Detecting Kinetic Occlusion

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

Abstract

Visual motion boundaries provide a powerful cue for the perceptual organization of scenes. Motion boundaries are present when surfaces in motion occlude one another. Conventional approaches to motion analysis have relied on assumptions of data conservation and smoothness, which has made analysis of motion boundaries difficult. We show that a common source of motion boundary, kinetic occlusion, can be detected using spatiotemporal junction analysis. Junction analysis is accomplished by utilizing distributed representations of motion used in models of human visual motion sensing. By detecting changes in the direction of motion in these representations, spatiotemporal junctions are detected in a manner which differentiates accretion from deletion. We demonstrate successful occlusion detection on spatiotemporal imagery containing occluding surfaces in motion.

1 Introduction

With motion boundaries as the sole cue, surfaces can easily be segmented, recognized, and perceived as having a distinct ordering in depth [19, 32]. Under natural conditions, motion boundaries are formed by occluding surfaces in motion. Relative motion of occluding surfaces generically results in accretion or deletion of surface texture of the *occluded* surface by the *occluding* surface. This basic phenomena, which we will call *kinetic occlusion*, is illustrated in Figure 1. Figure 1(a) depicts two frames of an image sequence where a tree occludes a flowerbed. As the tree moves in front of the flowerbed, portions of the flowerbed area are *accreted* and *deleted* due to relative motion between the tree and the flowerbed. Observers perceive distinct motion boundaries and an *ordinal* depth relationship between the two surfaces.

Kinetic occlusion, when viewed in space-time, possesses a structure which observers may detect and analyze. Figure 1(b) illustrates the spatiotemporal structure present in the stimulus of Figure 1(a); the (x, t) slice of (b) is obtained by stacking each frame in the image sequence of (a) to form an (x, y, t) volume, and slicing at a particular value of y. In space-time, there are spatiotemporal *junctions* indicative of the ordinal depth relationship between the two surfaces. That is, whenever spatiotemporal orientation *stops*, there is a strong cue to occlusion, i.e. *deletion* of surface texture; and, whenever spatiotemporal orientation *starts*, there is a strong cue to disocclusion, i.e. *accretion* of surface texture [6]. We have developed methods to detect kinetic occlusion and differentiate accretion and deletion.

With a detected motion boundary, interpretation of kinetic occlusion is straightforward: if the motion boundary moves consistently with either of the surfaces that define it, then that surface is the *occluding* surface [37]. If motion could be estimated accurately near motion boundaries, then edges could be detected in the velocity field. By deducing which side of the moving edge possessed motion consistent with the motion boundary, the ordinal depth relationship between two surfaces could be determined.

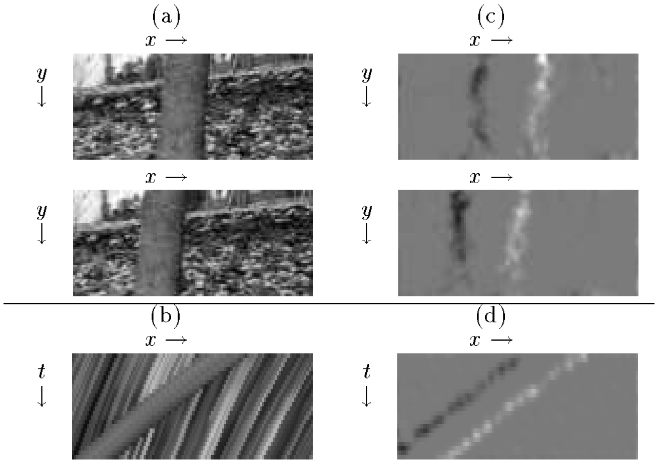


Figure 1: *Kinetic occlusion in space-time.*

However, motion estimation near motion boundaries is difficult. Conventional motion estimation algorithms assume that locally: (1) a single motion is present; (2) the intensity is constant in space time; (3) the flow field is smooth. These assumptions are violated in regions undergoing kinetic occlusion because locally: (1) there are at least two motions present; (2) intensity is not conserved; instead, texture of the occluded surface is accreted or deleted; (3) the flow field is discontinuous. Recently, many have proposed algorithms to counter these assumptions: regularizing with line processes [30, 36], generalizing local models [21, 33], using global parametric motion models [3, 5, 18, 40], and using robust estimation [4]. But all of these approaches pursue the goal of flow field estimation; few consider the spatiotemporal structure of kinetic occlusion. Exceptions are Zetsche et al. [41], who apply 3-d curvature operators directly to image sequences, and Bolles et al. [6], who bypass the detection problem that we address here, and offer a global, feature-based analysis of kinetic occlusion.

