hypothesis/Layer Selection

As described above, our method follows a hypothesize-and-test paradigm, generating a set of hypotheses which predict global velocity field structure, then testing which velocities

are actually supported in the data. In this section we present the mechanism that selects the subset of hypotheses which best accounts for the overall visual signal.

Previously, Pentland [16] proposed a mechanism to select a subset of hypotheses that was the minimum length description of an image, for the case of a boolean image and boolean shape primitives. As elaborated further in [5, 8] the selection criteria was derived from a minimal-length description criteria, which sought to find the shortest encoding of the signal. Using a simple support function based on the gradient constraint, we first applied this approach to the motion domain in [7].

The grouping mechanism adopted in [7] assumed a purely spatial notion of support and could not explicitly represent transparent phenomena, since the support was constrained to be spatially non-overlapping in the final solution. This paper extends the definition of support from an exclusively spatial notion, to include the velocity domain. The key insight is that when processing transparent motion displays, the support of a motion hypotheses should exist over both a region of space *and* velocity, so that more than one motion can be distinctly supported at a single spatial location.

Given a set of velocity fields $\mathcal{H}$:

$$\mathcal{H} = \{\vec{h_0}(x,y), \vec{h_1}(x,y), ..., \vec{h_M}(x,y)\} \tag{13}$$

we check to see if the local velocities have support in an image sequence. Each local hypothesis $\vec{h_i}(x,y)$ predicts a certain distribution over velocity space. We define $\mathcal{W}_{\vec{u}}(\vec{v})$ to be the expected support distribution over velocity space given a velocity $u$ and a uniform spatio-temporal energy distribution. For the third-order donut filters we use, the expected support given a known motion can be shown to be

$$\begin{aligned} \mathcal{W}_u(\vec{v}) &= E(s(\vec{v}) \mid \vec{u}) \\ &= \frac{1}{4} + \frac{1}{4}(\vec{u}\cdot\vec{v})^6 + \frac{1}{2}(1-(\vec{z}\cdot\vec{v})^2)^3 \end{aligned} \tag{14}$$

where $c = \vec{v} \times \vec{u}$ and $z = \vec{c} \times \vec{u} + c$. Figure 7 plots $W$ for an array of different velocities. As can be seen, $W$ is generally a unimodal, symmetric, "blob"-like function. The exact shape of $W$ is relatively unimportant; we could have used Gaussians or similarly shaped functions and achieved the same qualitative results reported below.

Our selection method proceeds by attempting to find the subset $\mathcal{L} \subset \mathcal{H}$ that best matches the velocity energy in the signal using the smallest number of hypotheses possible. We express this as an criterion function which encourages a match between the predicted velocity distribution and the data, but discounts the contribution of regions where there are conflicting (overlapping) hypotheses.

We attempt to find the subset $\mathcal{L}$ which maximizes

$$E(\mathcal{L}) = \sum_{\{h|h\in\mathcal{L}\}} \sum_{x,y,\vec{v}} s(x,y,\vec{v}) \mathcal{W}_{h(x,y)}(\vec{v}) \left(1 - \sum_{\{k|k\neq h, k\in\mathcal{L}\}} \mathcal{W}_{k(x,y)}(\vec{v})\right) - \alpha||\mathcal{L}|| \tag{15}$$

where $\alpha$ is a penalty that is incurred when an an additional layer is added to the representation. This equation encourages the selection of the set of layers whose correlation with

