To appear in a future issue of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI).
The paper has not been fully edited and thus the content is subject to change prior to final publication.

The Structure of Multiplicative Motions in Natural Imagery

Konstantinos G. Derpanis and Richard P. Wildes

***Abstract*—A theoretical investigation of the frequency structure of multiplicative image motion signals is presented, e.g., as associated with translucency phenomena. Previous work has claimed that the multiplicative composition of visual signals generally results in the annihilation of oriented structure in the spectral domain. As a result, research has focused on multiplicative signals in highly specialized scenarios, where highly structured spectral signatures are prevalent, or introduced a non-linearity to transform the multiplicative image signal to an additive one. In contrast, in this paper it is shown that oriented structure is present in multiplicative cases when natural domain constraints are taken into account. This analysis suggests that the various instances of naturally occurring multiple motion structures can be treated in a unified manner. As an example application of the developed theory, a multiple motion estimator previously proposed for translation, additive transparency and occlusion is adapted to multiplicative image motions. This estimator is shown to yield superior performance over the alternative practice of introducing a non-linear preprocessing step.**

***Index Terms*—Multiplicative motion, translucency, dynamic occlusion, pseudo-transparency, non-Fourier motion, spectral analysis, optical flow, multiple motion.**

I. Introduction

The study of spatiotemporal structure in vision is dominated by *optical flow* approaches. A fundamental assumption of these formulations is that a single motion is present within a finite image region of analysis. This single motion assumption commonly appears in the form of the conservation of some image feature property (e.g., brightness [20], phase [15], etc.). The performance of optical flow approaches, as measured on the synthetic *Yosemite* sequence [5], has steadily improved to the point where state-of-the-art algorithms obtain an impressive *average angular error* (AAE) of 2^o (equivalent to approximately 0.1 pixels). However, caution should be taken when using these results to predict performance with real-world imagery. As pointed out by several researchers (e.g., [3], [5]), the *Yosemite* sequence is a relatively simple example that does not contain significant multiple motion phenomena, as one frequently encounters in nature as dynamic occlusion, pseudo-transparency (e.g., partially obscuring foliage) and transparency/translucency (e.g., stained glass, atmospheric effects, lighting and shadows). To highlight these issues and others, there has been growing interest in the community in introducing challenging real-world data sets with ground truth [4], [24]. Not surprisingly, on these new data sets the state-of-the-art approaches perform relatively poorly in regions not conforming to the intrinsic assumptions of the optical flow algorithms.

The introduction of challenging test data sets highlights the limitations in the dynamic image models that underlie extant image motion estimators based on optical flow. From a practical point of view, some of these difficulties can be surmounted by making use of robust estimation procedures that treat data violations of standard optical flow assumptions as outliers [25]. Nevertheless, it remains highly desirable to develop models that capture the structure of spatiotemporal image data where optical flow assumptions fail. Such models can serve both to further the theoretical understanding of dynamic imagery as well as provide the basis for more sophisticated estimation procedures that yield accurate and precise estimates in application to the complexities of real-world data.

In this paper, a theoretical investigation is presented on the frequency structure of multiplicative spatiotemporal phenomena, such as translucency and dynamic occlusion. It appears that Fleet presented the earliest analysis of multiple motions in the frequency domain [14]. Subsequent research proposed computational schemes for recovering the image velocity of constituent components within this framework [6], [16]. These approaches focused on a world where constituent patterns were composed of a few spectral components (e.g., sinusoids and plaids). In the real-world, the spectra of image patterns is typically of a broadband nature [27]. For the case of dynamic occlusion with broadband signals, Yu et al. [30] demonstrated that the spectral features relied on in earlier work, [6], [16], are not reliable. Key to their analysis is understanding spatiotemporal structure in a more natural domain rather than in some highly contrived one. In the present paper, this idea is further pursued by introducing an additional natural domain constraint, that of non-negativity of the image signal and attenuation/transmittance in the signal composition stage. It will be demonstrated that the addition of this simple constraint imposes oriented spatiotemporal structure that was previously claimed to be lost in the multiplicative composition of multiple motions.

II. Preliminaries

A. Relevance of frequency analysis

The *Fourier transform* is a global transform and as such care must be taken in extrapolating results to local phenomena. A common property among the phenomena to be studied is that they are characterized by linear structures in the spectral domain. These structures represent idealizations. In practice, as a consequence of the *uncertainty principle* [8], these structures are subject to blurring by the window of analysis. The use of smooth windows (e.g., a *Kaiser window* [17]) can ameliorate this problem to some degree but will not remove it completely. The use of larger windows can also reduce this problem; however, this increases the possibility of mixing simple local structures. This dilemma represents an instance of the *generalized aperture problem* [21].

In order to apply the Fourier transform, the signal must conform to the *Dirichlet* conditions [8]. These conditions require that over any interval the signal is absolutely integrable, of bounded variation and that it has a finite number of discontinuities, each of which is finite. Since any measured physical signal satisfies these conditions, the analysis is ensured to hold for arbitrary natural image sequences.

K. Derpanis and R. Wildes are with the Department of Computer Science and Engineering, and Centre for Vision Research (CVR), York University, Toronto, Canada.
E-mail: {kosta,wildes}@cse.yorku.ca

B. Translation

Consider an image signal, $I(\mathbf{x}, t)$, parameterized in terms of spatial coordinates, $\mathbf{x} = (x, y)^\top$, and time, t, as it moves with velocity, $\mathbf{v} = (u, v)^\top$. The corresponding spectrum is given by [2], [13], [18], [29],

$$\tilde{I}(\mathbf{k}, \omega_t) = \tilde{I}(\mathbf{k})\delta(\mathbf{k}^\top \mathbf{v} + \omega_t), \tag{1}$$

where $\mathbf{k} = (\omega_x, \omega_y)^\top$ denotes the spatial frequency vector, ω_t the temporal frequency, $\tilde{}$ denotes the Fourier transform of the corresponding signal and $\delta()$ is the *Dirac delta* function. Geometrically, this can be interpreted as the spectrum being restricted to a plane through the origin with normal $(\mathbf{v}, 1)^\top$. In the 2D case, consisting of a single spatial dimension, x or y, and time, t, the planar spectra reduces to a line through the origin. This motion is often referred to as *first-order motion* or *Fourier motion*; whereas, the multiplicative types of motion stimuli of concern in this paper are often referred to as *non-Fourier motion* [16].