Distributed representations of motion, as used by biological visual systems, do not suffer from the problems of explicit velocity field construction. Rather than representing velocities explicitly, models of human visual motion sensing [13, 14, 34] use combinations of spatiotemporally oriented filters to construct distributed representations of velocity-tuned units. We utilize the flexibility of these representations to extract motion boundaries defined by kinetic occlusion. By constructing mechanisms which observe how distributed representations of motion change *over time*, we can detect kinetic occlusion. Here we extend the "end-stopping" mechanisms used in spatial

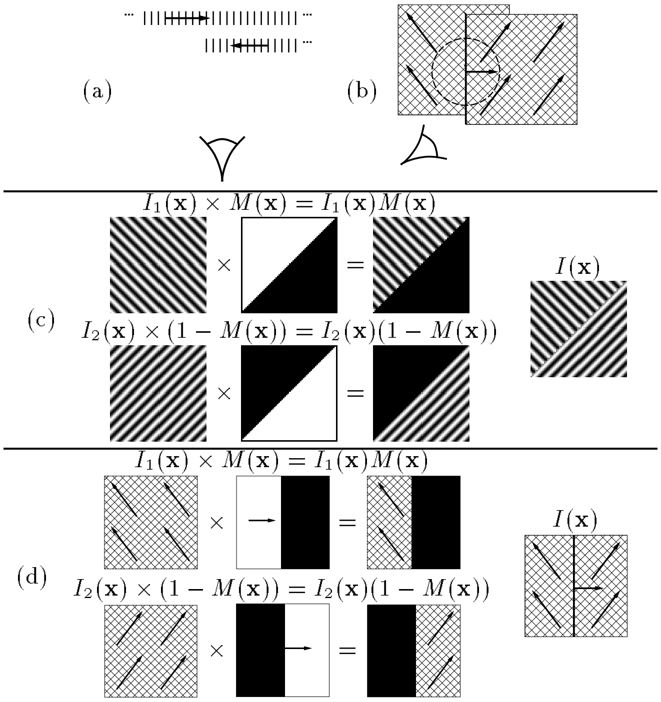


Figure 2: *Kinetic occlusion: Layers model* (a,c) where $\mathbf{x} = (x, t)$; (b,d) where $\mathbf{x} = (x, y, t)$.

junction analysis [11, 16, 31] to space-time.

By detecting motion boundaries, other visual modules may be aided: motion estimation can be improved by indicating where smoothing of motion estimates should not occur; and, as spatiotemporal junction analysis yields information similar to spatial junction analysis, our module can be used for surface segregation processes [9, 17, 27]. And, because our mechanisms are based on models of biological motion processing, we provide a model for one of most salient interactions between the motion processing pathway and the form processing pathway in biological visual systems.

2 Modeling Occlusion

We model the local structure present in occlusions directly by applying the "layers" model of Wang and Adelson [40]: a layer $I_n(\mathbf{x})$ attenuates the luminance of the layer $I_{n-1}(\mathbf{x})$ beneath it by a factor $\alpha_n(\mathbf{x})$ $(0 \leq \alpha(\mathbf{x}) \leq 1)$, and contributes its own emission $e_n(\mathbf{x})$. That is, if layer $n-1$ contains a luminance pattern $I_{n-1}(\mathbf{x})$, then the luminance pattern at layer n is:

$$I_n(\mathbf{x}) = \alpha_n(\mathbf{x}) I_{n-1}(\mathbf{x}) + e_n(\mathbf{x}) \tag{1}$$

The variable $\mathbf{x} = (x, y)$ models spatial stimuli, $\mathbf{x} = (x, t)$ models one-dimensional spatiotemporal stimuli, and $\mathbf{x} = (x, y, t)$ models two-dimensional spatiotemporal stimuli. Occlusion and transparency can be modeled through variation of the attenuation mask $\alpha_n(\mathbf{x})$.

With just two layers, $I_1(\mathbf{x})$ and $I_2(\mathbf{x})$, representing an occluded and occluding surface respectively, and an attenuation mask $M(\mathbf{x})$, a simpler model can be used:

$$I(\mathbf{x}) = I_1(\mathbf{x})M(\mathbf{x}) + I_2(\mathbf{x})(1 - M(\mathbf{x})) \tag{2}$$

We consider the above model in detail for one and two-dimensional spatiotemporal stimuli.

XT Occlusion. For one-dimensional kinetic occlusion, where $\mathbf{x} = (x, t)$, layers $I_n(\mathbf{x})$ are modeled as 1-d spatial surfaces $i_n(x)$ which move with *first order* motion v_n:

$$I_n(\mathbf{x}) = i_n(x + v_n t) \tag{3}$$

Occlusion is modeled with a step edge $u(x)$ translating with motion v_m:

$$M(\mathbf{x}) = u(x + v_m t) \tag{4}$$

Under natural conditions, the velocity of the occluder is identical to the velocity of the attenuation map, so $v_m = v_2$. Figure 2(a,c) illustrates the image formation process for an example stimulus.

XYT Occlusion. For two-dimensional kinetic occlusion, where $\mathbf{x} = (x, y, t)$, layers $I_n(\mathbf{x})$ are modeled as spatial surfaces $i_n(x, y)$ which move with first order motion $\vec{v}_n = [v_{xn}, v_{yn}]^T$:

$$I_n(\mathbf{x}) = i_n(x + v_{xn} t, y + v_{yn} t) \tag{5}$$

and occlusion is modeled with a spatially oriented step edge $u_{\theta_m}(x, y)$ translating with speed v_m and oriented at angle θ_m:

$$M(\mathbf{x}) = u_{\theta_m}(x + v_m \cos(\theta_m) t, y + v_m \sin(\theta_m) t) \tag{6}$$