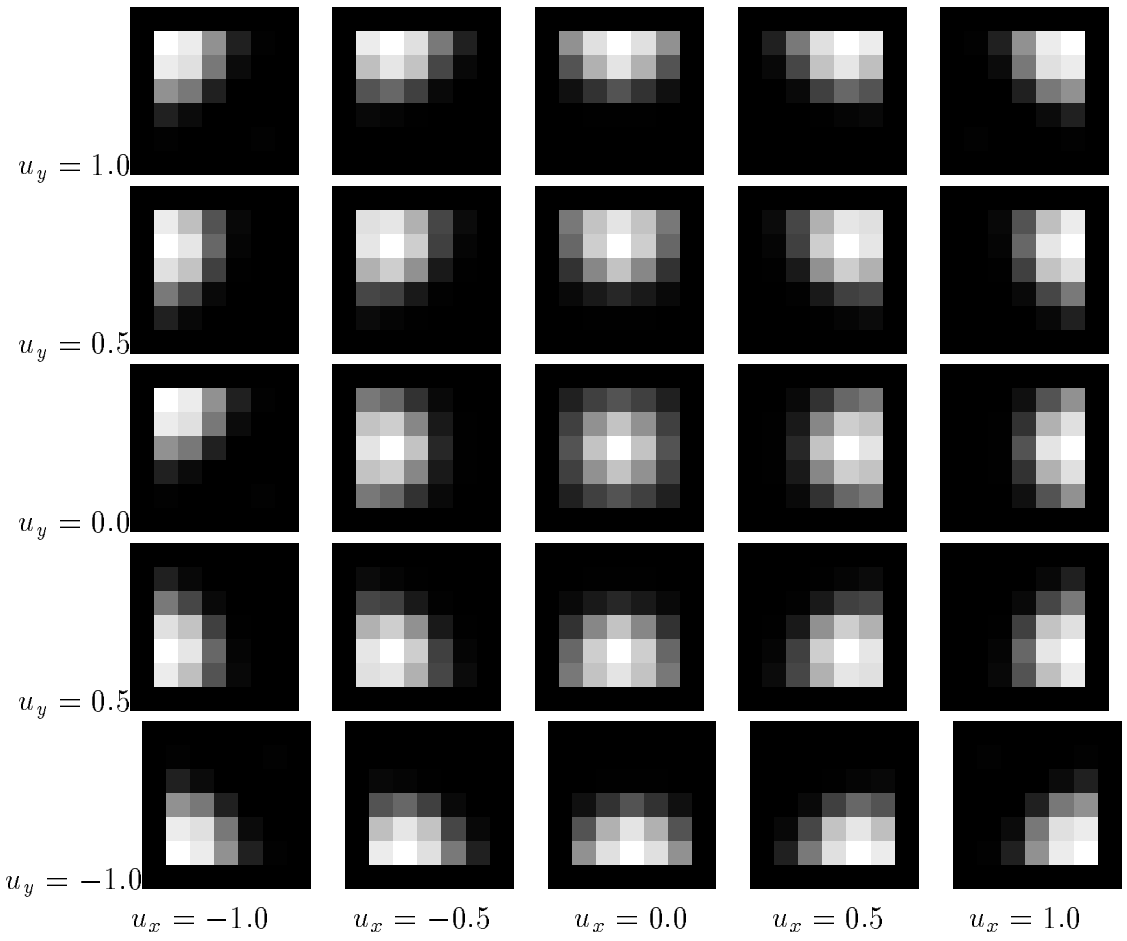


Figure 7: Plots of $W_{\vec{u}}(\vec{v})$. Each image shows $W$ for a particular $\vec{u}$, evaluated over the range of $-1.0 \leq v_x, v_y \leq 1.0$.

the observed velocity distribution is maximal, subject to a strong inhibitory term that prevents more than one layer from covering the same data, and a term that discourages layers with too little support (e.g. less than $\alpha$).

Since it has only first and second-order terms, this criterion function can be optimized using a number of standard techniques. One method that is biologically plausible is a distributed network optimization, based on a Hopfield-Tank network [10]. In this formulation, each hypothesis is associated with a particular network unit, whose activity represents (relative to a threshold) whether the hypothesis is included in the representation. Each unit produces a response that is the sum of its inputs passed through a sigmoid non-linearity. Units receive excitatory input proportional to the support they have in the data, and inhibit other hypotheses with similar support.

To maximize this numerically, we need to express $E(\mathcal{L})$ in a differentiable form. First, we represent $\mathcal{L}$ by enumerating those elements of $\mathcal{H}$ that are in $\mathcal{L}$ using a vector $a$. The sign of each $a_i$ indicates whether a layer hypothesis $l_i \in \mathcal{H}$ is an element of the layer subset $\mathcal{L} \subset \mathcal{H}$: a positive value means $l_i \in \mathcal{L}$, and a negative value means $l_i \notin \mathcal{L}$.