C. Generative model

For the cases of multiple motions considered in this paper, the following recursive procedure is used as the generative model for obtaining the final image from component layers [1]. Assume that the depth ordering of the N layers relative to the viewer is given, where the layer composition results are denoted $I_0(\mathbf{x}), \ldots, I_{N-1}(\mathbf{x})$. At each pixel, each layer, n, may partially transmit the total amount of light from the layers beneath it by a transmittance factor of $T_n(\mathbf{x})$, where $0 \leq T_n(\mathbf{x}) \leq 1$, and may contribute its own emission of quantity $E_n(\mathbf{x})$, where $E_n(\mathbf{x}) \geq 0$. The non-negativity of the emission term, $E_n(\mathbf{x})$, follows from the fact that it represents power per unit foreshortened area per unit solid angle (*radiance*) and thereby cannot take on negative values. The boundary cases of the transmittance factor consisting of zero and one, indicate that the light from the previous layers is fully attenuated and fully transmitted, respectively. The final composite image is the result of applying this process recursively from back-to-front, formally,

$$I_n(\mathbf{x}) = T_n(\mathbf{x})I_{n-1}(\mathbf{x}) + E_n(\mathbf{x}), \tag{2}$$

where $n \geq 0$ and $I_0(\mathbf{x}) \equiv E_0(\mathbf{x})$. Strictly speaking, the image signal is given in terms of irradiance, while $I_n(x)$ is given as scene radiance in the generative model, (2); however, since image irradiance is proportional to scene radiance [19], this distinction is neglected here, as it has been in developing other applicable analyses of transparency, e.g. [1], [28].

In the sequel, the generative model, (2), is used as a basis for understanding the frequency structure of various dynamic multiplicative phenomena. Without loss of generality, the number of layers under consideration will be restricted to two. As in [30], the focus here is on broadband signals, which is the typical case for real-world signals. A novel aspect of the present model in comparison to previous formulations used to understand the frequency structure of dynamic imagery [6], [14], [16], [30], is the explicit introduction of the non-negativity constraint of the image signal and transmittance. It will be demonstrated that enforcing this constraint yields oriented structure in the frequency domain that was conjectured in earlier work to be annihilated in the composition process [14], [16]. Interestingly, the constraint of non-negativity of the image signal has previously appeared in work concerning the simultaneous reconstruction of component images and recovery of motions in imagery containing reflections and transparency [28]; however, the authors did not pursue the implication of the non-negativity constraint on the explicit structure of the signal. In terms of the generative model, (2), this case corresponds to a spatially constant transmittance term.

III. Spectral analysis of multiplicative motion

A. Translucency

Assume that an image signal, $I_0(\mathbf{x})$, is viewed through a non-refractive translucent layer with transmittance $T_1(\mathbf{x})$. If components $I_0(\mathbf{x})$ and $T_1(\mathbf{x})$ are moving with velocities $\mathbf{v}_0$ and $\mathbf{v}_1$, respectively, using the generative model, (2), the image sequence signal can be written as

$$I_1(\mathbf{x}, t) = T_1(\mathbf{x} - \mathbf{v}_1 t)I_0(\mathbf{x} - \mathbf{v}_0 t). \tag{3}$$

From the generative model, the transmittance factor is strictly non-negative. Consequently, $T_1(\mathbf{x})$ can be reexpressed as the sum of a constant/DC term α and a zero mean signal, $T(\mathbf{x}) = T_1(\mathbf{x}) - \alpha$. Furthermore, to reflect the non-negative nature of image signals, $I_0(\mathbf{x})$ can be reexpressed as a constant/DC term β plus a zero mean signal, $I(\mathbf{x}) = I_0(\mathbf{x}) - \beta$. Including these constraints in (3), yields,

$$\begin{aligned} I_1(\mathbf{x}, t) &= \Big(\alpha + T(\mathbf{x} - \mathbf{v}_1 t)\Big)\Big(\beta + I(\mathbf{x} - \mathbf{v}_0 t)\Big) \\ &= \alpha\beta + \alpha I(\mathbf{x} - \mathbf{v}_0 t) + \beta T(\mathbf{x} - \mathbf{v}_1 t) \\ &\quad + T(\mathbf{x} - \mathbf{v}_1 t)I(\mathbf{x} - \mathbf{v}_0 t). \end{aligned} \tag{4}$$

From the standard *Fourier motion* result (Section II-B), the *superposition* property and the *convolution theorem* of the Fourier transform [8], it can be easily shown that the Fourier transform of (4) is

$$\begin{aligned} \tilde{I}_1(\mathbf{k}, \omega_t) &= \alpha\beta\delta(\mathbf{k}, \omega_t) \\ &\quad + \alpha\tilde{I}(\mathbf{k})\delta(\mathbf{k}^\top \mathbf{v}_0 + \omega_t) + \beta\tilde{T}(\mathbf{k})\delta(\mathbf{k}^\top \mathbf{v}_1 + \omega_t) \\ &\quad + \Big(\tilde{T}(\mathbf{k})\delta(\mathbf{k}^\top \mathbf{v}_1 + \omega_t)\Big) * \Big(\tilde{I}(\mathbf{k})\delta(\mathbf{k}^\top \mathbf{v}_0 + \omega_t)\Big), \end{aligned} \tag{5}$$

where $*$ symbolizes the convolution operator. Assuming broadband component signals, the first term corresponds to a DC term. The second and third terms correspond to two oriented spectral planes. Their normal vectors $(\mathbf{v}_0, 1)^\top$ and $(\mathbf{v}_1, 1)^\top$ denote their respective layer velocities. The final term corresponds to the convolution between two 3D planes that yields a non-oriented structure in the case of broadband signals. Finally, one can include the emission term, $E_1(\mathbf{x})$, that will result in the strengthening of the planar spectral structure of the translucent layer.