Figure 2(b,d) illustrates the image formation process for an example stimulus. Figure 3 shows a geometric decomposition of $\vec{v}_1$, $\vec{v}_m$ and $\vec{v}_2$. Unlike one-dimensional motion, motion boundaries may be present in the absence of any occlusion; it is straightforward to derive the conditions under which occlusion occurs. Referring to Figure 3, simple geometric reasoning shows that occlusion of layer $I_1(\mathbf{x})$ will occur if $\vec{v}_1^n \neq \vec{v}_m^n$, or equivalently, if $\vec{v}_1 \cdot \vec{n}_{\theta_m} \neq v_m$. Assuming that the mask moves consistently with the occluding surface, i.e. $\vec{v}_2 = \vec{v}_m$, occlusion occurs whenever $\vec{v}_1 \cdot \vec{n}_{\theta_m} \neq \vec{v}_2 \cdot \vec{n}_{\theta_m}$. Thus, if $\vec{v}_1 \neq \vec{v}_2$ but $\vec{v}_1 \cdot \vec{n}_{\theta m} = \vec{v}_2 \cdot \vec{n}_{\theta m}$, there is a motion boundary but no occlusion. Under such conditions, spatiotemporal junctions cannot be detected. In motion shear, for example, surface motions $\vec{v}_1$ and $\vec{v}_2$ move tangent to the motion boundary, but neither surface is occluded.

Because motion and kinetic occlusion possess a rich structure both in the spatiotemporal domain and in the spatiotemporal frequency domain [1, 10, 14], we can develop filtering methods with simple non-linearities to analyze kinetic occlusion. Using $\hat{I}(\mathbf{w})$ to denote the Fourier transform of $I(\mathbf{x})$, layers $I_n(\mathbf{x})$ moving with velocity $\vec{v}_n$ have power constrained to lie along the line or plane $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$; in (x, t), denoting $\mathbf{w} = (w_x, w_t)$ we have that:

$$\hat{I}_n(\mathbf{w}) = \hat{i}_n(w_x)\delta(\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w}) \tag{7}$$

$$\mathbf{n}_{\vec{v}_n} = \frac{1}{\sqrt{v_n^2 + 1}}[v_n, 1]^T \tag{8}$$

where $\mathbf{n}_{\vec{v}_n}$ is the unit normal to the line $w_x v_n + w_t = 0$ and $\delta(x)$ is a Dirac delta function; in (x, y, t), denoting $\mathbf{w} = (w_x, w_y, w_t)$, we have that:

$$\hat{I}_n(\mathbf{w}) = \hat{i}_n(w_x, w_y)\delta(\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w}) \tag{9}$$

$$\mathbf{n}_{\vec{v}_n} = \frac{1}{\sqrt{v_{xn}^2 + v_{yn}^2 + 1}}[v_{xn}, v_{yn}, 1]^T \tag{10}$$

where $\mathbf{n}_{\vec{v}_n}$ is the unit normal to the plane $w_x v_{xn} + w_y v_{yn} + w_t = 0$.

The task of *first order* motion analysis is to deduce as much information as possible about the layers $I_n(\mathbf{x})$; in motion analysis, this involves representing surface velocities $\vec{v}_n(\mathbf{x})$. Estimation of the first order motions $\vec{v}_n$ is possible by finding the lines or planes $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$ containing the most energy in (x, t) or (x, y, t). Alternatively, we

$$\vec{v}_m^n = v_m \vec{n}_{\theta_m}$$
$$\vec{n}_{\theta_m} = [\cos(\theta_m), \sin(\theta_m)]^T$$
$$\vec{v}_m = \vec{v}_m^n + \vec{v}_m^t$$
$$\vec{v}_1^n = (\vec{v}_1 \cdot \vec{n}_{\theta_m})\vec{n}_{\theta_m}$$
$$\vec{v}_2^n = (\vec{v}_2 \cdot \vec{n}_{\theta_m})\vec{n}_{\theta_m}$$

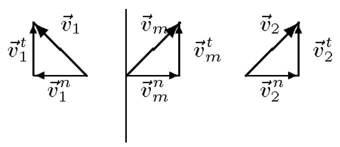


Figure 3: *XYT Kinetic occlusion: Vector decomposition.*

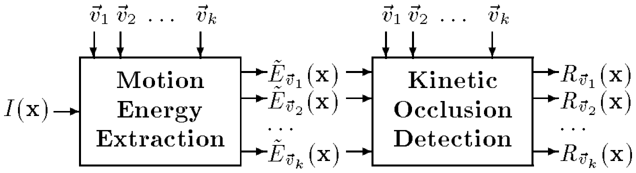


Figure 4: *Kinetic occlusion detection architecture.*

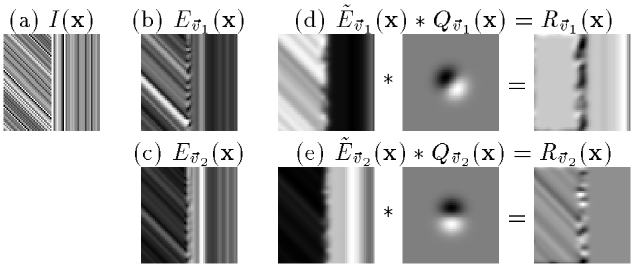


Figure 5: *XT Kinetic occlusion detection: Example.* Filters $Q_{\vec{v}_n}(\mathbf{x})$ are shown enlarged for visibility.

can construct distributed representations of motion with spatiotemporally oriented filters tuned to portions of the line in (x, t) [1] or plane in (x, y, t) [13, 14, 34].