We thus rewrite Eq. 15 as

$$E(\vec{a}) \;=\; \sum_i \sigma(a_i) \sum_{x,y,\vec{v}} s(x,y,\vec{v}) \mathcal{W}_{h_i(x,y)}(\vec{v}) \left(1 - \sum_{j \neq i} \sigma(a_j) \mathcal{W}_{h_j(x,y)}(\vec{v})\right) - \alpha \sum_i \sigma(a_i) \quad (16)$$

where $\sigma(a_i)$ is function which transforms the $a_i$ values into a multiplicative weight between 0 and 1. Ideally, $\sigma()$ would be the unit step function centered at the origin. However, this hard non-linearity makes it very difficult to maximize $E(\vec{a})$ numerically. Instead, we use the "softer" sigmoid function,

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (17)$$

The Hopfield-Tank network performs gradient ascent on $a_i$ through the following update rule:

$$\frac{da_i}{dt} = \frac{1}{K} \frac{dE(\vec{a})}{d\sigma(a_i)} =$$

$$= \frac{1}{K} \sum_{x,y,\vec{v}} s(x,y,\vec{v}) \mathcal{W}_{h_i(x,y)}(\vec{v}) \left(1 - \sum_{j \neq i} \mathcal{W}_{h_j(x,y)}(\vec{v}) \sigma(a_j)\right) - \alpha\sigma(a_i) \quad (18)$$

where $K$ is an integration constant. This can be expressed in matrix terms as

$$\vec{\nabla}\vec{a} \;=\; \vec{U} - \mathbf{T}\vec{A} \quad (19)$$

where $A_i = \sigma(a_i)$ and

$$U_i \;=\; \sum_{x,y,\vec{v}} s(x,y,\vec{v}) \mathcal{W}_{h_i(x,y)}(\vec{v})$$

and

$$T_{ij} \;=\; \sum_{x,y,\vec{v}} s(x,y,\vec{v}) \mathcal{W}_{h_i(x,y)}(\vec{v}) \mathcal{W}_{h_j(x,y)}(\vec{v}) \qquad i \neq j$$

except on the diagonal, where $T_{ii} = 0$. Since $\mathbf{T}$ is symmetric, the network will converge to a stable solution which is a local maxima of $E(\vec{a})$ [10].

### 4.1 Selection Example

First we show the results of our mechanism on an image sequence containing a single motion. The formulation presented above does not place any restrictions on the form of hypotheses; any vector field can be used. However, for clarity of presentation, we restrict ourselves to the case of planar translations for these example. Our array of hypotheses is the same as used in Figure 7, $\{-1 \leq v_x, v_y \leq 1\}$.

Figure 8(a) shows five frames from an image sequence containing a translating random noise pattern. Part (b) of this figure shows $\sum_{x,y} s(x, y, \vec{v})$; a single uni-modal peak is evident. In this case, our selection mechanism should select a single hypothesis from those in Figure 7 to account for the observed distribution. Figure 8(c) shows $\vec{U}$ and $\mathbf{T}$ for this example; $\vec{U}$ is displayed as an image, each of whose points is the product of the image in part (b) times each of the images in Figure 7. Each row and column of $\mathbf{T}$ should also be thought of as an image; each entry $T_{i,j}$ is computed by taking the product of the image in part (b) with two images ($i$, $j$) from Figure 7. This vector and matrix define a network as described above; one can see this network has both strong excitatory input and mutual cross-inhibition among the hypotheses centered over the peak of the distribution of $s$. Indeed, in this case the Hopfield-Tank network is, in essence, a winner-take-all network, and a single hypothesis survives. Part (d) shows the predicted support for the selected hypothesis; one can see it matches well the actual distribution in part (b). The output of the donut mechanism tuned to the predicted velocity is shown in part (e); there is a large energy response over the entire field, as the hypothesized motion is present throughout the scene.