Figure 1 illustrates the frequency spectra for the various terms of (5) and their compositional result. For illustrative

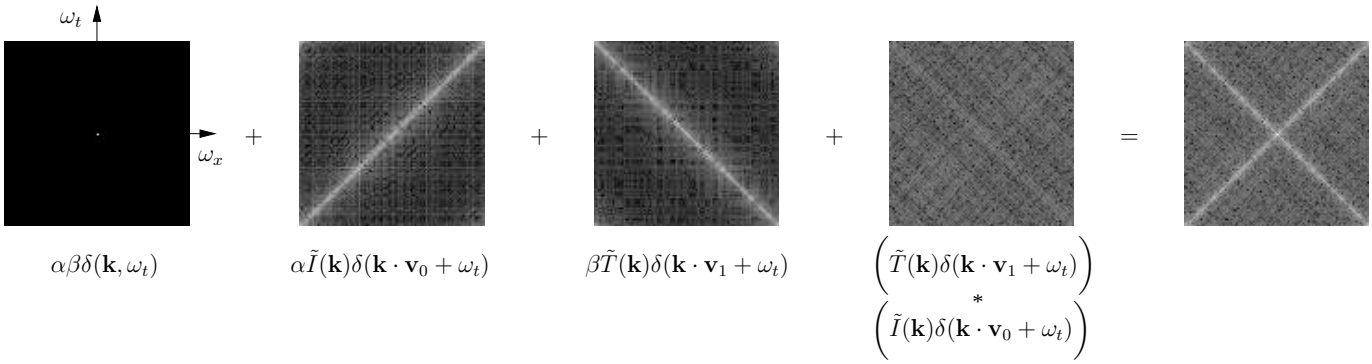


Fig. 1. Synthetic multiplicative transparency image sequence. The term-by-term composition of a multiplicative transparency is illustrated in the 2D, $\omega_x - \omega_t$, frequency domain (magnitude terms displayed), as given by (5). The white noise structures are moving in opposite directions with a speed of 1 pixel/frame. A *Kaiser window* in the spatiotemporal domain was used to reduce windowing distortions for all terms, except the DC term. For display purposes, the logarithm of the spectrum is displayed.

purposes attention is restricted to the 2D case (x-t). The constituent image signals are both white noise moving in opposite directions with a speed of 1 pixel/frame. As can be seen, enforcing the non-negativity constraint on the image signal by way of introducing DC components reveals the oriented structure of the constituent surfaces with the convolution (distortion) term acting as a non-oriented noise-like backdrop. In the case where both DC terms are zero, a clear violation of the non-negativity constraint, the translucency reduces to the non-oriented term,

$$\left(\tilde{T}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_1+\omega_t)\right) * \left(\tilde{I}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_0+\omega_t)\right). \quad (6)$$

In Fig. 2, a real translucency example is presented. This example consists of a painting moving behind a spatially varying translucent material, that is also in motion, and captured by a stationary video camcorder. In the epipolar slice image, two symmetric diagonal oriented structures are clearly evident. Correspondingly, the main power in the spectral domain is dominated by two lines through the origin. These structures are consistent with the constant leftward and rightward motions present within the analysis window.

Structured lighting and shadows can also be modeled as multiplicative motions as the local surface albedo determines the proportion of impinging light that is reflected. In Fig. 3, a real structured light example is presented. This example consists of a moving structured light illuminating a textured surface moving in the opposite direction, captured by a stationary video camcorder. In the epipolar slice image, two symmetric diagonal oriented structures are clearly evident. Correspondingly, the main power in the spectral domain is dominated by two lines through the origin. These structures are consistent with the constant leftward and rightward motions present within the analysis window.

Beauchemin and Barron [6] also considered the case of translucent materials. However, the authors focused on a special case consisting of a spatially constant translucent material, as opposed to spatially varying in the analysis above, that results in the following weighted superposition of signals,

$$I_1(\mathbf{x},t) = (1-\alpha)E_1(\mathbf{x}-\mathbf{v}_1) + \alpha I_0(\mathbf{x}-\mathbf{v}_0), \quad (7)$$

where α represents the constant translucency factor. This case is commonly referred to as *additive transparency*. By the superposition property of the Fourier transform, its spectrum consists of the sum of the translational spectra of the individual layers. With the exception of the distortion term, scaling factors and DC component, the spectra for additive and multiplicative transparency are identical. Computational schemes for dealing with the additive transparency case are presented in [26], [31].

B. Occlusion

In this section, the analysis of *dynamic occlusion* given in [30] is extended by enforcing the non-negativity constraint on image signals.

Assume that an image signal, $I_0(\mathbf{x})$, moves with a velocity $\mathbf{v}_0$ behind an opaque surface with transmittance $T_1(\mathbf{x})$ and emission $E_1(\mathbf{x})$ moving with velocity $\mathbf{v}_1$. Unlike the translucency case in Section III-A, the transmittance function is now binary (i.e., $T_1(\mathbf{x}) \in \{0,1\}$). The occlusion relationship between the two surfaces can be modeled as follow,

$$I_1(\mathbf{x},t) = \Big(1 - T_1(\mathbf{x}-\mathbf{v}_1t)\Big)E_1(\mathbf{x}-\mathbf{v}_1t) + T_1(\mathbf{x}-\mathbf{v}_1t)I_0(\mathbf{x}-\mathbf{v}_0t). \quad (8)$$

Next, let the non-negativity constraints be introduced to both the transmittance and emission terms. The transmittance $T_1(\mathbf{x})$ can be reexpressed as the sum of a constant/DC term α and a zero mean signal, $T(\mathbf{x}) = T_1(\mathbf{x}) - \alpha$. While the emission terms, $E_1(\mathbf{x})$ and $I_0(\mathbf{x})$ can be reexpressed as constant/DC terms, β and γ, plus zero mean signals: $E(\mathbf{x}) = E_1(\mathbf{x}) - \beta$ and $I(\mathbf{x}) = I_0(\mathbf{x}) - \gamma$. Introducing these constraints in (8) yields,

$$\begin{aligned} I_1(\mathbf{x},t) = & \Big(1 - \alpha - T(\mathbf{x}-\mathbf{v}_1t)\Big)\Big(\beta + E(\mathbf{x}-\mathbf{v}_1t)\Big) \\ & + \Big(\alpha + T(\mathbf{x}-\mathbf{v}_1t)\Big)\Big(\gamma + I(\mathbf{x}-\mathbf{v}_0t)\Big). \end{aligned} \quad (9)$$