The task of *second order* motion analysis is to deduce as much information as possible about the attenuation mask $M(\mathbf{x})$. This involves: (1) detection of occlusion; (2) determining the motion $\vec{v}_m^n$.

3 Detecting Occlusion

We focus on the detection problem exclusively here. The intuition behind our method of solving the occlusion detection problem is simple. There are three stages in our approach:

1. *Energy extraction.* We construct a distributed representation of motion $E_{\vec{v}_n}(\mathbf{x})$ which signals the presence of a given velocity $\vec{v}_n(\mathbf{x})$ at each point in space-time $\mathbf{x}$, for $n = 1 \ldots k$. We utilize "motion energy" mechanisms used in models of human visual motion sensing.

2. *Energy inhibition.* The energy extraction operation responds significantly to non-dominant motions. We inhibit $E_{\vec{v}_n}(\mathbf{x})$ with motion energy signals extracted for similar motions to produce signals $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ which respond only if the motion $\vec{v}_n(\mathbf{x})$ is dominant.

3. *Occlusion detection.* To detect occlusion, i.e. spatiotemporal junctions, each signal $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ is filtered with a first derivative of Gaussian $Q_{\vec{v}_n}(\mathbf{x})$ steered *orthogonally* to the direction of motion to construct signals $R_{\vec{v}_n}(\mathbf{x})$ which detect changes in the direction of motion.

Our architecture for kinetic occlusion detection is shown in Figure 4. We illustrate the basic operations in (x, t) first, and will describe the details behind energy extraction, energy inhibition, and occlusion detection in (x, t) and (x, y, t) in the following sections. Figure 5(a) shows a stimulus where there are two dominant motions, $v_1 = 1$ and $v_2 = 0$, corresponding to the occluded and occluding surface respectively. The energy $E_{\vec{v}_n}(\mathbf{x})$ for both motions is shown in Figure 5(b,c). Responses to non-dominant motions are significant; detecting changes in the direction of motion in this signal would yield false responses. With a simple inhibition scheme, however, we can construct signals $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ which respond only in regions where the motion $\vec{v}_n$ is dominant, as shown in Figure 5(d,e). Filtering these signals with $Q_{\vec{v}_n}(\mathbf{x})$ detects occlusion at the motion boundary. The signal $R_{\vec{v}_1}(\mathbf{x})$ indicates occlusion at the motion boundary for rightward motion; it is negative because surface texture is being deleted at the boundary. The signal $R_{\vec{v}_2}(\mathbf{x})$ generates spurious responses because static motion is not occluded.

3.1 Energy Extraction

To construct signals which respond for the presence of a given motion, we utilize motion energy schemes used in models of biological vision. For a given velocity $\vec{v}_n$ that contains power along the line/plane $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$, extracting motion energy $E_{\vec{v}_n}(\mathbf{x})$ is possible by filtering the stimulus $I(\mathbf{x})$ with banks of spatiotemporal filters centered on the line/plane [1, 13, 14].

XT Energy extraction. In (x, t), energy is extracted at a given scale on the line $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$ by filtering the signal $I(\mathbf{x})$ with a quadrature pair of filters, $G_\theta(\mathbf{x})$ and $H_\theta(\mathbf{x})$, tuned to orientation $\theta = \tan^{-1}(v_n)$. By squaring their responses, a phase-invariant "motion energy" signal $E_{\vec{v}_n}(\mathbf{x})$ can be obtained [1]:

$$E_{\vec{v}_n}(\mathbf{x}) = (I(\mathbf{x}) * G_\theta(\mathbf{x}))^2 + (I(\mathbf{x}) * H_\theta(\mathbf{x}))^2 \qquad (11)$$

Gabor functions and derivative of Gaussians are convenient forms for energy extraction; we use derivatives of Gaussians because they are steerable [12].

XYT Energy extraction. In (x, y, t), distributed representations of motion are constructed via a two-stage process, as suggested by [24]. 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. The above operations model the V1-MT "Fourier" motion pathway: simple cells in V1 are modeled with spatiotemporally oriented filters, complex cells in V1 are modeled by squaring operations [7, 8]; velocity-tuned cells in MT are modeled by summing the response of complex cells with frequency selectivities centered on a plane in the frequency domain for a given velocity [14, 34].

We utilize the method developed in [34] to extract energy along planes in the frequency domain using tools of steerable filtering [12, 20]. We summarize the method below. Convolution of an image sequence $I(\mathbf{x})$ with a set of basis filters $G_n(\mathbf{x})$ allows for steerable filtering to any portion of the frequency domain $\mathbf{w}_i = \pm[w_1, w_2, w_3]^T$, where $|\mathbf{w}| = 1$, with interpolation functions $f_n(\mathbf{w})$:

$$F_{\mathbf{w}_i}(\mathbf{x}) = \sum_n f_n(\mathbf{w}_i)(G_n(\mathbf{x}) * I(\mathbf{x})) \qquad (12)$$

Figure 6(a) summarizes steering functions for Nth order derivative of Gaussians ($N = \{1, 2\}$). Filter coefficients used in all of our experiments can be found in [28]. To tile the plane $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$, energy is extracted with $N + 1$

