COMPUTER VISION, GRAPHICS, AND IMAGE PROCESSING 45, 227–250 (1989)

Motion Detection in Spatio-Temporal Space*

SHIH-PING LIOU AND RAMESH C. JAIN

Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109

Received August 13, 1987; revised August 22, 1988

We present an analysis of existing motion detectors for determining desirable characteristics of a motion detector. A spatio-temporal surface type inseparable model is then proposed for motion detection. Based on this model, we analyzed mathematically how the geometry of the intensity hypersurface gives information about motion in image. The local motion information, obtained from the parameters of the Monge patch approximating the intensity hypersurface in the spatio-temporal space, may be used for segmentation of dynamic scenes. Motion detection results for real sequences show the robustness of this detector. © 1989 Academic Press, Inc.

1. INTRODUCTION

When there is a relative motion between the observer and some object in the environment, a changing pattern of light falls upon the retina (projection surface). The resulting *optical flow* field, perceived due to the temporal variation in the brightness pattern, carries rich information not only about the motion but also about the 3-dimensional structure of the scene. The recovery of 3D structure and motion from optical flow on a projection surface is one of the most challenging tasks that computer vision research is confronted with. To achieve such goals, an accurate *image flow*, the projection of the 3D instantaneous velocity field on the projection surface, is always required. However, in some cases, it is sufficient to discover only certain properties of the image flow field rather than to measure it completely. For example, it might be desirable to quickly respond to a moving object. In such case, motion must be detected but not necessarily measured. This motion detection problem has recently attracted several researchers in psycho-physics [1, 5, 6, 8–10, 13, 18, 19, 22, 25] and computer vision [11, 15].

The conventional method in motion analysis is first to treat the images as 2D signals sampled at discrete times and then to compute motion from the relations among the image reflectance model, the types of motion, the structure of the environment, and the change of intensity values. Most of these approaches use just two or three frames of a sequence, disregarding the information about the motion of objects in dynamic scenes. Even if a longer sequence is used, the spatial information in many cases, due to the existence of unmodelled camera rotation, is more noise-sensitive than the temporal information. These aspects motivate our research on the theory and the scheme of spatio-temporal approaches.

From a spatio-temporal point of view, time-varying stimuli may be pictured as occupying a 3-dimensional space, in which x and y are two spatial dimensions and t is the temporal dimension. Considering an edge moving through space and time as

*We gratefully acknowledge the support of National Science Foundation Grant DCR8517251.

0734-189X/89 $3.00

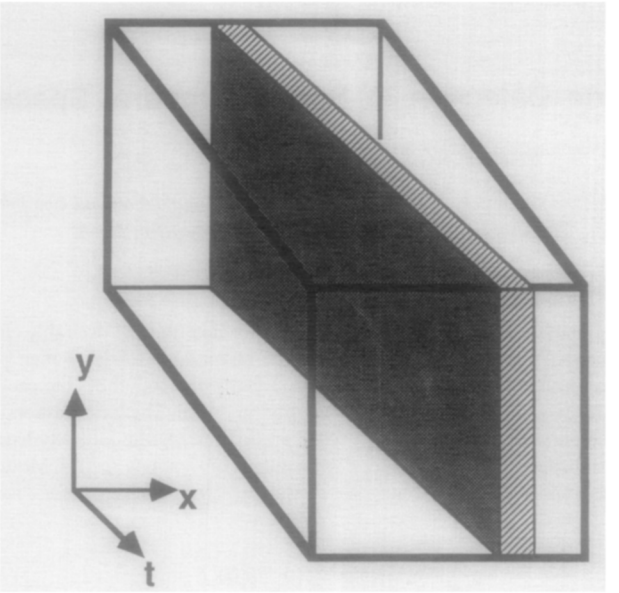

FIG. 1. The 3D spatio-temporal solid.

shown in Fig. 1, its trajectory will sweep out a 2D surface in space-time. The first-order discontinuities of this spatio-temporal solid directly form this 2D surface and the orientation of its corresponding tangent plane at every point indicates the speed and the direction of motion at that point. Therefore, the problem of detecting motion can be thought of as a problem of detecting the orientation of tangent planes in spatio-temporal space [1, 11, 22, 25, 28]. Recently, Baker, Bolles, and Marimont [7] explore a technique for detecting the 3D zeros of the Laplacian of a chosen 3D Gaussian in the spatio-temporal space to recover the structure of a known camera motion. However, this scheme is restricted to a linear camera motion and will fail in the scene which contains moving objects.

This paper considers the image irradiance function as a 3-dimensional scalar field and the intensity surface as a Monge patch in the neighborhood of a spatio-temporal point. The parameters of the Monge patch are then used for detection of motion. Haralick, Lee, and Joo [12] propose a facet approach to optical flow in which not only locally conserved intensity along the path of a trajectory but constant image velocity over small periods of time are assumed. The mathematical analysis we will use, however, does not involve any of these assumptions.

The rest of this paper is organized as follows. In Section 2, we review some of the earlier spatio-temporal motion detectors. We present a spatio-temporal surface model for analyzing the motion detection problem in Section 3. This analysis reaches a conclusion similar to [24]. Section 4 shows that a robust motion detector should be based on the spatio-temporal surface model. A motion detector is proposed in Section 5 and its robustness is shown by considering a few motion sequences in Section 6. Finally, the conclusion is drawn in Section 7. A very brief review of relevant concepts of hyperspaces is given in an Appendix for ready reference.

2. SPATIO-TEMPORAL APPROACHES

The spatio-temporal filter-type models for motion detection have received remarkably widespread use in biological and machine studies of early visual processing. In biological studies they are used as qualitative behavioral models to accommodate data from both psychophysical studies and single-cell electrophysiological recordings on the retina, lateral geniculate nucleus, and visual cortex. In computer vision, they have received extensive use as an initial stage of spatio-temporal filtering in approaches to motion detection and a multiple channel band-pass representation for dynamic scenes.

The essential concept of this type of models is explored by Watson and Ahumada [25]. Suppose an arbitrary monochromatic space-time image is represented as a function $c(x, y, t)$ over some interval which specifies the contrast at each point x, y and time t. Its Fourier transform is denoted by $\tilde{c}(u, v, w)$. Under the translation at constant velocity $\mathbf{r} = (r_x, r_y)$, its transform is then

$$c(x - r_x t, y - r_y t, t) \rightarrow_3 \tilde{c}(u, v, w + r_x u + r_y v),$$

where $\rightarrow_3$ indicates the 3D Fourier transform. Geometrically, image motion changes the static-image transform, which lies in the u, v plane, into a spectrum that lies in an oblique plane through the origin. This property is called *temporal modularity*, which will play an important role in later discussions.

Barlow and Levick [3] suggest that neurons in the rabbit retina have directional selectivity and they work as shown in Fig. 2a. In their model, receptors A and B sample adjacent retinal regions. B's output is delayed and subtracted from A's

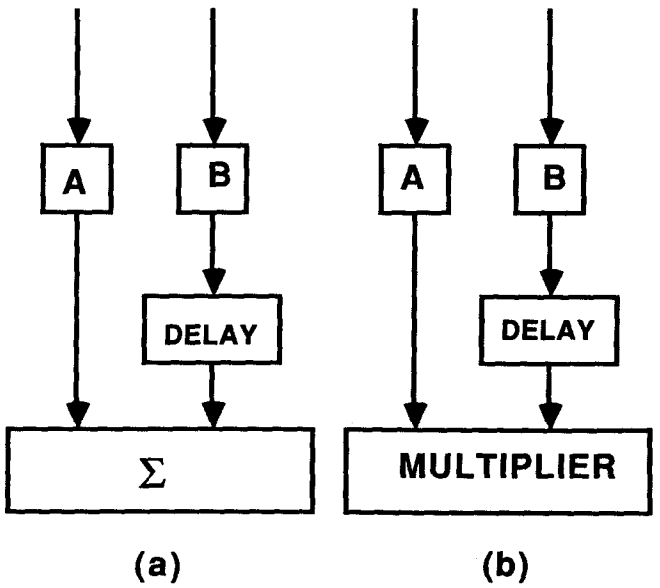

FIG. 2. (a) Barlow and Levick's model. (b) Reichardt's model.

output. Reichardt [19] proposes a model of motion detection by the beetle's eye, in which B's delay output is multiplied by A's output (see Fig. 2b). In this case, the multiplication is used as an operator to measure correlation.