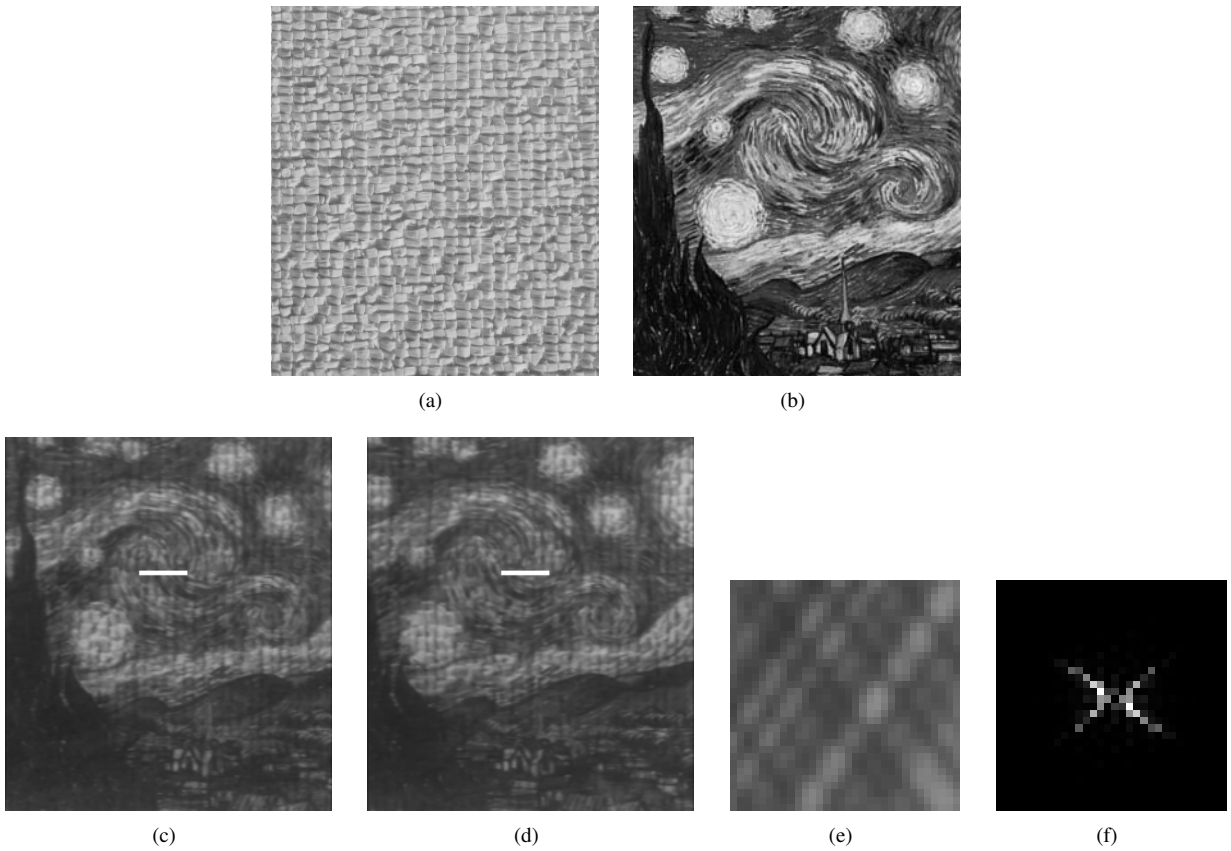


Fig. 2. Real translucency example. (a) and (b) depict the "raffia weave" texture from the Brodatz database [9] and van Gogh's *Starry Night* painting, respectively, used to form the constituent layers of the translucency example. (c) and (d) represent the first and last frames of a 32 frame image sequence depicting the *Starry Night* painting (i.e., opaque surface) moving behind an acetate (i.e, translucent material) depicting the "raffia weave" texture, captured by a stationary video camcorder. The surfaces are moving in opposite directions (leftward and rightward) with approximately equal speed. The movements were generated using computer-controlled translating stages. The white solid lines overlayed for display purposes only in (c) and (d) denote the 32 pixel (horizontal) spatial extent of the analysis window; the temporal extent of the analysis window is 32 frames. (e) The epipolar slice of the sequence along the analysis window; the spatial and temporal axes point rightward and downward, respectively. (f) The 2D power spectrum of the *Kaiser windowed* analysis region; the origin of the spectrum lies in the middle of the image. For display purposes, the DC component has been removed.

The corresponding Fourier transform can be written as,

$$\begin{aligned}\tilde{I}_1(\mathbf{k},\omega_t) = & \\ & \Big((1-\alpha)\beta+\alpha\gamma\Big)\delta(\mathbf{k},\omega_t) \\ & +\alpha\tilde{I}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_0+\omega_t) \\ & +(1-\alpha)\tilde{E}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_1+\omega_t) \\ & +(\gamma-\beta)\tilde{T}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_1+\omega_t) \\ & -\Big(\tilde{T}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_1+\omega_t)\Big)*\Big(\tilde{E}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_1+\omega_t)\Big) \\ & +\Big(\tilde{T}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_1+\omega_t)\Big)*\Big(\tilde{I}(\mathbf{k})\delta(\mathbf{k}^\top\mathbf{v}_0+\omega_t)\Big).\end{aligned} \tag{10}$$

The final step consists of defining a transmittance function. Following [14], [30], the two-dimensional *Heaviside function* (i.e., unit step) is used for the support of the occluder,

$$T_1(\mathbf{x}) = \begin{cases}1, & \mathbf{x}^\top\hat{\mathbf{n}} \geq 0 \\ 0, & \text{otherwise'}\end{cases} \tag{11}$$

where $\hat{\mathbf{n}}$ denotes the unit normal vector to the occluding boundary. Note that the DC term of (11) is given by $\alpha = 1/2$. The assumption of a linear occluding boundary can be justified on the grounds that the region of analysis that straddles the boundary is generally much smaller than the constituent surfaces.

