Computer Vision and Image Understanding **82,** 1–32 (2001)
doi:10.1006/cviu.2000.0900, available online at http://www.idealibrary.com on IDEAL®

The Statistics of Optical Flow[1]

Cornelia Fermüller, David Shulman, and Yiannis Aloimonos

Center for Automation Research, University of Maryland, College Park, Maryland 20742-3275
E-mail: fer@cfar.umd.edu, yiannis@cfar.umd.edu, dshulman@cyc.com

Received May 23, 2000; accepted December 6, 2000

When processing image sequences some representation of image motion must be derived as a first stage. The most often used representation is the optical flow field, which is a set of velocity measurements of image patterns. It is well known that it is very difficult to estimate accurate optical flow at locations in an image which correspond to scene discontinuities. What is less well known, however, is that even at the locations corresponding to smooth scene surfaces, the optical flow field often cannot be estimated accurately.

Noise in the data causes many optical flow estimation techniques to give biased flow estimates. Very often there is consistent bias: the estimate tends to be an underestimate in length and to be in a direction closer to the majority of the gradients in the patch. This paper studies all three major categories of flow estimation methods—gradient-based, energy-based, and correlation methods, and it analyzes different ways of compounding one-dimensional motion estimates (image gradients, spatiotemporal frequency triplets, local correlation estimates) into two-dimensional velocity estimates, including linear and nonlinear methods.

Correcting for the bias would require knowledge of the noise parameters. In many situations, however, these are difficult to estimate accurately, as they change with the dynamic imagery in unpredictable and complex ways. Thus, the bias really is a problem inherent to optical flow estimation. We argue that the bias is also integral to the human visual system. It is the cause of the illusory perception of motion in the Ouchi pattern and also explains various psychophysical studies of the perception of moving plaids.

Key Words: analysis of optical flow estimation algorithms; bias; optical illusion.

CONTENTS

[1] The support of this research by the Office of Naval Research under Contract N00014-95-1-0521 is gratefully acknowledged, as is the help of Sara Larson in preparing this paper.

1077-3142/01 $35.00

3. *Correlation methods.*
4. *An explanation of optical illusions.*
5. *Computational models of motion processing.*
6. *Conclusions.*

1. THE PROBLEM

1.1. Errors Matter

A serious problem with optical flow computation is that the flow must be estimated using noisy data, and it is often not possible to accurately estimate the noise parameters. Because there is noise, the estimated and the actual flows can be different. Worse, the estimate often is biased; the expected value of the difference between the actual and estimated flow is not zero. Confidence limits are also difficult to predict, and they matter because the variance of the flow can be large.

It will be shown here that many commonly used methods for computing optical flow are biased. It is difficult to correct for this bias because it is difficult to estimate the noise parameters. It might be possible if the parameters were static, but instead they change in unpredictable and complex ways. If we had enough data, we could use various statistical techniques to estimate the parameters; for example, we could use maximum likelihood. But when we first view a scene, there is not enough data. When the environment changes, when the lighting conditions have changed recently or there has been a significant recent change in orientation which produces a change in the accuracy of the constraint that corresponding points have the same intensity, there is not enough data about the current values of the noise parameters.

In this paper we analyze all three major classes of optical flow algorithms: gradient methods, frequency-domain methods, and correlation methods. We analyze both linear and nonlinear estimation techniques, as well as some robust methods. In some of our analyses we do not make either a Gaussian or an asymptotic assumption.

1.2. Bias

There has been previous work on optical flow that analyzes error. Examples are [20, 34, 39, 45]. However, it has not been widely noticed by the computer vision community that optical flow estimates can be biased. It has been pointed out in [30] that optical flow estimated using gradient methods is biased: estimates tend to be underestimates. As we shall show here, even the estimates of the direction of flow are not unbiased. We also demonstrate here that it is not just gradient methods that result in biases. The mathematics of frequency domain methods is not very different than that of gradient methods and similar biases arise. Finally, we present a model that shows how even correlation methods can be biased and tend toward underestimation. Thus all these methods produce biases. It is not often appreciated how difficult these biases are to correct.

We conclude by arguing that more robust and more qualitative methods should be used for estimating optical flow and that the estimation of optical flow should be combined with the estimation of three-dimensional information.