Next we show the results of our mechanism on a stimulus with two motions. Figure 9(a) shows five frames from an image sequence containing two translating random noise patterns. Part (b) of this figure shows $\sum_{x,y} s(x, y, \vec{v})$; two peaks are evident in the distribution. Thus, our selection mechanism should select two hypotheses from those in Figure 7 to account for the observed distribution. Figure 8(c) shows $\vec{U}$ and $\mathbf{T}$ for this example; there are two groups of hypotheses with strong excitatory input and mutual cross-inhibition in this network. There are two disjoint winner-take-all networks that are (approximately) formed in this case, and two hypotheses are selected. Part (d) shows the predicted support for the selected hypotheses; one can see their sum matches well the actual distribution in part (b). The output of the donut mechanisms tuned to the predicted velocities is shown in part (e); there is a large energy response over the entire field in both, as both hypothesized motions are present throughout the scene. Finally, we show the results using the image shown in Figure 1. As Figure 10 shows, the results are the same as the previous figure, with the exception that the donut mechanism output shown in part (d) only covers the regions of the scene where the motion is actually present.

## 5 Conclusion

We have presented a model for recovering a layered representation of transparent motion stimuli. Our method uses a new approach to the separation of transparent motions, using

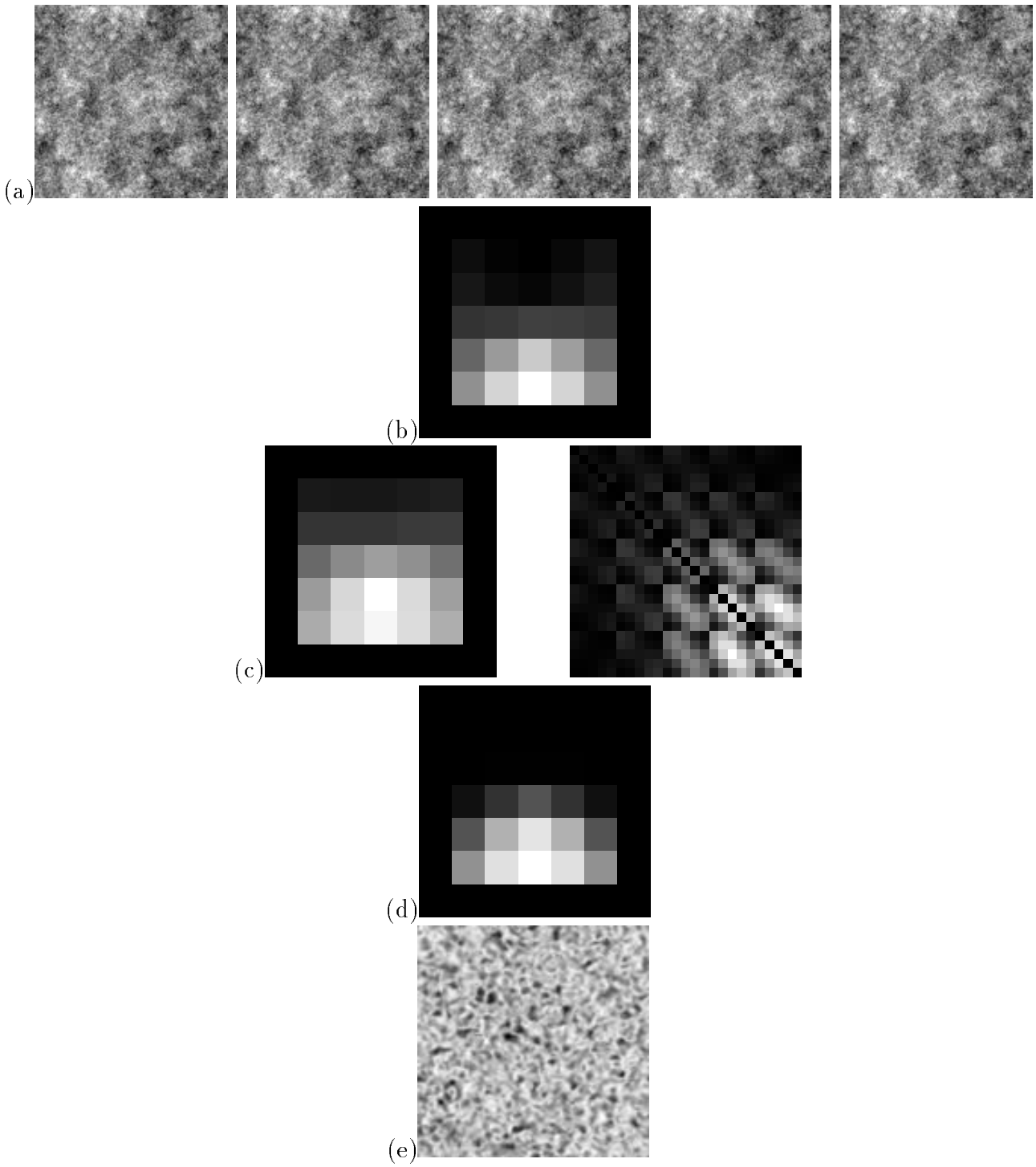


Figure 8: Selection of single layer when single motion is present in random noise stimulus; see text for full description of figure.

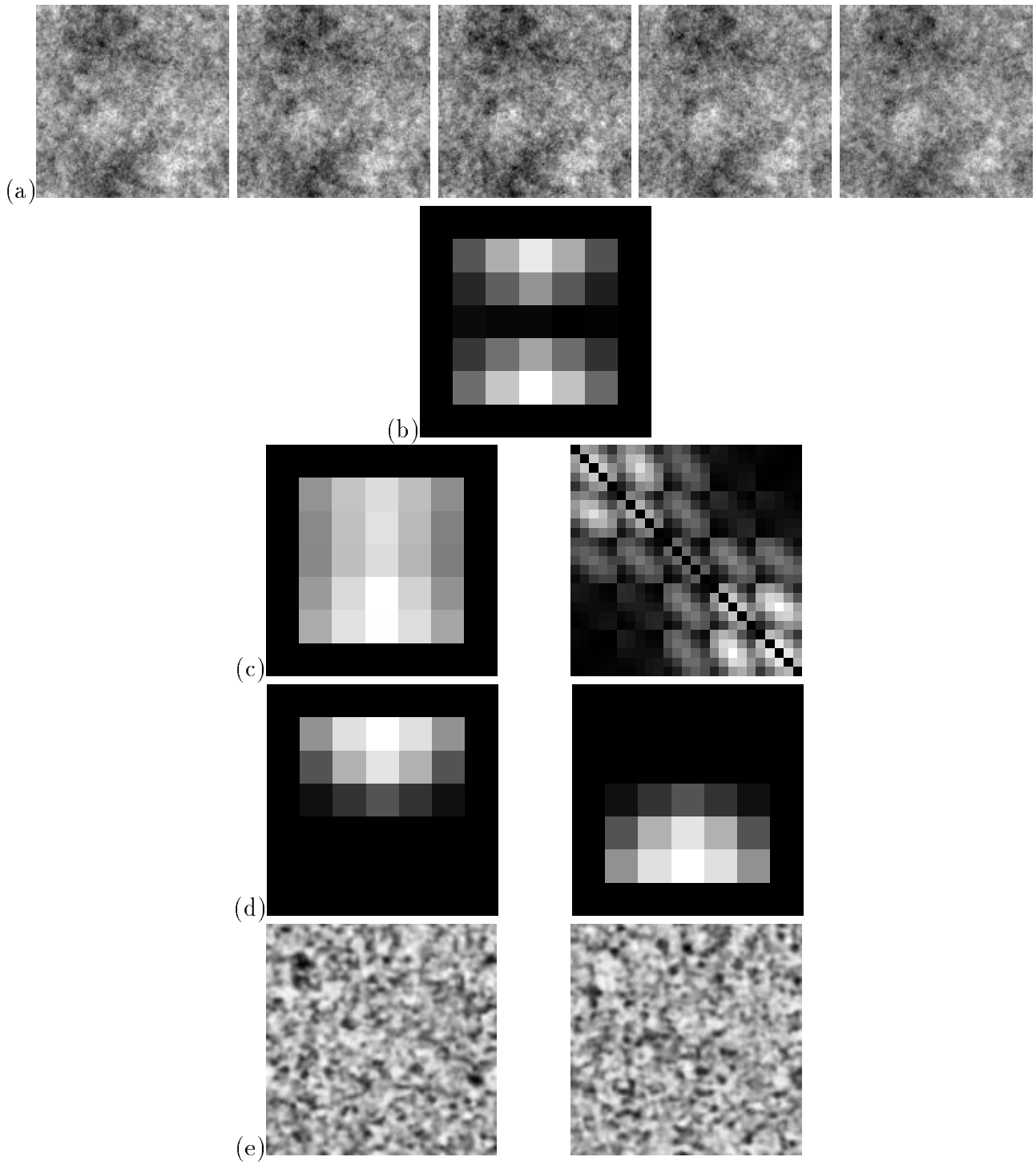


Figure 9: Selection of two layers when two motions are present in random noise stimulus; see text for full description of figure.