With the occlusion model fully specified, (10), it can be interpreted with broadband signal components. The first term corresponds to a DC component. The second and third terms correspond to the scaled spectral planes of the occluded and occluder signals, respectively. Their normal vectors $(\mathbf{v}_0, 1)^\top$ and $(\mathbf{v}_1, 1)^\top$ denote their respective layer velocities. It is interesting to point out here that in the present derivation both the occluder and occluded signals explicitly appear as separate terms (ignoring scale and bias); whereas, in the original derivation of Eq. (12) in [30] only the occluder signal appears undistorted. The oriented structure of the occluder signal is reinforced by the fourth and fifth terms. Observing that the motion of the Heaviside function is an instance of the *aperture problem* [19], the final term corresponds to a convolution between a 3D line and a 3D plane. This lone term contributes to a distortion from the ideal case of superposition between two oriented planes. As pointed out in [30], the influence of the convolution of the line corresponds to a hyperbolic

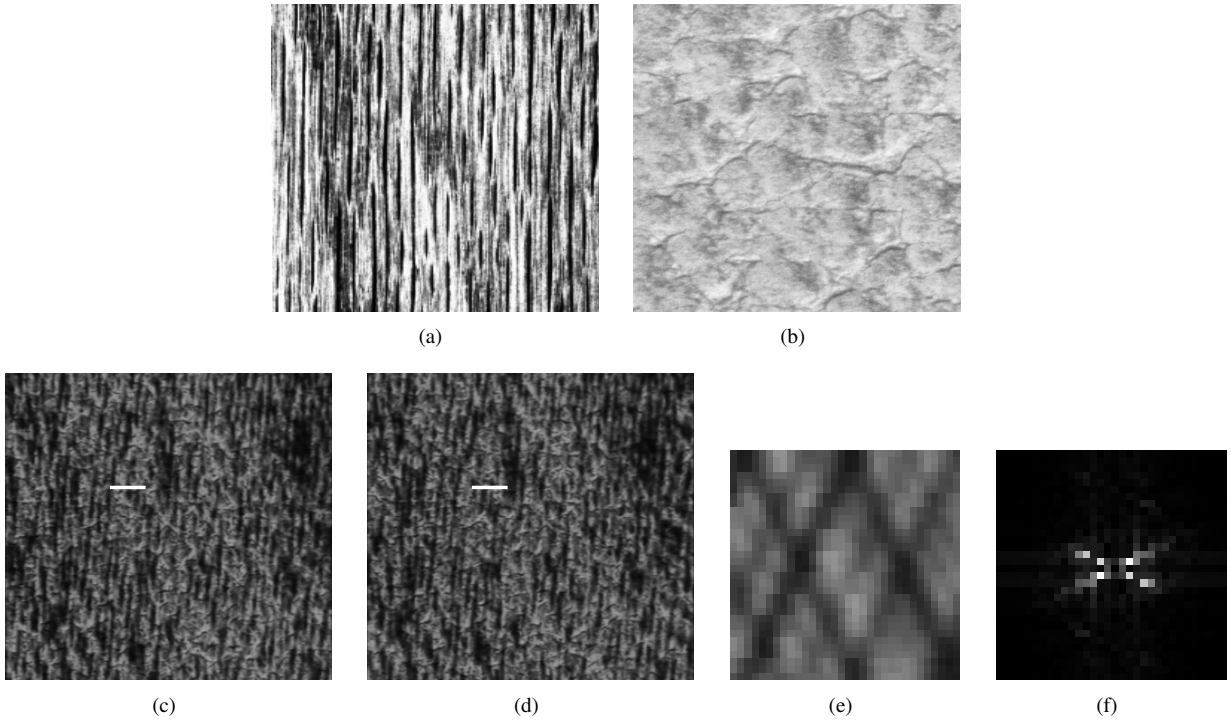


Fig. 3. Real structured light example. (a) and (b) depict the "wood grain" and "pigskin" textures, respectively, from the Brodatz database [9], used to form the constituent layers of the structured light example. (c) and (d) represent the first and last frames of a 32 frame image sequence depicting a moving structured light pattern projected (using an LCD projector) onto an opaque moving surface, captured by a stationary video camcorder. The structured light and opaque surface depict the "wood grain" and "pigskin" textures, respectively. The surfaces are moving in opposite directions (leftward and rightward) with approximately equal speed. The movement of the opaque surface was generated using a computer-controlled translating stage. The white solid lines overlayed for display purposes only in (c) and (d) denote the 32 pixel (horizontal) spatial extent of the analysis window; the temporal extent of the analysis window is 32 frames. (e) The epipolar slice of the sequence along the analysis window; the spatial and temporal axes point rightward and downward, respectively. (f) The 2D power spectrum of the *Kaiser windowed* analysis region; the origin of the spectrum lies in the middle of the image. For display purposes, the DC component has been removed.

distortion that can be assumed negligible as compared to noise. Importantly, the main energy of the spectrum lies on the two spectral planes given by the motion of the occluder and occluded signals.