Van Saten and Sperling [22] present an elaboration of Reichardt's model in which a local correlation is performed across space and time. This model consists of two subunits tuned to opposite directions, each of which performs a spatial and temporal linear filtering. The outputs of the filters are multiplied and then integrated. Their detectors are designed so that those sensitive to high temporal frequencies are less sensitive to high spatial frequencies and vice versa. This model does not attempt to preserve temporal modularity. Therefore, it is quite susceptible to variations in the contrast of image components at different orientations and directions.

Adelson and Bergen [1] propose an energy model. They formulate the problem of detecting motion as the problem of detecting orientation in space-time. The motion energy is computed by integrating over time the outputs of a set of linear filters tuned in spatial frequency. The output from this model is shown to be the same as that from Reichardt-type model, except for the scale factor.

Watson and Ahumada [25] propose a model of human visual-motion sensing, in which they preserve the temporal modularity and note that it directly codes the image-velocity components. At the first stage of their model there is a set of spatial-frequency-tuned, direction-selective linear sensors. The second stage refers to a process in which these components are resolved to measure the velocity of image motion at each of spatial locations and spatial frequencies. A number of interesting results show the qualitative agreement with human perception.

Fleet [9] and Fleet *et al.* [10] present the center-surround (CS) model, an extension of the spatial DOG (difference of Gaussian) model to include time-dependent behavior. The first interesting aspect of this model is its inseparable spatio-temporal behavior. Second, this model is designed to have simple and desirable band-pass signal characteristics that might lead to a representation for visual information based on local-frequency analysis. The behavior of the CS model is shown in close agreement with the data from a variety of neurophysiological and perceptual experiments.

Recently, similar to the scheme in [25], Heeger [13] presents a model for computing ideal image velocity. This model uses 3D (space-time) Gabor filters to sample the power spectrum and, by combining the outputs of several such filters, estimates the image velocity field. Experiments are performed on a wide variety of real images as well as sine-grating plaid patterns. The results show the ability of this model in dealing with the aperture problem.

In contrast to the above biological models, there is another type of model based on the Marr–Hildreth theory of edge detection [17]. Marr and Ullman [18] propose a motion detector in which they computed the time derivative of the Laplacian of Gaussian. They apply this scheme to several real images and demonstrate its directional selectivity. Although physiological support is claimed, this detector still produces spurious motion of the stationary background. Buxton and Buxton [5, 6] extend the Marr–Hildreth theory of edge detection to design a spatio-temporal filter and speculate on the possibility of extracting depth information from the edges tracked over time. They also discuss the effects associated with the choice of metric in their spatio-temporal filter. Their scheme is applied to both simulated and

artificial data. The results show that image flow can be accurately computed only from moving edge features in an image sequence.

3. PROBLEM ANALYSIS

The conventional approach adopts either the assumption of constant brightness [14] or the assumption of constant velocity [25] in approaching the motion detection problem. Here, instead of making assumptions to simplify the problem, this section discusses how the geometry of the intensity hypersurface gives information about motion in image.

3.1. *A Spatio-Temporal Surface Model*

Since a digital image is a discrete sampling of a continuous function of 2D spatial variables, the underlying continuous intensity surface in the neighborhood of a point can always be constructed as the best approximation to the intensity values. Considering a spatio-temporal solid, we can also obtain the underlying continuous 3D intensity hypersurface in the spatio-temporal neighborhood of a point whenever the sampling rate with respect to the degree of fitting is reasonable. Therefore, we consider the image irradiance function denoted by $E(x, y, t)$ as a scalar field

$$E: \mathbf{U} \to \mathbf{R} \qquad \mathbf{U} \subseteq \mathbf{R}^3$$

with at least first-order continuous partial derivatives in any 3-ball $B(\mathbf{a}), \mathbf{a} \in \mathbf{U}$. In the neighborhood of a spatio-temporal point, the intensity surface can then be considered as a Monge patch (graph surface),

$$\mathbf{X}(x, y, t) = (x, y, t, E(x, y, t))$$

which is an obvious 3-surface in $\mathbf{R}^4$. To simplify the notation, the unit vector in the direction of $\mathbf{x}$ in $\mathbf{R}^{n+1}$ will be denoted by $\hat{\mathbf{x}}$.

3.2. *Analysis*

To understand how the geometry of the intensity hypersurface (3-surface) gives information about motion in image, let us consider a spatial point whose trajectory on the intensity hypersurface is represented by a parametrized 3-surface curve in $\mathbf{R}^4$ as

$$\alpha(t) = (x(t), y(t), t, E(x, y, t)). \tag{1}$$

According to Eq. (18) in the Appendix, its velocity vector at time t is

$$\mathbf{T}(t) = (\dot{x}, \dot{y}, 1, E_x\dot{x} + E_y\dot{y} + E_t). \tag{2}$$

Note that a spatio-temporal trajectory is coded either by α_{x-y-t} or by $\mathbf{T}_{x-y-t}$,[1] and both of them are unknown. Therefore, it is desirable having constraints on either α_{x-y-t} or $\mathbf{T}_{x-y-t}$. For simplicity, we begin the analysis with a case where the

[1] V_{x-y-t} indicates the projected vector of V onto the spatio-temporal $x - y - t$ subspace, a space formed by the unit vectors along the x, y, t axes.

image motion is in only one spatial direction (x direction) and then extend the model to general x-y motion.

3.2.1. *ID Case.* At any arbitrary surface point (x_0, t_0), the unit surface normal is

$$\hat{\mathbf{N}}(x_0, t_0) = \frac{1}{\sqrt{1 + E_x^2 + E_t^2}}(-E_x, -E_t, 1). \tag{3}$$

Now, since the curve of the trajectory of any ID point can be parametrized by t as

$$\alpha(t) = (x(t), t, E(x, t)), \tag{4}$$

its velocity vector at time t is then

$$\mathbf{T}(t) = (\dot{x}, 1, E_x\dot{x} + E_t). \tag{5}$$

Without any further knowledge, $\hat{\mathbf{N}} \cdot \hat{\mathbf{T}} = 0$ is the only local information. Now, consider the projected vector of both $\alpha(t)$ and $\mathbf{T}(t)$ onto the spatio-temporal plane. If they are denoted by $\alpha_{x-t}(t)$ separately, $\mathbf{T}_{x-t}(t)$ is still tangent to this plane curve $\alpha_{x-t}(t)$ at every point. Thus, the direction of this velocity vector $\mathbf{T}_{x-t}$ can be estimated by finding the orthogonal direction of the projected unit surface normal $\mathbf{N}_{x-t}$ as shown in Fig. 3.[2] However, it is well known from analytic geometry that $\mathbf{N}_{x-t}$ and $\mathbf{T}_{x-t}$ will not be, in general, orthogonal to each other, unless either the unit surface normal or the tangent vector is parallel to the $x - t$ plane. In other words, $\hat{\mathbf{T}}_{x-t}$ can be correctly estimated only if the angle between $(0, 0, 1)$ and either $\hat{\mathbf{N}}(x, t)$ or $\hat{\mathbf{T}}(t)$ is close to $\pi/2$. In the first case, this is related to the places of step edges; in the second case this condition represents the assumption of constant brightness [14].

3.2.2. *2D Case.* In the general 2D motion case, the unit normal at an arbitrary surface point (x_0, y_0, t_0) is

$$\hat{\mathbf{N}}(x_0, y_0, t_0) = \frac{1}{\sqrt{1 + E_x^2 + E_y^2 + E_t^2}}(-E_x, -E_y, -E_t, 1). \tag{6}$$

These two projected vectors of $\mathbf{N}$ and $\mathbf{T}$ onto the $x - y - t$ subspace after normalization, denoted by $\hat{\mathbf{N}}_{x-y-t}$ and $\hat{\mathbf{T}}_{x-y-t}$ separately, are

$$\hat{\mathbf{N}}(x_0, y_0, t_0)_{x-y-t} = \frac{1}{\sqrt{E_x^2 + E_y^2 + E_t^2}}(-E_x, -E_y, -E_t, 0) \tag{7}$$

and

$$\hat{\mathbf{T}}(t_0)_{x-y-t} = \frac{1}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}}(\dot{x}, \dot{y}, 1, 0). \tag{8}$$

[2] For other sinusoid wave, please refer to [23] for details.