(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	$F_{\mathbf{W}_i}(\mathbf{x}) = I_{xx}(\mathbf{x})w_1^2 + 2I_{xy}(\mathbf{x})w_1w_2 + 2I_{xt}(\mathbf{x})w_1w_3 + I_{yy}(\mathbf{x})w_2^2 + 2I_{yt}(\mathbf{x})w_2w_3 + I_{tt}(\mathbf{x})w_3^2$ $\mathbf{w}_1 = \mathbf{w}_a, \mathbf{w}_2 = \frac{1}{2}\mathbf{w}_a + \frac{\sqrt{3}}{2}\mathbf{w}_b, \mathbf{w}_3 = \frac{\sqrt{3}}{2}\mathbf{w}_a + \frac{1}{2}\mathbf{w}_b$

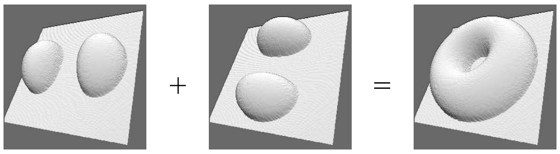
(b) 

Figure 6: *XYT motion energy extraction.*

weightings centered at locations $\mathbf{w}_i$, as summarized in Figure 6(a). By filtering the image sequence $I(\mathbf{x})$ with filters with frequency selectivities centered at those locations, and subsequently applying squaring operations, we extract energy $E_{\vec{v}_n}(\mathbf{x})$ along a ring on the plane via:

$$E_{\vec{v}_n}(\mathbf{x}) = G_0(\mathbf{x}) * \sum_{i=1}^{N+1} (F_{\mathbf{W}_i}(\mathbf{x}))^2 \qquad (13)$$

High frequency components introduced by squaring are diminished through subsequent convolution with a 3-d Gaussian $G_0(\mathbf{x})$. Figure 6(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.

3.2 Energy Inhibition

The above energy extraction scheme responds significantly to non-dominant motions; detecting changes in the direction of motion would result in false responses. We wish to construct inhibited motion energy signals $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ which respond only if $\vec{v}_n$ is dominant at point $\mathbf{x}$. Only motion which is locally dominant can be made to respond in $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ with simple non-linear operations on energy signals extracted for similar motions:

$$\tilde{E}_{\vec{v}_n}(\mathbf{x}) = f\left(\frac{E_{\vec{v}_n}(\mathbf{x})}{\sigma + \sum_{\delta\vec{v}} E_{\vec{v}_n+\delta\vec{v}}(\mathbf{x})}\right) \qquad (14)$$

where $f(x)$ is a sigmoid function:

$$f(x) = \frac{1}{1 + e^{-c(x-\tilde{x})}} \qquad (15)$$

which is near 1 if $x \gg \tilde{x}$ and near 0 if $x \ll \tilde{x}$. The above operation implements a soft local winner-take-all mechanism, i.e. $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ only responds if the energy $E_{\vec{v}_n}(\mathbf{x})$ extracted at a velocity $\vec{v}_n$ is significantly greater than the energy $E_{\vec{v}_n+\delta\vec{v}}(\mathbf{x})$ extracted at neighboring motions $\vec{v}_n + \delta\vec{v}$. At high contrasts, division by neighboring orientations results in contrast invariance. At low contrasts, σ dominates, so the response is near zero. We have found that a winner-take-all scheme is necessary; simple normalization by local energy, such as in the model of [15], does not provide the level of inhibition that we need.

Choosing the neighborhood of $\vec{v}_n + \delta\vec{v}$ is straightforward. The unit normal $\mathbf{n}_{\vec{v}_n}$ to the line/plane $\mathbf{n}_{\vec{v}_n} \cdot \mathbf{w} = 0$ is used to find two/four unit normals $\mathbf{n}_{\vec{v}_n+\delta\vec{v}}$ at fixed angles to $\mathbf{n}_{\vec{v}_n}$. Each normal vector $\mathbf{n}_{\vec{v}_n+\delta\vec{v}}$ defines a neighboring line/plane $\mathbf{n}_{\vec{v}_n+\delta\vec{v}} \cdot \mathbf{w} = 0$ whose energy $E_{\vec{v}_n+\delta\vec{v}}(\mathbf{x})$ is extracted in the same manner.

3.3 Occlusion Detection

Given signals $E_{\vec{v}_n}(\mathbf{x})$ which respond only if motion $\vec{v}_n$ is dominant locally, we construct a signal $R_{\vec{v}_n}(\mathbf{x})$ which signals the disappearance or appearance of first order motion $\vec{v}_n$:

$$R_{\vec{v}_n}(\mathbf{x}) = Q_{\vec{v}_n}(\mathbf{x}) * \tilde{E}_{\vec{v}_n}(\mathbf{x}) \qquad (16)$$

where $Q_{\vec{v}_n}(\mathbf{x})$ is a first derivative of a Gaussian steered [12] *orthogonally* to the direction of motion. In (x, t):

$$Q_{\vec{v}_n}(\mathbf{x}) = [G_x(\mathbf{x}), G_t(\mathbf{x})]^T \cdot \mathbf{n}_{\vec{v}_n} \qquad (17)$$

and $G_x(\mathbf{x})$ and $G_t(\mathbf{x})$ are derivatives of 2-d Gaussians in the x, y, and t directions. In (x, y, t):

$$Q_{\vec{v}_n}(\mathbf{x}) = [G_x(\mathbf{x}), G_y(\mathbf{x}), G_t(\mathbf{x})]^T \cdot \mathbf{n}_{\vec{v}_n} \qquad (18)$$

and $G_x(\mathbf{x})$, $G_y(\mathbf{x})$, and $G_t(\mathbf{x})$ are derivatives of 3-d Gaussians in the x, y, and t directions.