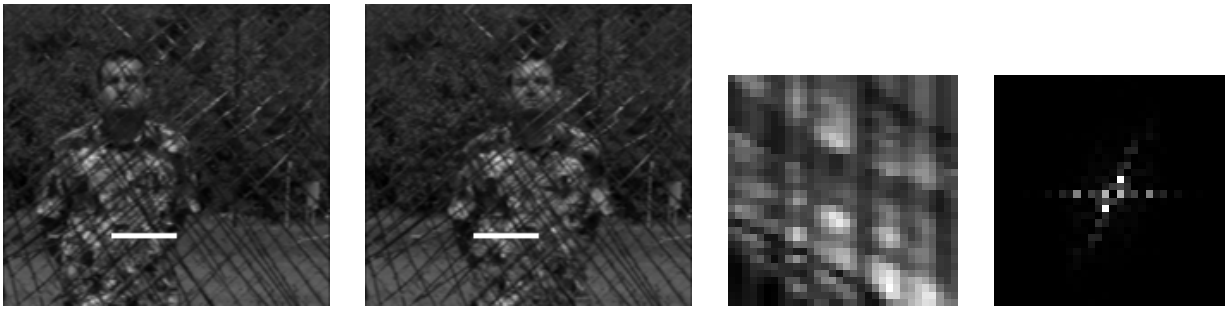
C. Pseudo-transparency

Pseudo-transparency (also commonly referred to as *diaphanous* or gauzy/sheer transparency) can also be accommodated by the model, (10). This spacetime structure corresponds to the case where the holes in a perforated occluder are below the observer's spatial resolution limit [22]. In other words, across the region of concern each analysis window contains both the foreground and background. A prime example of this case in the real-world is viewing a moving object through some fragmented surface, such as a fence, leafless bush, grassy field, etc. Assuming that the binary transmittance function of the occluder is broadband, reflecting its typically "complex" nature, (10) can again be interpreted as two oriented spectral planes through the origin reflecting the velocities of the two surfaces, where the distortion in the last term, as in the case of translucency, corresponds to a non-oriented noise backdrop.

In Fig. 4, a real pseudo-transparency example is presented. This example consists of a person moving rightward behind a stationary chain linked fence, captured by a stationary video camcorder. In the epipolar slice image, diagonal and vertical oriented structures are clearly evident. Correspondingly, the main power in the spectral domain is dominated by two lines through the origin. These structures are consistent with the constant rightward motion and static structures present within the analysis window.

IV. Example application: Multiple motion recovery

As an example application of the theoretical analysis developed in this paper, this section considers the problem of multiple motion recovery with multiplicatively combined image motion signals. In order to recover the multiple motions, the approach of Yu et al. [32] is adapted. This approach originally was proposed in the context of translation, additive transparency and occlusion image motion signals, where component motions were taken as giving rise to corresponding planes in the frequency domain. Accordingly, the basic idea behind the approach is to simultaneous fit a set of planes to the 3D power spectrum of the input image sequence to estimate the component motions. Significantly, previous analyses of multiple motions suggests that such an approach will fail in application to multiplicatively combined signals, as the component oriented structures would have been annihilated. In contrast, the present analysis suggests that such a method can be applied directly to the input signal, as the orientation



(a) (b) (c) (d)

Fig. 4. Real pseudo-transparency example. (a) and (b) represent the first and last frames of a 32 frame image sequence depicting a person moving rightward behind a stationary chain linked fence, captured by a stationary video camcorder. The white horizontal lines overlayed for display purposes only in (a) and (b) denote the 32 pixel (horizontal) spatial extent of the analysis window; the temporal extent of the analysis window is 32 frames. (c) The epipolar slice of the sequence along the analysis window; the spatial and temporal axes point rightward and downward, respectively. (d) The 2D power spectrum of the *Kaiser windowed* analysis region; the origin of the spectrum lies in the middle of the image. For display purposes, the DC component has been removed.

structure is preserved, and that is the basis of the present approach. For the sake of keeping the current paper self-contained, the spectral fitting method is summarized next. The approach breaks down into two parts: First, the input image signal is mapped to the frequency domain and distortions suppressed; second, Expectation-Maximization (EM) [10] is applied to estimate the component velocities.

Mapping to the frequency domain is accomplished via application of a windowed Fourier transform over the spatiotemporal region of interest. This processing is accomplished using a Kaiser windowed Fourier transform, as used elsewhere in the current paper. Next, a 3D low-stop filter is applied to mitigate the effects of distortions at low-frequencies. The frequency response of the low-stop filter is defined as,

$$L(\mathbf{k},\omega_t) = \frac{1}{\alpha + G(\mathbf{k},\omega_t;\mu_0,\sigma_0^2)} - \frac{1}{\alpha + G(0,0,0;\mu_0,\sigma_0^2)}, \tag{12}$$

where $G(\mathbf{k},\omega_t;\mu,\sigma^2)$ denotes the 3D Gaussian in the spectral domain with mean value, $\mu_0 = (0,0,0)^\top$, and variance, $\sigma_0^2 = \pi/16$. The parameter α, which acts as a signal pedestal, is set to 0.1.

With the data so transformed, the second part of the method consists of iterating between an expectation step (E-step) and a maximization step (M-step), until the velocity estimates converge. Assuming that there are two motions in the composition, and beginning with arbitrary initial motion values, $\breve{\mathbf{u}}_1 = (u_1, v_1)^\top$ and $\breve{\mathbf{u}}_2 = (u_2, v_2)^\top$ [1], the E-step assigns weights $w_{i,1}$ and $w_{i,2}$ to the i-th point as follows,

$$w_{i,1} = \frac{1}{1 + e^{-(r_{i,2} - r_{i,1})/\sigma^2}} \tag{13}$$

$$w_{i,2} = \frac{1}{1 + e^{-(r_{i,1} - r_{i,2})/\sigma^2}} \tag{14}$$

where

$$r_{i,1} = a_i^2(\omega_{i,x}u_1 + \omega_{i,y}v_1 + \omega_{i,t})^2 \tag{15}$$

$$r_{i,2} = a_i^2(\omega_{i,x}u_2 + \omega_{i,y}v_2 + \omega_{i,t})^2, \tag{16}$$

and a_i denotes the amplitude of the i-th point in the spectral domain. These weights represent the membership probability for each point.

Given the weights from the E-step, the M-step solves the following two (weighted) linear systems in a least-squares manner,

$$w_{i,1}a_i\omega_{i,x}u_1 + w_{i,1}a_i\omega_{i,y}v_1 + w_{i,1}a_i\omega_{i,t} = 0 \tag{17}$$

$$w_{i,2}a_i\omega_{i,x}u_2 + w_{i,2}a_i\omega_{i,y}v_2 + w_{i,2}a_i\omega_{i,t} = 0. \tag{18}$$

In contrast to the approach suggested in the present paper, i.e., analyzing the spacetime oriented structure directly, some previous research has proposed preprocessing image sequence data with a logarithmic transformation to deal with multiplicative transparency, e.g., [7], [23]. Under such a transformation, the multiplicative composition is changed to an additive one and subsequent processing proceeds much the same as it would for additive transparency (i.e., consideration of multiple dominant orientations). [2] To conclude this section, an empirical comparison is made between analyzing the signal directly (as suggested in this paper) and the logarithmic preprocessing step for multiple motion recovery with multiplicative image signals. For both cases, motion estimates are recovered using the spectral plane fitting method [32], as summarized above.