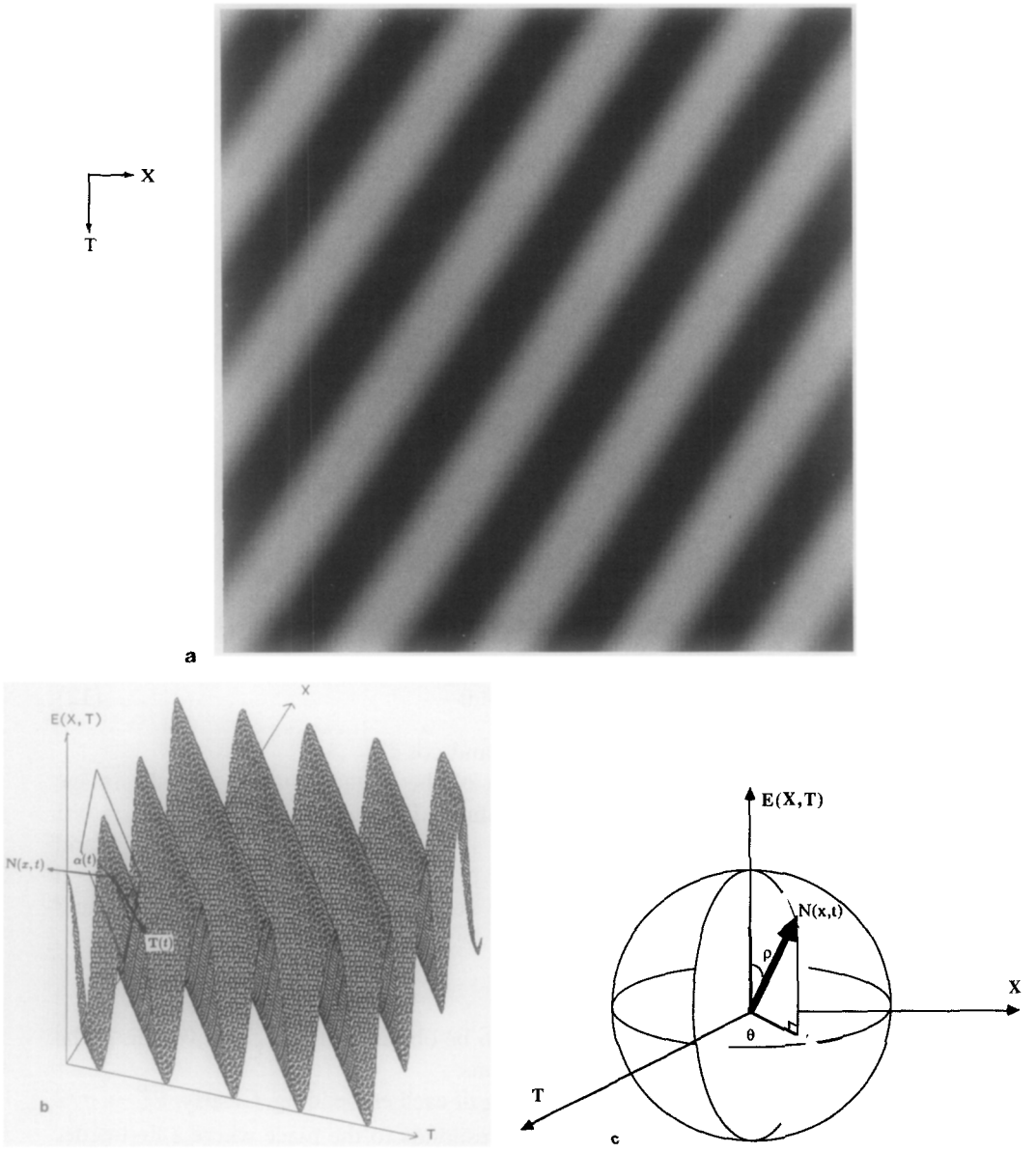

FIG. 3. (a) A left-drifting sinusoid, $E(x, t) = 127.0 + 127.0\cos(2\pi f_0 x + 2\pi w_0\, dt)$ with $f_0 = w_0 = 0.024$ and $d = 1.5$. (b) A perspective plot of (a). (c) Geometric representations of the angles.

In order to measure how *parallel* a vector is to a subspace, we define the *embedding angle* $\theta_{\mathbf{v}}^{\mathscr{R}}$ of a vector $\mathbf{v}$ with respect to a n-dimensional subspace $\mathscr{R}$ in $\mathbf{R}^{n+1}$ as

$$\cos^{-1}\frac{\mathbf{v}\cdot\mathbf{N}_{\mathscr{R}}}{\|\mathbf{v}\|}.$$

where $\mathbf{N}_{\mathscr{R}}$ is the unit vector normal to the n basis of $\mathscr{R}$.

If we denote the $x - y - t$ subspace by $\mathcal{U}$, then

$$\cos \theta_{\mathbf{N}}^{\mathcal{U}} = \frac{1}{\sqrt{1 + E_x^2 + E_y^2 + E_t^2}}, \qquad 0 \le \theta_{\mathbf{N}}^{\mathcal{U}} < \frac{\pi}{2}, \tag{9}$$

$$\cos \theta_{\mathbf{T}}^{\mathcal{U}} = \frac{E_x\dot{x} + E_y\dot{y} + E_t}{\sqrt{1 + \dot{x}^2 + \dot{y}^2 + \left(E_x\dot{x} + E_y\dot{y} + E_t\right)^2}}, \qquad 0 < \theta_{\mathbf{T}}^{\mathcal{U}} < \pi. \tag{10}$$

Similar to the discussion in the previous section, we can also estimate the direction of $\mathbf{T}_{x-y-t}$ by finding the orthogonal direction of $\mathbf{N}_{x-y-t}$. Note that how close the estimated direction is to the direction of $\mathbf{T}_{x-y-t}$ depends on how small the right-hand side of the following equation is

$$\hat{\mathbf{N}}_{x-y-t} \cdot \hat{\mathbf{T}}_{x-y-t} = -\cos \theta_{\mathbf{N}}^{\mathcal{U}} \sqrt{1 + \frac{1}{E_x^2 + E_y^2 + E_t^2}} \times \frac{1}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}} \left(E_x\dot{x} + E_y\dot{y} + E_t\right). \tag{11}$$

This equation is derived directly from

$$\hat{\mathbf{N}} \cdot \hat{\mathbf{T}} = 0 \tag{12}$$

by rewriting it in terms of $\hat{\mathbf{N}}_{x-y-t} \cdot \hat{\mathbf{T}}_{x-y-t}$ and $\cos \theta_{\mathbf{N}}^{\mathcal{U}}$.

In real images, all the partial derivatives of the image function and the image velocities inside the region of an object are finite. Therefore, $\hat{\mathbf{N}}_{x-y-t} \cdot \hat{\mathbf{T}}_{x-y-t} \approx 0$ if and only if either $\theta_{\mathbf{N}}^{\mathcal{U}} \approx \pi/2$ or $E_x\dot{x} + E_y\dot{y} + E_t \approx 0$. In other words, $\hat{\mathbf{N}}_{x-y-t}$ and $\hat{\mathbf{T}}_{x-y-t}$ are orthogonal if and only if either $\hat{\mathbf{N}}$ or $\hat{\mathbf{T}}$ is embedded in the $x - y - t$ subspace. By *embedding*, we mean that the vector can be represented by a linear combination of the basis of $\mathcal{U}$.

3.3. *Conclusion*

First, note that no motion information can be obtained if $\mathbf{N}_{x-y-t} = \mathbf{0}$. This is the case when an image has uniform illuminations.

Second, let us discuss the physical meaning of each embedding. Clearly, $\theta_{\mathbf{N}}^{\mathcal{U}} \to \pi/2$ if and only if $\sqrt{E_x^2 + E_y^2 + E_t^2} \to \infty$. It corresponds to the place where a first-order discontinuity occurs in the spatio-temporal solid. This situation could never happen because the solid is not constructed that way. However, the relation between $\theta_{\mathbf{N}}^{\mathcal{U}}$ (embedding angle) and $\sqrt{E_x^2 + E_y^2 + E_t^2}$ (gradient magnitude) in Eq. (9) states that there exists quite good approximation even with small gradient magnitude. In 2D spatial image, an edge corresponds to local discontinuities of various order in the intensity surface of a scene. For instance, a *step edge* is a first-order discontinuity. From now on, we will use the term *step hyperedge* to mean a place where $\theta_{\mathbf{N}}^{\mathcal{U}} \approx \pi/2$ in the spatio-temporal solid.

As $\theta_{\mathbf{T}}^{\mathcal{U}} \to \pi/2$, the fourth term of $\mathbf{T}(t)$ approaches 0. This is the case when the intensity value is locally conserved along the path of a trajectory of a spatial point, which is the same as the assumption of constant brightness [20].