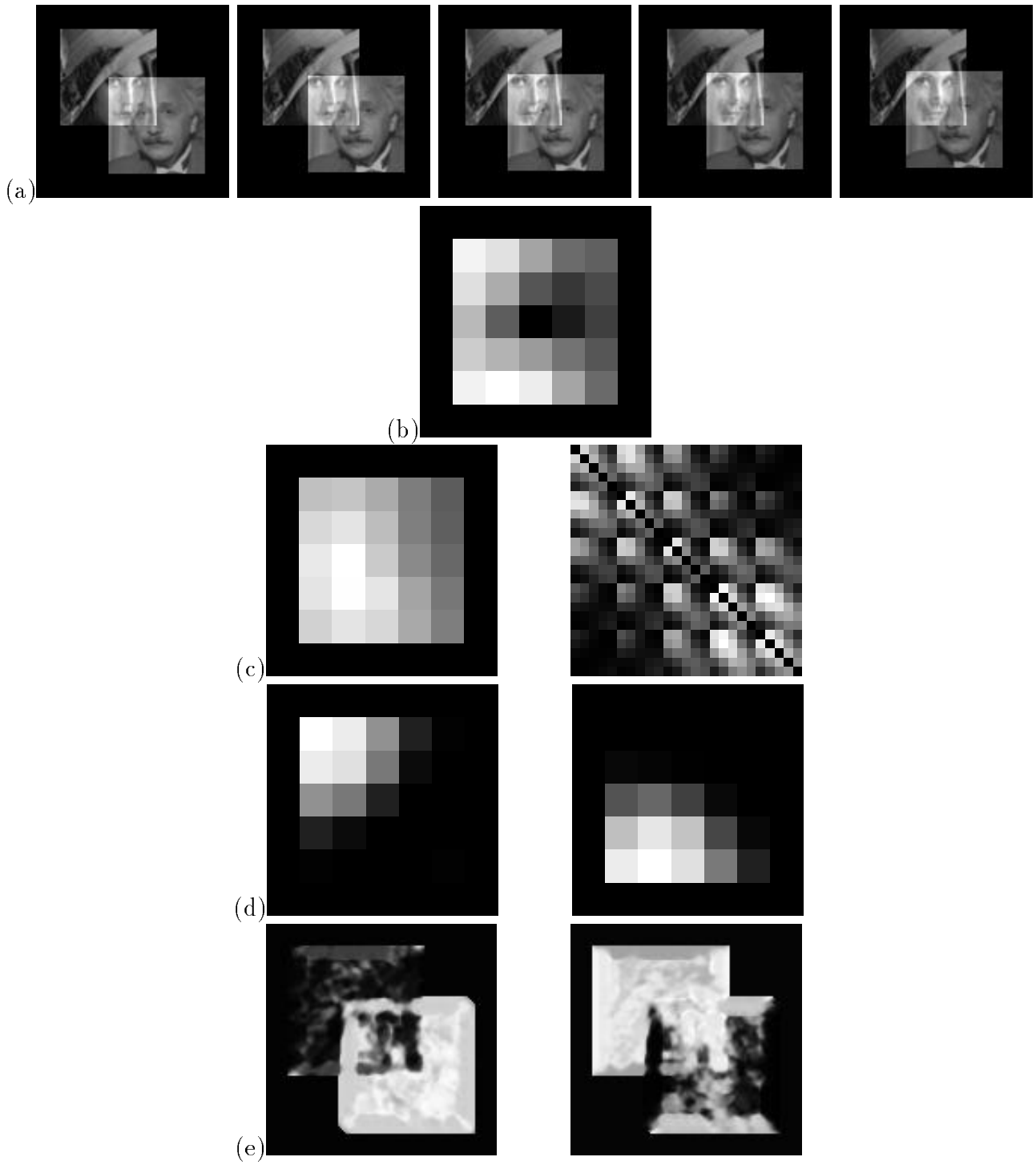


Figure 10: Selection of two layers when two motions are present in structured face image stimulus; see text for full description of figure.