The first comparison considered a synthetic signal consisting of two (non-negative) white noise signals that have been combined multiplicatively, analogous to the pattern used to generate Fig. 1. The component signals translate with velocities $\mathbf{u}_1 = (1,1)^\top$ and $\mathbf{u}_2 = (1,-1)^\top$. The spatiotemporal support of the input signal was $32 \times 32 \times 32$. For this case, the spectral plane fitting algorithm applied directly to the input signal successfully converged to velocity estimates of $\breve{\mathbf{u}}_1 = (0.924, 0.992)^\top$ and $\breve{\mathbf{u}}_2 = (0.987, -0.973)^\top$ after 6 iterations. In contrast, the logarithmically preprocessed signal converged to incorrect results of $\breve{\mathbf{u}}_1 = (1.118, -0.374)^\top$ and $\breve{\mathbf{u}}_2 = (0.183, 1.054)^\top$. Several additional runs of the algorithm applied to the logarithmically preprocessed signal

[1] In the following, $\breve{}$ is used to distinguish empirically recovered estimates.

[2] Interestingly, Langley [23] asserted a positivity constraint on the component signals of multiplicative motions. This was for the purpose of ensuring that the logarithmic operation was defined. The author did not, however, study the implications of such a constraint on the explicit structure of the signal.

were attempted by randomly varying the initial motion values without success. As a control, the spectral plane fitting method also was run directly on an additive combination of the same component signals used to construct the multiplicative signal. In this case, the algorithm converged to $\breve{\mathbf{u}}_1 = (0.904, 0.995)^\top$ and $\breve{\mathbf{u}}_2 = (0.995, -0.941)^\top$, which is close to both the ground truth and the result of directly processing the input multiplicative transparency signal. Taken together, these results suggest that the spectral plane fitting method is directly applicable to both additive and multiplicatively combined motion signals; however, preprocessing with the logarithmic transformation significantly damages performance.

As a second comparison, the real translucency example in Fig. 2 was considered. A $32 \times 32 \times 32$ spacetime volume around the region marked with the white line overlaid on the figure was used as input. While effort was made to move the component surfaces in opposite horizontal directions with approximately the same speed and maintain minimal vertical motion, no strict ground truth is available; so, only qualitative observations can be made. Here, the spectral plane fitting algorithm applied directly to the input signal converged to the velocities of $\breve{\mathbf{u}}_1 = (-0.619, 0.007)^\top$ and $\breve{\mathbf{u}}_2 = (0.710, -0.005)^\top$, which qualitatively is consistent with the input. In contrast, the logarithmically preprocessed signal erroneously converged to the velocity estimates of $\breve{\mathbf{u}}_1 = (-0.307, 0.048)^\top$ and $\breve{\mathbf{u}}_2 = (0.739, -0.058)^\top$. Again, the spectral plane fitting method was run numerous additional times on the logarithmically preprocessed signal while randomly varying the initial motion values without a change in the converged result. These results further demonstrate practical relevance of the present analysis of multiple motions for application to real imagery.

What is the cause of the relatively poor performance of the logarithmically transformed imagery? An explanation can be had by observing that the logarithmic transformation is compressive and thereby reduces the dynamic range of the imagery to which it is applied. In the case of the natural imagery example, the constituent surface patterns have a relatively small dynamic range even prior to the transformation. After application of the logarithmic transformation, the structure in the power spectrum attributable to the *Starry Night* painting apparently is not reliably discernable from the noise in the input signal; consequently, its motion component is poorly estimated. Similarly, in application to the synthetic imagery, the logarithmic transformation compresses the dynamic range of the signal and neither of the velocities are estimated accurately. Overall, it is seen that the method based on the analysis presented in the present paper is not only simpler than the alternative (it requires no logarithmic preprocessing), it also produces more reliable results.

V. Discussion

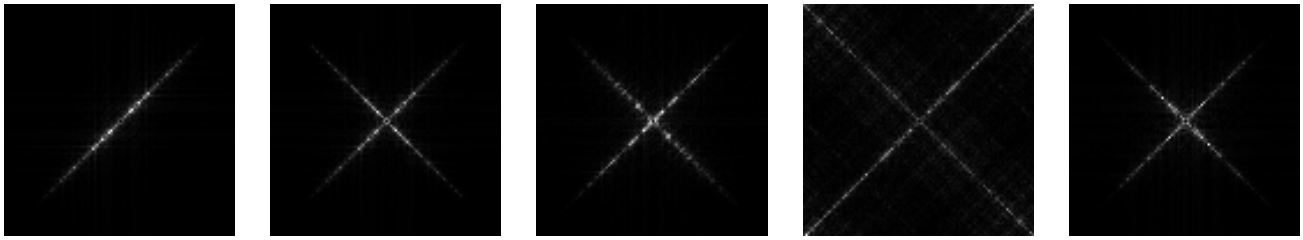
The contributions of this paper are both theoretical and practical. From a theoretical point of view, the analysis shows that five major classes of image motion patterns can be treated in a unified manner simply through consideration of the physical constraint that natural image signals cannot take on negative values. In particular, the cases of translational motion, additive transparency, dynamic occlusion, pseudo-transparency and multiplicative transparency are all characterized by dominant planes through the origin in the spectral domain, where the planes are indicative of the individual component motion patterns (see Fig. 5).