TABLE 1
Four classes of Motion Detectors

Types	Filter	Surface
Separable	[1, 19, 22]	[11, 14]
Inseparable	[9, 10, 13, 25]	Desirable

It is noteworthy that this formulation and the formulation in [24] show essentially the same conclusion in different ways. However, our formulation is much simpler and broader in scope because there are no additional assumptions made in our model either for how image is formed or for how motion is conducted.

4. CRITERIA FOR A GOOD MOTION DETECTOR

In this section, we shall discuss motion detection from the point of view of surface model vs filter model as well as separability vs inseparability. We conclude that surface type motion detectors with inseparable spatio-temporal behaviors are desirable. However, this design strategy has not been used by any of the researchers in the literature (see Table 1).

4.1. *Surface Model vs Filter Model*

The filter type models have been used in a variety of motion detectors to describe various cell types and seem to receive both quantitative and qualitative supports from psychophysics as well as neurophysiology. Unfortunately, this type of model suffers from several fundamental problems. First, almost all filter type models use the assumption of constant velocity [25] which is not general enough.

Second, this type of model has the *scale space* problem, a problem of separating events at different scales [26]. Filters should possess the ability to localize spatial/temporal events. This criterion excludes the use of a global frequency analysis of images to detect image features and then introduces the sense of a tuned filter to respond to different frequency-band events. Thus, detectors need to be tuned to different spatio-temporal orientations but various scales of resolutions as well. The discussion on how a chosen scale affects a spatio-temporal frequency passband can be found in [25]. On the other hand, surface type models can avoid the scale space problem by using a global analysis like the variable-order surface fitting method in [4].

Third, the filter type approaches are suitable for analysis of steady state behavior of dynamic systems, but not for the analysis of transient behavior. The importance of the transient behavior led to a complete turnaround in the techniques used by control system analysts. In computer vision, it is rare to find situations where image sequences can be satisfactorily analyzed using filter-type techniques. The filters that give excellent performance for regular uniform motion, such as motion of sinusoidal gratings, will not be adequate for the analysis of abruptly changing motion of objects in real scenes.

4.2. *Separability vs Inseparability*

The earliest support for the inseparability of retinal mechanisms comes from several studies in the early 1960s, while the actual dependence of spatial frequency sensitivity on temporal frequency was first suggested by Enroth-Cugell and Robson [8]. They observed that at low spatial frequencies the cells have band-pass temporal characteristics, yet at higher spatial frequencies they have low-pass temporal characteristics.

A filter $H(\mathbf{x}, t)$ is called *separable* if it can be expressed as the product of a purely spatial part, $S(\mathbf{x})$ and a purely temporal part, $T(t)$ as

$$H(\mathbf{x}, t) = S(\mathbf{x})T(t).$$

If this property is not satisfied, the filter is said to be *inseparable*.

The notion of a spatio-temporal filter with separable/inseparable spatio-temporal behavior can be extended to the analysis of the spatio-temporal surface type motion detectors. We call a surface type motion detector $\mathscr{H}(\nabla_{xy}, \nabla_t)$ *separable* if it can be expressed as the product of a purely spatial part, $\mathscr{S}(\nabla_{xy})$ and a purely temporal part, $\mathscr{T}(\nabla_t)$ as

$$\mathscr{H}(\nabla_{xy}, \nabla_t) = \mathscr{S}(\nabla_{xy})\mathscr{T}(\nabla_t),$$

where ∇_{xy} and ∇_t are the terms containing the spatial and temporal derivatives of any order. Otherwise, this detector is said to be *inseparable*.

For instance, the *difference technique* [15], based on the detection of significant temporal changes, is an example of separable surface type detectors. Another example is the time-varying edge detector proposed by Haynes and Jain [11] because the time-varying edginess was defined as

$$\text{moving\_edginess} = \sqrt{E_x^2 + E_y^2} \times |E_t|, \tag{13}$$

where E_x, E_y, and E_t are the spatial and temporal derivatives at the point (x, y) in the frame of the sequence at time t. Finally, the magnitude of estimated normal velocity, revealing the minimum image velocity at a point, can also measure the amount of motion. However, since these techniques treat space and time differently, they cause the difficulty in detecting real motion if the frames are not correctly registered.

Similar problems resulting from detectors with separable spatio-temporal behaviors were also reported from filter-type approaches [1, 22]. This set of motion detector, whether filter type or surface type, respond strongly to places where the frequencies or gradients are in a certain range; therefore, they are not suitable for detecting motion in general. The filter-type detectors with separable behavior cannot preserve temporal modularity, while surface-type detectors with separable behavior cannot discriminate rapidly intensity-varying pixels from slowly moving edges.

5. PROPOSED MOTION DETECTOR

If motion is detected locally (small compared to the overall contour), the only information that can be extracted is the estimated normal velocity, the motion component perpendicular to the estimated local orientation of the element. As we increase the size of the local window (for example, at corner point), further

information about the actual image velocity might be obtained. We have discussed that an estimated normal velocity approaches a true normal velocity only at step hyperedges. Besides, such computation usually falls apart in regions having uniform illuminations.

To resolve the problem resulting from the use of the magnitude of estimated normal velocity as motion measurement, let us first consider motion detection at step edges in the 1D motion case. Note that at step edges the image gradient vector (an approximation of the projected intensity surface normal) can be considered parallel to the $x - t$ plane. The angle between this vector and the unit vector along the t axis, therefore, is very good for measuring the amount of 1D motion. Similarly, in 2D motion case, the angle between the image gradient vector and the unit vector along the t axis can be used for measuring the amount of 2D motion. For convenience, this angle will be referred to the *temporal direction* throughout the rest of this paper.

There are also some mathematical motivations for using the magnitude of temporal direction instead of the magnitude of estimated normal velocity as a motion measure. First, this angle together with the spatial orientation uniquely represents the direction of the associated tangent plane and thus is a very useful source of information for further processing. Second, the output from this measure falls into the range $(-\pi/2, \pi/2)$. Therefore, it resolves the problem of having unbounded range resulting from the use of magnitude of normal velocity. Based on the above, we suggest the following steps:

1. Measure the gradient E_x, E_y, E_t.
2. For each frame, compute the gradient magnitude M and the three angles[3] defined by
   - $\sin^{-1}(E_t/M)$ (temporal direction θ)
   - $\sin^{-1}(E_x/\sqrt{E_x^2 + E_y^2})$ (spatial direction ϕ)
   - $\cos^{-1}(1/\sqrt{1 + M^2})$ (embedding angle ρ).
3. Report θ to measure motion, but only at places where the *confidence measure*, ρ, is close to 90°.

Note that the above steps report essentially the same information as any 3D edge operator does. However, two distinctive differences should be pointed out. First, since the local motion information is reliable only at step hyperedges, i.e., the place where its embedding angle ρ is close to 90° (as shown in Eq. (11)), the information at rest of image places is useless. Second, the magnitude of the temporal direction θ is proposed to measure motion. In addition to the advantages we discussed earlier, this measure also has the advantage of having an inseparable spatio-temporal behavior; therefore, is a desirable motion detector according to our criteria.

6. EXPERIMENTAL RESULTS

The following sequences of dynamic scenes have been used to test our proposed motion detector:

1. ROAD1 sequence. An ideal outdoor scene is shown in this sequence, which was taken by a stationary camera on one side of the road. A car is turning right in

[3]To see the physical meanings of these angles, please refer to the 2D case as shown in Fig. 3.

FIG. 4. Frame 1 of ROAD1 sequence.

FIG. 5. Frame 5 of ROAD1 sequence.

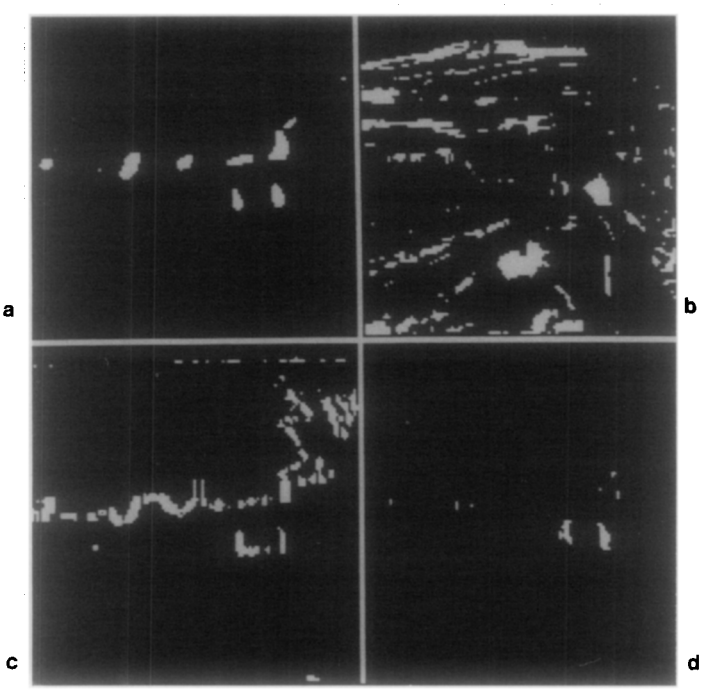

FIG. 6. Results of four motion detectors, ROAD1 sequence.