By detecting changes in the direction of motion, the *sign* of $R_{\vec{v}_n}(\mathbf{x})$ differentiates accretion from deletion. Deletion of surface texture is reflected by negative responses in $R_{\vec{v}_n}(\mathbf{x})$; accretion of surface texture is reflected by positive responses.

In [28], we provide an in depth interpretation of this occlusion model and detection algorithm. We show that the image formation model involves *modulation* of information about the mask $M(\mathbf{x})$ along the line/plane corresponding to the occluded surface $I_1(\mathbf{x})$. We show that the energy extraction operations commonly employed in the models of human visual motion sensing used here *demodulate* both the information about the mask $M(\mathbf{x})$ and the layer $I_1(\mathbf{x})$; the energy inhibition and occlusion detection operations used here isolate information about the mask.

4 Experiments

We have applied our kinetic occlusion detection algorithm to a battery of two-dimensional synthetic spatiotemporal stimuli as well as a few real stimuli. Further experimentation is reported in [28].

In our experiments, simulations were simplified tremendously by only detecting kinetic occlusion for k dominant motions. Dominant motions $\vec{v}_n(\mathbf{x})$ are extracted via the k-means clustering technique used in [40]; the technique uses k-means clustering to local affine parameter fits of optical flow to find k dominant motions described by:

$$\vec{v}_n(\mathbf{x}) = \begin{bmatrix} a_x(t) + a_{xx}(t)x + a_{xy}(t)y \\ a_y(t) + a_{yx}(t)x + a_{yy}(t)y \end{bmatrix} \qquad (19)$$

where $\vec{a}_n(t) = [a_x, a_{xx}, a_{xy}, a_y, a_{yx}, a_{yy}]^T$ represents affine parameter fits for a cluster at time t. Any parametric scheme (c.f. [3, 5, 18]) could easily be substituted.

Figure 7 shows stages of our occlusion detection algorithm on a 64x32x36 random-dot stimulus where a rightward moving circle occludes a leftward moving square, both of which occlude a static background. Two representative (x, y) frames, and an (x, t) slice are shown in each part. In (a), we depict an idealized segmentation of the random dot stimulus shown in (b). There are three dominant motions ($k = 3$) corresponding to the three layers of the background, the square, and the circle ($\vec{v}_1 = [0, 0]^T, \vec{v}_2 = [-1, 0]^T, \vec{v}_3 = [1, 0]^T$).

Figure 7(c,d,e) shows the motion energy $E_{\vec{v}_n}(\mathbf{x})$ extracted for the three layers. Significant responses are present in regions where the motion is not dominant. Detecting changes in the direction of motion would result in spurious responses.

Figure 7(f,g,h) shows the inhibited motion energy signals $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ extracted with our local winner-take-all mechanism. Now, responses are present only in regions where the motion $\vec{v}_n(\mathbf{x})$ is dominant.

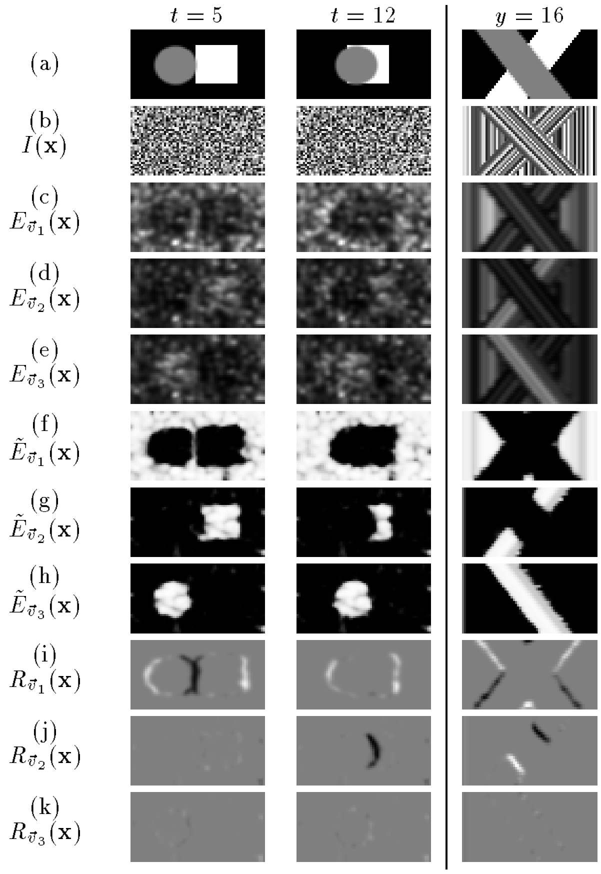


Figure 7: *XYT Kinetic occlusion detection: Example.*

Figure 7(i,j,k) shows occlusion detected $R_{\vec{v}_n}(\mathbf{x})$ by filtering $\tilde{E}_{\vec{v}_n}(\mathbf{x})$ orthogonal to the direction of motion. Positive responses indicate that the texture of the background is being accreted, whereas negative responses indicate the texture of the posterior surface is being deleted. In (i), responses in $R_{\vec{v}_1}(\mathbf{x})$ are obtained wherever the static background is accreted or deleted due to the foreground motions of the circle and square. In the region directly above and below the square and circle, there is a motion boundary but no kinetic occlusion, so little response is obtained. In (j), responses in $R_{\vec{v}_2}(\mathbf{x})$ are obtained when the square is occluded by the circle. In (k), responses in $R_{\vec{v}_3}(\mathbf{x})$ are negligible because the circle is never occluded.