velocity selective mechanisms which can test for the presence of a motion despite transparency. Using global motion hypotheses and a selection mechanism based on finding a parsimonious subset of these hypotheses, we are able to successfully decompose motion stimuli that contain additively combined transparent layers. This mechanism uses units which are tuned for a particular *global* flow field, and which have inter-unit inhibition based on similarity of the global fields.

# References


[1] E. H. Adelson. Layered representations for motion sequences. Technical Report TR-181, Vision and Modeling Group Technical Report, MIT Media Lab, December 1991.

[2] E. H. Adelson and P. Anandan. Ordinal characteristics of transparency. Technical Report TR-150, Vision and Modeling Group Technical Report, MIT Media Lab, July 1990.

[3] J. R. Bergen, P. J. Burt, K. Hanna, R. Hingorani, P. Jeanne, and S. Peleg. Dynamic multiple-motion computation. In Y. A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*, pages 147–156. Elsevier Science Publishers B.V., 1991.

[4] M. Black and P. Anandan. Constraints for the early detection of discontinuity from motion. In *Proc Eighth National Conf on Artificial Intelligence*, pages 1060–1066, July 1990.

[5] T. Darrell and A. P. Pentland. Segmentation by minimal description. In *Proceedings Third Intl. Conference on Computer Vision*, Osaka, Japan, December 1990.

[6] T. Darrell and A. P. Pentland. On the representation of occluded shapes. In *Proceedings IEEE Conference on Computer Vision*, Maui, Hawaii, June 1991.

[7] T. Darrell and A. P. Pentland. Robust estimation of a multi-layer motion representation. In *Proceedings IEEE Workshop on Visual Motion*, Princeton, October 1991.

[8] T. Darrell and A. P. Pentland. Cooperative Robust Estimation Using Layers of Support, revised submission to TPAMI, June 1993

[9] T. Darrell and E. P. Simoncelli "Nulling" Filters and the Separation of Transparent Motions In *Proceedings IEEE Conf. Computer Vision and Pattern Recognition*, New York, June 1993, also available in extended form as MIT Media Lab Vision Group TR-198.

[10] J. J. Hopfield, and D. W. Tank, Neural computation of decisions in Optimization Problems. *Biological Cybernetics,* Vol. 52, pp. 141-152. 1985.

[11] M. Husain, S. Treue, and R. Andersen. Surface interpolation in three-dimensional structure-from-motion perception. *Neural Computation*, 1:324–333, 1989.

[12] M. Irani and S. Peleg. Image sequence enhancement using multiple motions analysis. Technical Report 91-15, Department of Computer Science Technical Report, The Hebrew University of Jerusalem, Israel, December 1991.

[13] P. Meer. Robust regression methods for computer vision: A review. *Intl. J. Computer Vision*, 6:60–70, 1991.

[14] J. Mulligan. Motion transparency is restricted to two planes. *Investigative Opthalmology and Visual Science Supplement (ARVO)*, 33:1049, 1992.

[15] H. H. Nagel. On the estimation of optical flow: relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[16] A. P. Pentland. Part Segmentation for Object Recognition. *Neural Computation*, 1:82-91, 1989.

[17] M. Shizawa and K. Mase. Simultaneous multiple optical flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Atlantic City, June 1990.

[18] M. Shizawa and K. Mase. A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, Maui, June 1991.

[19] E. P. Simoncelli and E. H. Adelson. Probability distributions of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, June 1991.

[20] E. P. Simoncelli. Distributed Representation and Analysis of Visual Motion. Ph.D Thesis, MIT EECS Dept., MIT Media Lab TR-209, Janurary 1993.

[21] J. Y. A. Wang and E. Adelson. Layered representations for Motion Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York, June 1993.

[22] A. B. Watson and A. J. Ahumada. A look at motion in the frequency domain. In J. K. Tsotsos, editor, *Motion: Perception and representation*, pages 1–10. 1983.