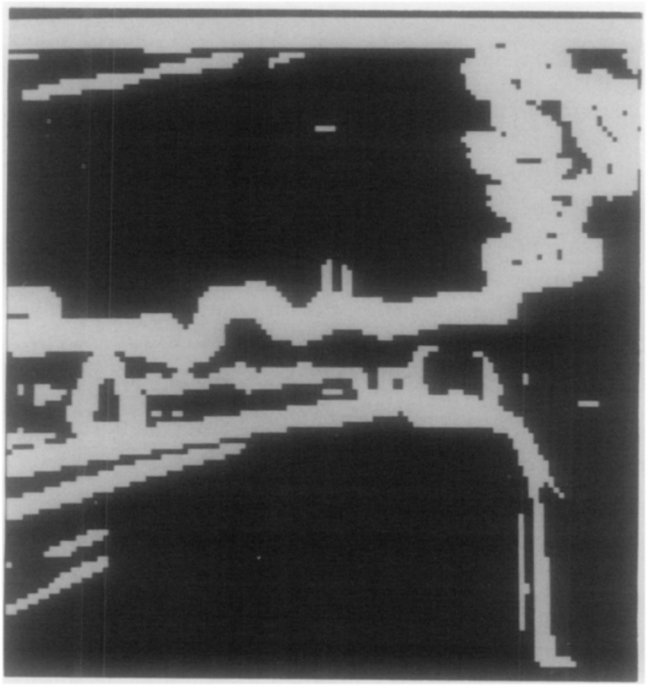

FIG. 7. Confidence measure, ROAD1 sequence.

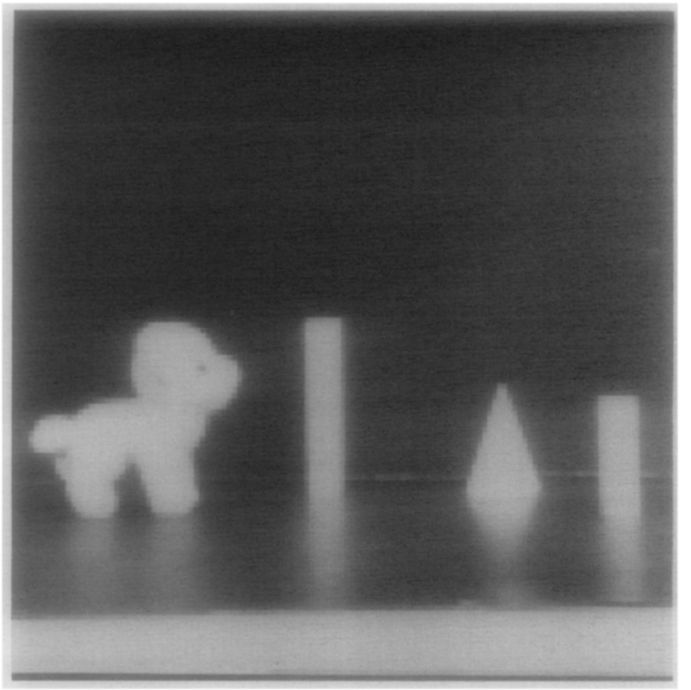

FIG. 8. Frame 1 of LAB sequence.

the foreground. The object motion depicted here contains rotational components (Figs. 4–7).

2. LAB sequence. This sequence was taken by a camera moving toward the center of the image. The background is mostly empty and homogeneous. Two blocks in the center are stationary and the one on the right is moving left. The toy dog on the left of the image is moving right. They are all on a table. Two stationary blocks in the middle have no motion at the edges, while the table shows a movement of either 0 or 1 pixel at the edge because it is closer to the camera. (Figs. 8–11).

3. ROAD2 sequence. This sequence was obtained from Martin-Marietta. In this sequence, the camera was mounted on a slowly moving vehicle. There are two moving cars in the scene. One car is on the far front of the camera and the other is passing by from the left. Since the camera is moving slowly, the motion between it and the background is not significant. (Figs. 12–15).

Four different detectors were first implemented and tested to verify the theoretical results we derived earlier. These detectors are

1. detector using temporal derivative E_t,
2. detector using estimated normal speed $|E_t| / \sqrt{E_x^2 + E_y^2}$,
3. detector using moving edginess $|E_t| \sqrt{E_x^2 + E_y^2}$, and
4. detector using our proposed method.

To estimate the gradient, we first smoothed each image by simply averaging the intensity values in a $3 \times 3 \times 3$ window and then applied a triquadratic least-squares

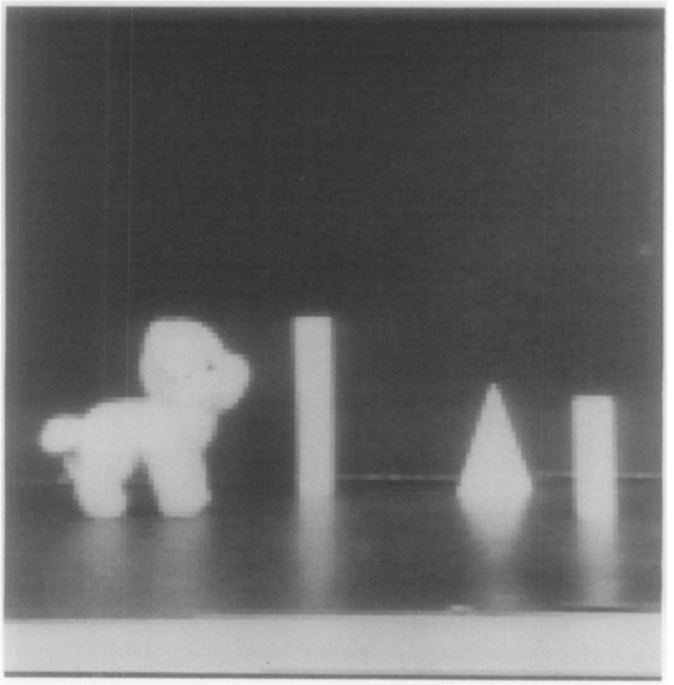

FIG. 9. Frame 5 of LAB sequence.

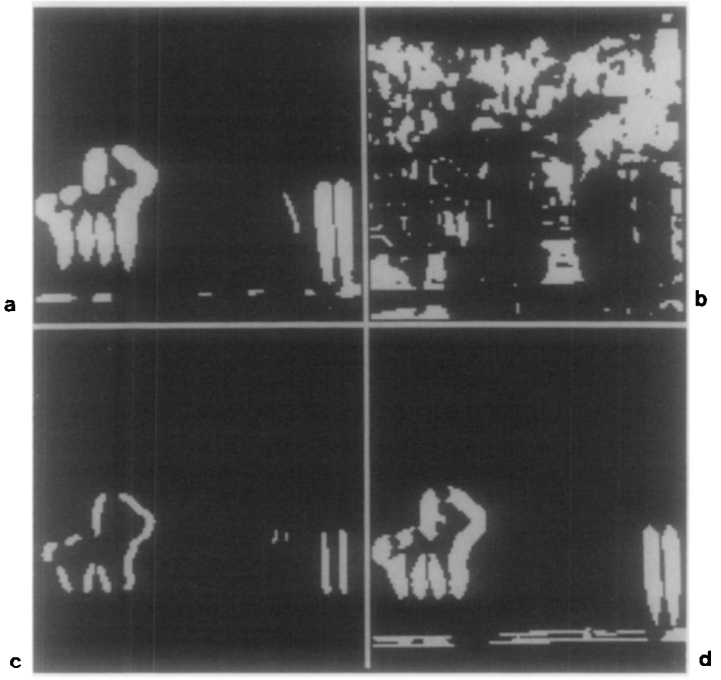

FIG. 10. Results of four motion detectors, LAB sequence.