Figure 1(c,d) shows the application of our algorithm to the stimulus of Figure 1(a,b), where a tree occludes a flowerbed. The motion of the tree and flowerbed is represented accurately by the affine model we have adopted. Figure 1(c) shows detected occlusion $R_{\vec{v}_1}(\mathbf{x})$ for the flowerbed layer. At the right side of the tree, the tree disoccludes the flowerbed, so $R_{\vec{v}_1}(\mathbf{x})$ responds positively at the motion boundary. At the left side of the tree, the tree occludes the flowerbed, so $R_{\vec{v}_1}(\mathbf{x})$ responds negatively. In Figure 1(d), a spatiotemporal (x, t) slice of $R_{\vec{v}_1}(\mathbf{x})$ reveals the robustness of the response over time. The response of $R_{\vec{v}_2}(\mathbf{x})$ is negligible in comparison, so we do not show it here.

5 Discussion

There are a variety of additional directions that can be taken to extend our kinetic occlusion detection architecture. First, considerable improvements can be made on the details of our detection methods alone. Other mechanisms which signal the presence of a given orientation and motion can be easily constructed. Combining techniques used in spatial junction analysis [11, 16, 31] and our spatiotemporal junction analysis in a unified framework is in order. Second, we are currently using non-causal operations. An asymmetry in detecting appearance and disappearance is expected; considering the temporal dynamics of distributed winner-take all mechanisms could handle this asymmetry in part. Third, while we have explored occlusion exclusively here, other motion stimuli, such as multiplicative transparency, can also be considered as second order motion stimuli [10]; providing a general model of second order motion analysis is desirable. Most importantly, our current analysis is purely local; no spatiotemporal propagation has been attempted. Solutions to spatial contour completion [17, 27] are likely to be extendable to spatiotemporal surface completion.

Since our models of first order motion analysis are based on models of human visual motion sensing, we suggest that we have modeled the computations necessary for one of the *interactions* between the motion and form processing streams in primate vision (c.f. [39]). Lesion studies in monkeys and humans offer a means of deducing which visual areas might be involved in detecting motion boundaries. Ablation of area MT causes monkeys' performance on motion-defined shape discrimination tasks to be severely impaired [23], while psychophysical experiments on humans with parietotemporal lesions indicates a double disassociation between motion discontinuity detection and motion coherence discrimination [38]. This evidence suggests that there are separate neural substrates for representing discontinuities and computing motion, probably involving MT/MST.

Selectivity for the orientation of motion boundaries, defined by kinetic occlusion *and* relative motion, has been tested in visual areas V1, V2, V3, and MT through single cell recordings. We should recognize, of course, that motion boundary extraction can be based on both relative motion cues *and* occlusion cues. Mechanisms which may process relative motion cues, such as the center-surround mechanisms proposed by [26], have been found (c.f. [2]), thus complicating interpretations of physiological data. While V1 and MT cells are not selective to the orientation of kinetic boundaries, a considerable percentage of cells in V2 and V3 possess identical orientation selectivity for both luminance and kinetic-defined boundaries [22, 29]. Furthermore, cells in ventral MST have been shown to have their responses modulated by accretion and deletion cues [35]. Since our model only involves spatiotemporally filtering distributed representations of motion believed to be present in MT, a reasonable conjecture is that the change detection operation we advocate is done with a feedback projection from a "motion" area with velocity-tuned cells, such as MT and/or MST, to one or more "form" areas, such as V2, V3, and/or V4 representing boundary information.

Biological modeling considerations aside, we have demonstrated feasibility of detecting motion boundaries defined by kinetic occlusion, and offer a new tool which can be used in a variety of computer vision tasks. First, surfaces can be segregated on the basis of kinetic occlusion. Detecting edges in velocity fields [37] and detecting appearance and disappearance in token-tracking schemes [25] is difficult; we offer an alternate possibility. Second, by discontinuing use of optical flow upon occlusion detection, and continuing use of optical flow upon disocclusion detection, we can potentially track surfaces *through* occlusion. Segmentation and tracking tools such as deformable contours and surfaces may use our detected motion boundaries as an image potential. Furthermore, with a detected ordinal cue to depth, we can determine the depth ordering of surfaces locally: our detected occlusion can be used to activate direction of figure representations, as is done in the models of [9, 17]. A global depth ordering of surfaces can also be attempted from these local cues [40]. Finally, the detection of motion boundaries can allow for improved motion estimation: in addition to activating line processes due to velocity gradients, we can activate them by detecting kinetic occlusion.

Motion estimation algorithms, relying on "constant intensity" and "single velocity" assumptions, have always had difficulty near motion boundaries. Providing a better model of motion boundaries is of paramount importance to two-dimensional motion analysis. Unlike conventional approaches to kinetic occlusion analysis, our approach recognizes the spatiotemporal structure present in occlusion. We provide a new tool for motion analysis and surface segmentation processes, and provide a model for one of the interactions between the motion and form processing streams in biological visual systems.