From a practical point of view, with the common structure of various multiple motion patterns revealed, correspondingly unified image processing and inference mechanisms can be developed. Such developments can remove the need for operations that proceed on a case-by-case basis, including potentially complicated integration mechanisms. As an example, a spectral plane fitting mechanism, previously demonstrated with respect to translation, additive transparency and occlusion [32], was generalized in Section IV of this paper to apply to multiplicative multiple motion estimation as well. More generally, the theoretical developments can serve to motivate further techniques for image sequence processing and interpretation based on spatiotemporal orientation measurements, irrespective of whether multiple motions are present or not and irrespective of whether multiple motions are combined additively or multiplicatively. For example, recent techniques for segregating and delineating boundaries between a wide range of juxtaposed spatiotemporal patterns in image sequences based on spacetime orientation measurements has its theoretical foundation in the present analysis of multiple motions [11], [12]. Along these lines, a practical limitation that will enter into the application of such techniques will arise from how fine grained a distinction can be made between multiple orientations in visual spacetime, $I(\mathbf{x}, t)$.

As discussed in Section IV, an alternative to the approach suggested in the present paper is to preprocess the input image sequence with a logarithmic transformation to deal specifically with the case of multiplicative transparency [7], [23]. Such an approach has significant limitations. First, it suggests that multiplicative transparency be dealt with as a special case, including attendant issues of integration with results produced in terms of other motion classes. Second, the logarithmic transformation is compressive and thereby will result in a significant reduction of the signal-to-noise ratio. Indeed, the practical ramification of such reductions were seen in the experiments reported in Section IV. Moreover, the analysis presented in this paper shows that such a transformation of the image data is not needed, as the physical nature of the signal already ensures that the data is amenable to a uniform treatment in terms of orientation processing for an important range of image motion patterns: single translation, additive transparency, dynamic occlusion, pseudo-transparency and multiplicative transparency.

Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors thank M. Spetsakis and the anonymous reviewers for their helpful feedback.



(a) (b) (c) (d) (e)

Fig. 5. Comparison of 2D spectra of various spacetime structures with white noise sequences. Each of the patterns are windowed with a *Kaiser window* prior to applying the Fourier transform. Each of the component layers move with a speed of 1 pixel/frame; motion is in opposite directions in the case of multiple components. (a)-(e) The magnitude spectra for: (a) translational motion, (b) additive transparency, (c) dynamic occlusion, (d) pseudo-transparency and (e) multiplicative transparency; the origin of the spectrum lies in the middle of the image. The pseudo-transparency case was realized using white noise patterns for the emission terms and a numerically rounded low-pass white noise pattern for the transmittance factor. Each of these spacetime structures are characterized by oriented lines passing through the origin, where their orientation reflects the speed and direction of motion of the constituent layers. For display purposes, the DC components have been removed.

REFERENCES

[1] E.H. Adelson and P. Anandan. Ordinal characteristics of transparency. In *AAAI Workshop on Qualitative Vision*, pages 77–81, 1990.

[2] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America - A*, 2(2):284–299, February 1985.

[3] I. Austvoll. A study of the Yosemite sequence used as a test sequence for estimation of optical flow. In *Scandinavian Conference on Image Analysis*, 2005.

[4] S. Baker, S. Roth, D. Scharstein, M.J. Black, J.P. Lewis, and R. Szeliski. A database and evaluation methodology for optical flow. In *IEEE International Conference on Computer Vision*, 2007.

[5] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, February 1994.

[6] S.S. Beauchemin and J.L. Barron. The frequency structure of 1D occluding image signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):200–206, February 2000.

[7] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(19):886–896, September 1992.

[8] R.N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, New York, 2000.

[9] P. Brodatz. *Textures*. Dover, New York, 1966.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society-B*, 39(1):1–38, 1977.

[11] K.G. Derpanis and R.P. Wildes. Detecting spatiotemporal structure boundaries: Beyond motion discontinuities. In *Asian Conference on Computer Vision*, 2009.

[12] K.G. Derpanis and R.P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[13] M. Fahle and T. Poggio. Visual hyperacuity: Spatio-temporal interpolation in human vision. *Proceedings of the Royal Society of London - B*, 213:451–477, November 1981.

[14] D.J. Fleet. *Measurement of Image Velocity*. Kluwer, Norwell, 1992.

[15] D.J. Fleet and A.D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, August 1990.

[16] D.J. Fleet and K. Langley. Computational analysis of non-Fourier motion. *Vision Research*, 34(22):3057–3079, November 1994.

[17] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978.

[18] D.J. Heeger. Model for the extraction of image flow. *Journal of the Optical Society of America - A*, 2(2):1455–1471, August 1987.

[19] B.K.P. Horn. *Robot Vision*. MIT Press, Cambridge, 1986.

[20] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, August 1981.

[21] A.D. Jepson and M.J. Black. Mixture models for optical flow computation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–761, 1993.

[22] D. Kersten. Transparency and the cooperative computation of scene attributes. In M. Landy and J.A. Movshon, editors, *Computational Models of Visual Processing*, pages 209–228. Cambridge, MA: MIT Press, 1991.

[23] K. Langley. Computational models of coherent and transparency plaid motion. *Vision Research*, 39:87–108, January 1998.

[24] C. Liu, W.T. Freeman, E.H. Adelson, and Y. Weiss. Human-assisted motion annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[25] P. Meer, D. Mintz, D.Y. Kim, and A. Rosenfeld. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70, April 1991.

[26] M. Shizawa and K. Mase. Principle of superposition: A common computational framework for analysis of multiple motion. In *Motion Workshop*, pages 164–172, 1991.

[27] E.P. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, May 2001.

[28] R.S. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 246–253, 2000.

[29] A.B. Watson and A.J. Ahumada. Model of human visual-motion sensing. *Journal of the Optical Society of America-A*, 2(2):322–342, February 1985.

[30] W. Yu, G. Sommer, S. Beauchemin, and K. Daniilidis. Oriented structure of the occlusion distortion: Is it reliable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1286–1290, September 2002.

[31] W. Yu, G. Sommer, and K. Daniilidis. Multiple motion analysis: In spatial or in spectral domain? *Computer Vision and Image Understanding*, 90(2):129–152, May 2003.

[32] W.C. Yu, K. Daniilidis, S. Beauchemin, and G. Sommer. Detection and characterization of multiple motion points. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 171–177, 1999.