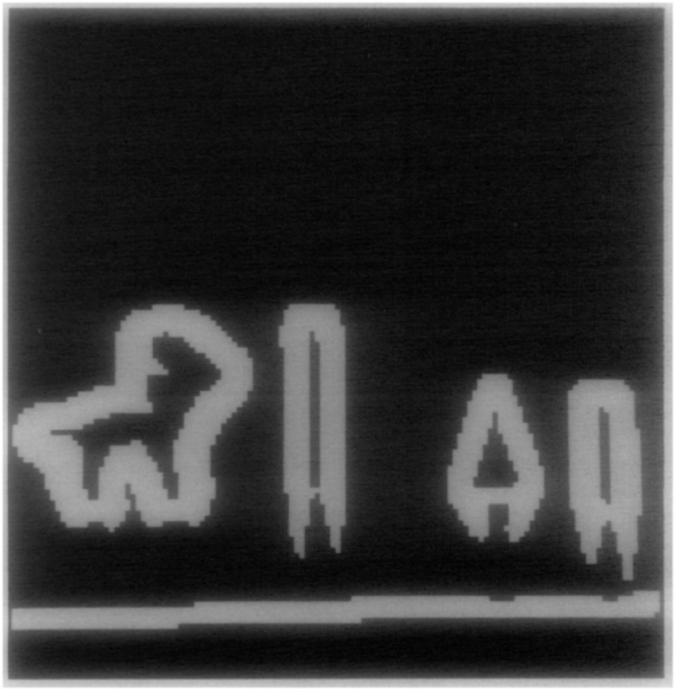

FIG. 11. Confidence measure, LAB sequence.

FIG. 12. Frame 1 of ROAD2 sequence.

FIG. 13. Frame 5 of ROAD2 sequence.

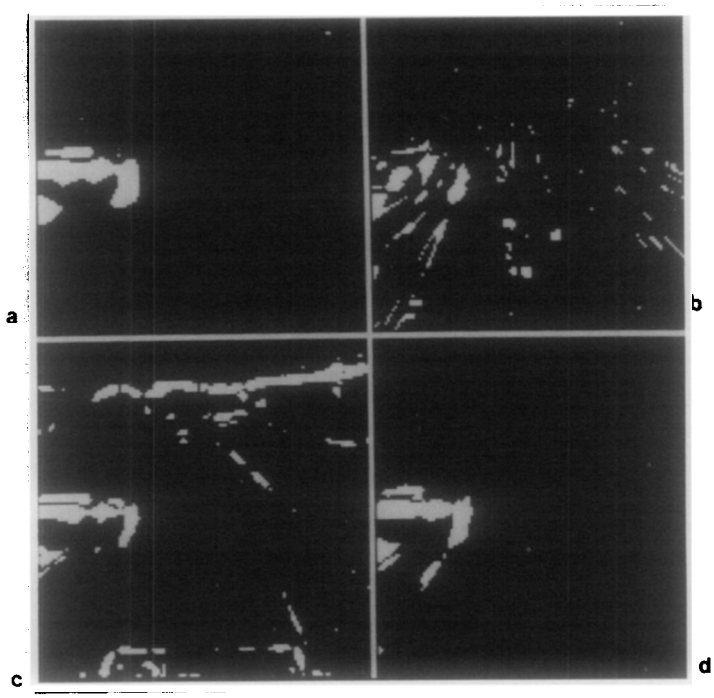

FIG. 14. Results of four motion detectors, ROAD2 sequence.

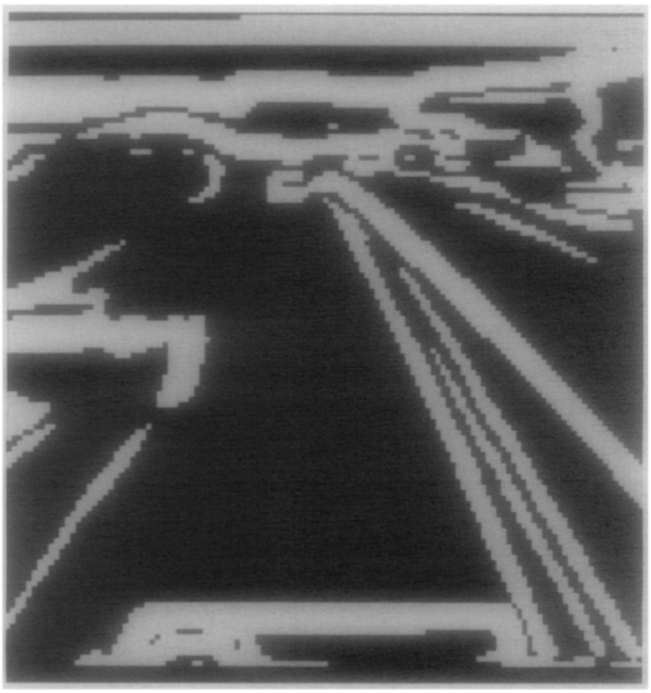

FIG. 15. Confidence measure, ROAD2 sequence.

surface fitting for a $5 \times 5 \times 5$ window, which is a direct extension of the biquadratic fitting used in [4]. The image gradients were estimated from the coefficients of the resulting triquadratic polynomial and were used for implementing all the four detectors. For our experimental results, we used a threshold 80° for the embedding angle in step 3 of our proposed method.

The results after applying all four detectors are shown in Figure 6, 10, and 14 for ROAD1, LAB, and ROAD2 sequences separately. The confidence measure was estimated in the third frame of each sequence and all pixels whose confidence measure above our preset threshold are shown white in Figs. 7, 11, 15. In each of Figs. 6, 10, and 14, (a), (b), (c), and (d) are the best thresholded results associated with detector (1), (2), (3), and (4).

Empirically, the determination of the image velocity in image areas having zero or very small spatial gradients is usually difficult. Some of the approaches even exclude those points from the solution process [6, 18, 27]. On the other hand, at a point along or close to a contour, the velocity can be obtained rather accurately after applying either an iterative or a minimization method. As is shown in our experiments, the performance of our confidence measure agrees qualitatively with these conclusions drawn from empirical analysis.

As evidenced from the results shown in Figs. 6b, 10b, and 14b, the detector using the magnitude of estimated normal velocity performed obviously the worst. This is because the detector has trouble in responding properly to somewhat homogeneous areas and then makes the responses from the rest of areas difficult to be judged on the same scale. In the following discussions, only detectors (1), (3), and (4) are considered.

In the ROAD1 sequence, the contour which separates the sky and the trees is part of the stationary background. Along this contour, detectors (1) and (3) report motion which is obviously incorrect. This is because the frames are misregistered and these two detectors usually have difficulty in differentiating moving edges from rapidly intensity-varying pixels. Similarly, the motion responses along the contour on top of the ROAD2 sequence from detector (3) seem to be influenced by the same problem. This problem of misregistration, however, does seem to have only a little effect on the performance of our detector. In addition, detector (3) also responds incorrectly to the bottom area of the ROAD2 sequence. This area is part of the moving vehicle on which the video camera is mounted. Finally, in analyzing the LAB sequence, the results obtained from detectors (1) and (3) seem to be very close. In this sequence, the two blocks in the center are almost stationary, whereas the table on the bottom area shows slight movement. These two detectors both fail in giving motion responses to the table without inferring motion from the two blocks.

As evidenced from the results shown in Figs. 6d, 10d, and 14d, our proposed detector appears to perform remarkably well in all three sequences. Its responses seem to be considerably closer to the *perceptible motion for the human observer*. Some minor false responses from the stationary background, however, still occur in the ROAD1 sequence. This is because the frames are seriously misregistered. Such problems are pervasive in the field of visual motion analysis.