Acknowledgements

This work was sponsored by the Television of Tomorrow Consortium of the MIT Media Lab. Edward Adelson provided much insights, advice and support. The tools provided by Eero Simoncelli, John Wang, and William Freeman proved invaluable. William Freeman, Yair Weiss, Martin Szummer, David Askey, and Ujjaval Desai provided helpful suggestions on various versions of this document.

References

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

[2] J. Allman, F. Miezin, and E. McGuiness. Direction and velocity-specific responses from beyond the classical receptive field in MT. *Perception*, 14:105–126, 1985.

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

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

[5] M. Black and A. Jepson. Estimating optical flow in segmented images using variable order parametric models with local deformations. In *Workshop on Non-Rigid Motion and Articulated Objects*, Austin, TX, 1994.

[6] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *IJCV*, 1(1):7–56, 1987.

[7] G. DeAngelis, I. Ohzawa, and R. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex - I. General characteristics and postnatal development. *Journal of Neurophysiology*, 69(4):1091–1117, 1993.

[8] R. Emerson, J. Bergen, and E. Adelson. Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research*, 32(2):203–18, 1992.

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

[10] D. Fleet and K. Langley. Computational analysis of non-Fourier motion. *Vision Research*, 34(22):3057, 1994.

[11] W. Freeman. *Steerable filters and the local analysis of image structure*. PhD thesis, MIT Dept. Media Arts and Sciences, September 1992.

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

[13] N. Grzywacz and A. 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.

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

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

[16] F. Heitger, L. Rosenthaler, R. von der Heydt, E. Peterhans, and O. Kubler. Simulation of neural contour mechanisms: From simple to end-stopped cells. *Vision Research*, 32(5):963–981, 1992.

[17] 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.

[18] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 12(1):1–16, 1994.

[19] G. A. Kaplan. Kinetic disruption of optical texture: The perception of depth at an edge. *Perception and Psychophysics*, 6(4):193–198, 1969.

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

[21] K. Langley, D. Fleet, and T. Atherton. Multiple motions from instantaneous frequency. In *CVPR*, pages 846–849, 1992.

[22] V. Marcar and A. Cowey. Do cells in area V2 respond to the orientation of kinetic boundaries? In *Society of Neuroscience Abstracts*, volume 18, page 1275, 1992.

[23] V. Marcar and A. Cowey. The effect of removing superior temporal cortical motion areas in the macaque monkey: II. Motion discrimination using random dot displays. *European Journal of Neuroscience*, 4:1228–1238, 1992.

[24] A. Movshon, E. Adelson, M. Gizzi, and W. Newsome. The analysis of moving visual patterns. *Experimental Brain Research*, 11:117–152, 1986.

[25] K. Mutch and W. Thompson. Analysis of accretion and deletion at boundaries in dynamic scenes. *IEEE PAMI*, 7(2):133–138, 1985.

[26] K. Nakayama and J. Loomis. Optical velocity patterns, velocity-sensitive neurons, and space perception: A hypothesis. *Perception*, 3:63–80, 1974.

[27] M. Nitzberg, D. Mumford, and S. Shiota. *Filtering, Segmentation and Depth*, volume 622. Springer-Verlag, 1993.

[28] S. Niyogi. Kinetic occlusion. Technical Report 319, MIT Media Lab Vision and Modeling, 1995. ftp://whitechapel.media.mit.edu/pub/tech-reports

[29] E. Peterhans and J. Baumann. Elements of form processing from motion in monkey prestriate cortex. In *Society of Neuroscience Abstracts*, volume 20, page 1053, 1994.

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

[31] L. Rosenthaler, F. Heitger, O. Kubler, and R. von der Heydt. Detection of general edges and keypoints. In *ECCV*, pages 78–86, 1992.

[32] C. S. Royden, J. F. Baker, and J. Allman. Perceptions of depth elicited by occluded and shearing motions of random dots. *Perception*, 17:289–296, 1988.

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

[34] E. Simoncelli. *Distributed Representations of Motion*. PhD thesis, MIT Dept. of Electrical Engineering, January 1993.

[35] Y. Sugita and K. Tanaka. Occlusion-related cue used for analysis of motion in visual cortex. *NeuroReport*, 2:751–754, 1991.

[36] D. Terzopoulos. Regularization of inverse problems involving discontinuities. *IEEE PAMI*, 8:413–424, 1986.

[37] W. Thompson, K. Mutch, and V. Berzins. Dynamic occlusion analysis in optical flow fields. *IEEE PAMI*, 7(4):374–383, 1985.

[38] L. Vaina, N. Grzywacz, and R. Kikinis. Segregation of computations underlying perception of motion discontinuity and coherence. *NeuroReport*, 5(17):2289, 1994.

[39] D. C. van Essen, C. Anderson, and D. Felleman. Information processing in the primate visual system. *Science*, 255(5043):419–422, 1992.

[40] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Image Processing*, 3(5):625–638, 1994.

[41] C. Zetzsche, E. Barth, and J. Berkmann. Spatiotemporal curvature measures for flow field analysis. In *SPIE Geometric Methods in Computer Vision*, volume 1570, pages 337–350, 1991.