We also implemented the detector proposed by Marr and Ullman [18] for comparison. The time derivative was implemented by simple difference as mentioned in [16]. As can be seen in Figs. 16–17 and Figs. 18–19 for the results from the ROAD1 and LAB sequences, respectively, black represents motion to the left

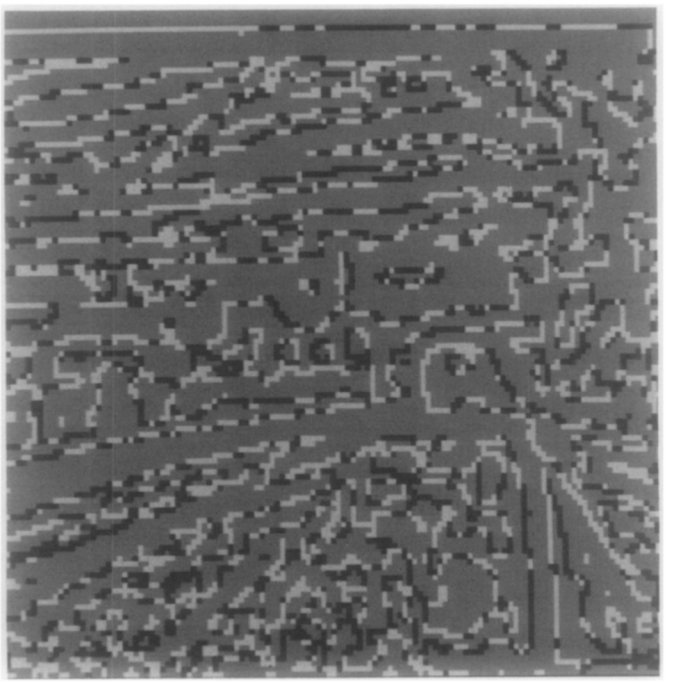

FIG. 16. Results of Marr and Ullman's detector, ROAD1 sequence, $\sigma = 1.5$.

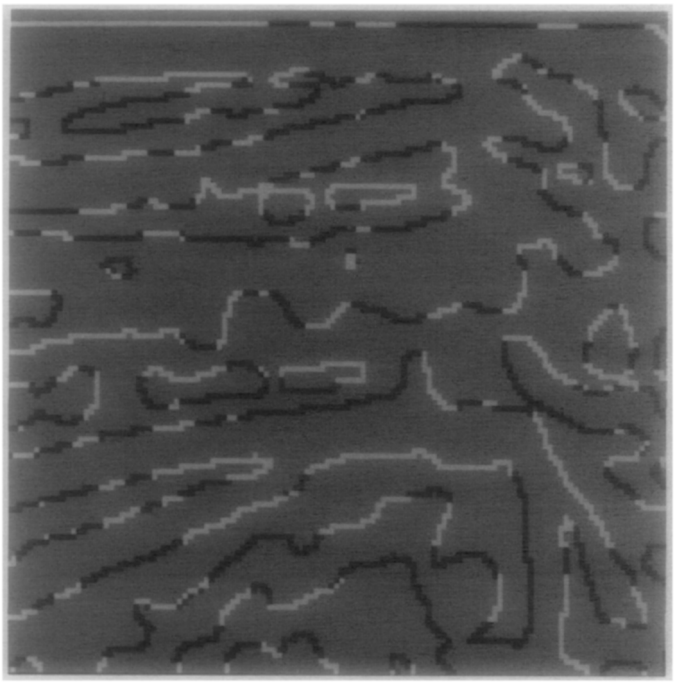

FIG. 17. Results of Marr and Ullman's detector, ROAD1 sequence, $\sigma = 3.0$.

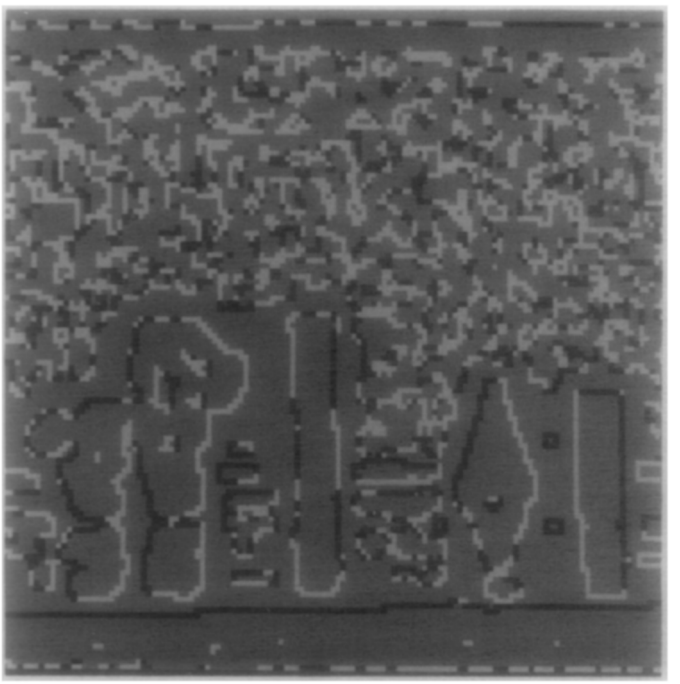

FIG. 18. Results of Marr and Ullman's detector, LAB sequence, $\sigma = 1.5$.

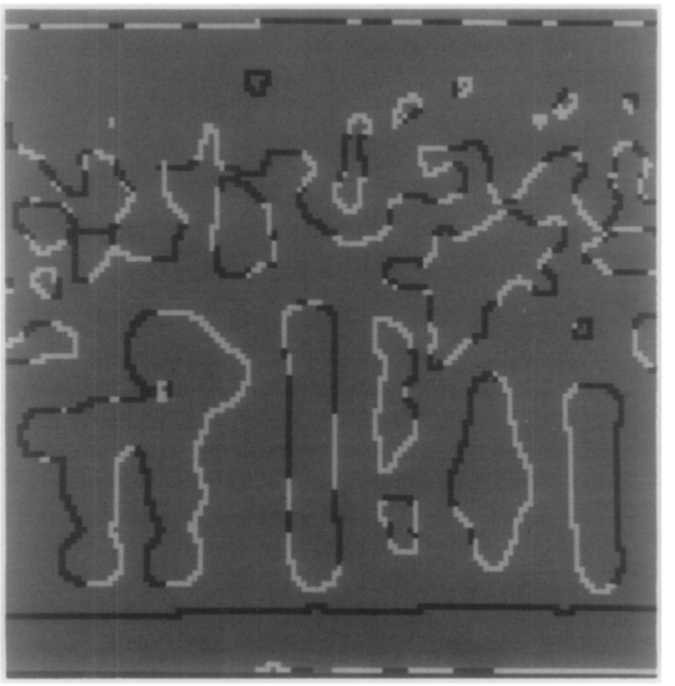

FIG. 19. Results of Marr and Ullman's detector, LAB sequence, $\sigma = 3.0$.

and white represents motion to the right, relative to its contrast. Because of the unavoidable registration errors and the nature of this measure, this method *almost always* reports spurious motion to the stationary background.

7. CONCLUSIONS

We assume that we are given discrete samples of an underlying spatio-temporal intensity hypersurface with perhaps some spatial as well as temporal noise added. Unlike the conventional approach, which tries to consider how the change of surface in the 3-dimensional space affectes the gray-value intensities in the 2-dimensional image domain, our approach seeks the information the characteristics of the intensity hypersurface can reveal about the 3-dimensional structure and motion. The experimental results obtained by applying our proposed motion detector to three real-world image sequences indicates that the motion detection based on a spatio-temporal surface model and an inseparable measure is feasible and provides excellent results.

Our discussions of the mathematical derivation can be concluded as follows:

1. Based on our analysis, the motion constraint equation seems to be the only available local information. If we use additional assertions such as *3D zero crossings are formed by the trajectory of 2D zero-crossing contours* and *the image irradiance has the property of linear variation along the orientation of step hyperedge*, we can compute the exact 3-surface normal once the 2D zero-crossing contour surface is found and the true motion constraint equation should be corrected as

$$E_x''\dot{x} + E_y''\dot{y} + E_t'' = 0, \tag{14}$$

where E'' is the 3D Laplacian operator and (E_x'', E_y'', E_t'') is the gradient of E''. Note that this is essentially the equation used by Buxton and Buxton [6].

2. As opposed to the traditional derivation of motion constraint equation, we find the equation itself is only an extension of the 2D edge constraint, and further show, from an intensity surface point of view, the equation only holds at either step hyperedges or places with uniform illumination along the path of a trajectory.

3. Based on the idea from surface model, we also define a confidence measurement to measure the correctness of the constraint equation. It has been shown a useful measure in interpreting image gradient information. A similar idea has been reported in [2] to measure the confidence of the computed dense displacement field.

Now, we turn to the question of what other areas this model can be applied to. The segmentation algorithm developed by Besl and Jain [4] for range images has also been successfully applied to a variety of intensity images. It seems that the conjecture of *surface coherence* is applicable to the intensity images as well. Therefore, the first application is to look into the issue of *hypersurface coherence* in spatio-temporal space with the aim to develop a robust dynamic scene segmentation algorithm.

Second, since the image illumination is usually conserved over small periods of time if there is no relative motion, any motion-varying surface characteristics can be used to measure motion. For instance, Marr and Ullman's motion detector can be thought of using the surface curvature as a motion measurement. This model will be helpful in discovering other kinds of surface characteristics suitable for motion detection.

APPENDIX: A REVIEW OF GEOMETRY IN HYPERSPACE

In general, we can define *a surface of dimension n*, in $\mathbf{R}^{n+1}$ as a nonempty subset S of $\mathbf{R}^{n+1}$ of the form $S = f^{-1}(c)$, where $f: U \to \mathbf{R}$ (U open in $\mathbf{R}^{n+1}$) is a smooth function with the property that $\nabla f(p) \neq \mathbf{0}\ \forall p \in S$ and c is a real constant.

The *n-plane* $a_1x_1 + \cdots + a_{n+1}x_{n+1} = b$ can also be defined, for $\mathbf{0} \neq (a_1, a_2, \ldots, a_{n+1}) \in \mathbf{R}^{n+1}$ and $b \in \mathbf{R}$, as the level set $f^{-1}(b)$, where $f(x_1, \ldots, x_{n+1}) = a_1x_1 + \cdots + a_{n+1}x_{n+1}$. Note that an *n-plane* is an *n-surface* for each $b \in \mathbf{R}$. For instance, a 1-plane is usually called a line in $\mathbf{R}^2$, a 2-plane is usually called a plane in $\mathbf{R}^3$ and an n-plane for $n > 2$ is sometimes called a *hyperplane* in $\mathbf{R}^{n+1}$.

Given any two vectors in $\mathbf{R}^{n+1}$, their inner product can be defined, using the standard dot product on $\mathbf{R}^{n+1}$. The length $\|\mathbf{v}\|$ of a vector $\mathbf{v}$ and the angle $\mathbf{v}\angle\mathbf{w}$ between two vectors $\mathbf{v}$ and $\mathbf{w}$ can then be defined by

$$\|\mathbf{v}\| = (\mathbf{v} \cdot \mathbf{v})^{1/2} \tag{15}$$

$$\mathbf{v}\angle\mathbf{w} = \cos^{-1} \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|}, \qquad 0 \leq \theta < \pi. \tag{16}$$

Therefore, the standard definition of the orthogonality can be extended to the hyperspace.

Each *n-surface* S has at each point $p \in S$ a *tangent space*, which is an n-dimensional vector subspace of the space $\mathbf{R}^{n+1}$ of all vectors at p orthogonal to $\nabla f(p)$.

Therefore, the tangent space of any n-surface in $\mathbf{R}^{n+1}$ at p forms a hyperplane with dimension n.

Let $g: U \to \mathbf{R}$ be a smooth function on the open set U in $\mathbf{R}^n$ and let $\phi: U \to \mathbf{R}^{n+1}$ be defined by $\phi(u_1, \ldots, u_n) = (u_1, \ldots, u_n, g(u_1, \ldots, u_n))$. The unit normal along ϕ is given by

$$\mathbf{N}(p) = \frac{1}{\left(\sum_{k=1}^{n} g_{u_k}^2 + 1\right)^{1/2}} \left(-\frac{\partial g}{\partial u_1}, \ldots, -\frac{\partial g}{\partial u_n}, 1 \right). \tag{17}$$

A parametrized curve in $\mathbf{R}^{n+1}$ is a smooth function $\alpha: I \to \mathbf{R}^{n+1}$, where I is some open interval in $\mathbf{R}$. The velocity vector at time t of the parametrized curve $\alpha: I \to \mathbf{R}^{n+1}$ is the vector at $\alpha(t)$ defined by

$$\dot{\alpha}(t) = \frac{d\alpha}{dt}(t). \tag{18}$$

This vector is tangent to the curve α at $\alpha(t)$.

Finally, the projection of any vector $\mathbf{b}$ in $\mathbf{R}^{n+1}$ onto an arbitrary subspace of dimension n is given by

$$p = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b \tag{19}$$

where each vector in the column space of A is a basis in this subspace.

REFERENCES

1. E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion, *J. Opt. Soc. Amer.* **2**, No. 2, 1985, 284–299.
2. P. Anandan, Computing dense displacement fields with confidence measures in scenes containing occlusions, in *Proceedings, Image Understanding Workshop, New Orleans, 1984,* pp. 236–246.
3. H. B. Barlow and W. R. Levick, The mechanism of directionally selective units in rabbit's retina, *J. Physiol. London* **178**, 1965, 477–504.
4. P. J. Besl and R. Jain, Segmentation through variable-order surface fitting, *IEEE Trans. Pattern Anal. Mach. Intell.* **10**, No. 2, 1988, 167–192.
5. B. F. Buxton and H. Buxton, Monocular depth perception from optical flow by space-time signal processing, *Proc. Roy. Soc. London B* **218**, 1983, 27–47.
6. B. F. Buxton and H. Buxton, Computation of optical flow from the motion of edge features in image sequences, *Image Vision Comput.* **2**, No. 2, 1984, 59–75.
7. R. C. Bolles, H. H. Baker, and D. H. Marimont, Generalizing epipolar-plane image analysis for non-orthogonal and varying view directions, in *Proceedings, Image Understanding Workshop, Los Angeles, Feb. 23–25, 1987,* pp. 843–848.
8. C. Enroth-Cugell and J. G. Robson, The contrast sensitivity of retinal ganglion cells of the cat, *J. Physiol.* **187**, 1966, 517–552.
9. D. J. Fleet, *The Early Processing of Spatio-Temporal Visual Information*, M.Sc. thesis, Dept. of Computer Science, Univ. of Toronto, 1984.
10. D. J. Fleet, A. D. Jepson, and P. E. Hallet, *A Spatio-Temporal Model for Early Visual Processing*, RCBV-TR-84-1, Research in Biological and Computational Vision, Univ. of Toronto, 1984.
11. S. Haynes and R. Jain, Time-varying edge detection, *Comput. Graphics Image Process.* **21**, 1983, 345–367.
12. R. M. Haralick, J. S. Lee, and H. N. Joo, The facet approach to optical flow, manuscript, 1984.
13. D. J. Heeger, Optical flow from spatiotemporal filters, in *Proceedings, First Int. Conf. on Computer Vision, 1987,* pp. 181–190.
14. B. K. P. Horn and B. G. Schunck, Determining optical flow, *Artif. Intell.* **17**, 1981, 185–203.

15. R. Jain, Dynamic scene analysis using pixel-based processes, *IEEE Comput.* Aug. 1981, 12–18.
16. D. Marr, *Vision*, Freeman, New York, 1982.
17. D. Marr and E. Hildreth, The theory of edge detection, *Proc. Roy. Soc. London B* **207**, 1980, 187–217.
18. D. Marr and S. Ullman, Directional selectivity and its use in early visual processing, *Proc. Roy. Soc. London B* **211**, 1981, 151–180.
19. W. Reichardt, Autocorrelation, a principle for the evaluation of sensory information by the central nervous system, in *Sensory Communication* (W. A. Rosenblith, Ed.), Wiley, New York, 1961.
20. B. G. Schunck, The image flow constraint equation, *Comput. Vision Graphics Image Process.* **35**, 1986, 20–46.
21. I. K. Sethi and R. Jain, Finding trajectories of feature points in a monocular image sequences, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-9**, No. 1, 1987, 56–73.
22. J. P. H. Van Santen and G. Sperling, Temporal covariance model of human motion perception, *J. Opt. Soc. Amer.* **1**, No. 5, 1984, 451–473.
23. J. P. H. Van Santen and G. Sperling, Elaborated Reichardt detectors, *J. Opt. Soc. Amer.* **2**, No. 2, 1985, 300–321.
24. A. Verri and T. Poggio, Against quantitative optical flow, in *Proceedings, First Int. Conf. on Computer Vision, 1987*, pp. 171–180.
25. A. B. Watson and A. J. Ahumada, Jr., Model of human visual-motion sensing, *J. Opt. Soc. Amer.* **2**, No. 2, 1985, 322–342.
26. A. P. Witkin, Scale-space filtering, in *Int. Joint Conf. on Artif. Intell., 1983* Vol. 2, pp. 1019–1022.
27. M. Yachida, Determining velocity maps by spatio-temporal neighborhoods from image sequences, *Comput. Vision Graphics Image Process.* **21**, 1983, 262–279.
28. S. W. Zucker and L. Iverson, From orientation selection to optical flow, *Comput. Vision Graphics Image Process.* **37**, 1987, 196–220.