1.3. Optical Illusions

The inevitability of bias provides an explanation of certain well-known optical illusions. In particular, we provide a computational model for the Ouchi illusion.

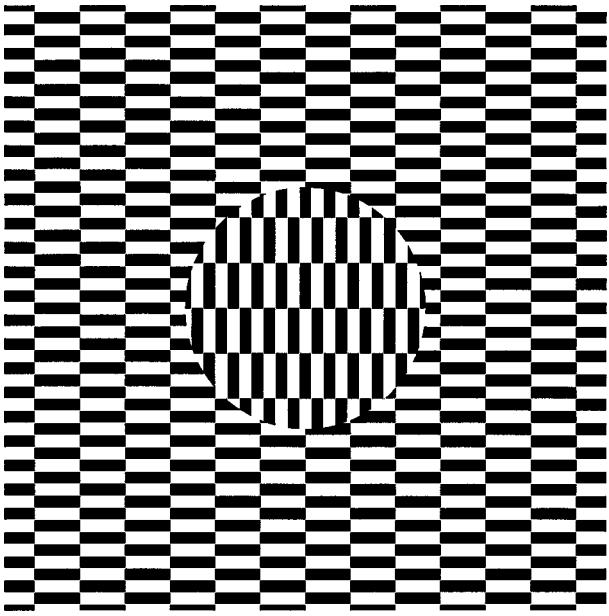

FIG. 1. A pattern similar to one by Ouchi [31].

The striking illusion discovered in 1977 by the graphic artist H. Ouchi consists of two black and white rectangular checkerboard patterns oriented in orthogonal directions—a background orientation surrounding an inner ring (Fig. 1). Small retinal motions, or slight movements of the paper, evince a segmentation of the inset pattern and motion of the inset relative to the surround. The illusion occurs for a variety of viewing distances and angles. Some observers report an apparent depth discontinuity, with the center floating as it moves above the background [36].

Our explanation of the illusion lies in the estimation of differently biased flow vectors in the two patterns. Because of the sparse spatial frequencies in these checkerboard patterns the bias is highly pronounced. In the following two different 3D motions are derived which cause the inset to move relative to the surround. Our model for explaining this illusion is given in Section 4 along with a set of illustrations.

1.4. A Formulation of the Flow Estimation Problem

The rest of this paper is a detailed exploration and examination of the themes mentioned above. In order to proceed further, it will be necessary to discuss in somewhat more technical detail exactly what is meant by flow and the kinds of methods for estimating flow that we are interested in analyzing.

It is assumed that a sequence of two or more images of a scene is available. If in the real world, whatever is at point P_1 at time t_1 is found at point P_2 at time t_2, and point p_i is the image of real-world point P_i in image i, then image points p_1 and p_2 will be said to be corresponding points.

A basic assumption is that there exists some attribute that has the same value I at the two corresponding image points, p_1, p_2. The value I might be the intensity of light at a point p in the image or the intensity of light of a given frequency. But I might also be something somewhat less local, such as the (weighted) average value of light intensity in

some region R of the image. In computing the average, points that are nearest to p have the most weight. The value I might be assumed exactly known, or I might be modeled as being measured with a certain amount of error. The error may often be very large, but it is assumed that the value of I is the same at corresponding points and that this value is known with reasonable accuracy at a significant number of points. (Later we discuss more general constraints where the I's at corresponding points are not equal, but instead there is a linear relation between the values of I at corresponding points.) Of course, despite the long tradition in computer vision of assuming exactly equal values at corresponding points, there is no meaningful physical quantity that is going to have exactly the same value at two corresponding points. But the statistical modeling below will take into account error in the constraint and realize that some of the error is due not to noisy observation but to imperfect constraint. Given the difficulty of making locally accurate flow estimations and the large amount of noise involved in any case, it is reasonable to work with the assumption of equal I at corresponding points.

If p_1 at time t_1 and p_2 at time t_2 are corresponding points, there is a two-dimensional motion from p_1 to p_2. It makes sense to speak of the velocity of this motion. If the difference between the times t_1 and t_2 is infinitesimal, one speaks of the instantaneous two-dimensional velocity or optical flow.

The constraint that some attribute have the same value at corresponding points is not enough. Other constraints must be employed in order to actually estimate flow. These additional constraints usually amount to models of the 2D velocity field, for example, constraints on the sizes of the derivatives, or parametric models of the velocity field.

To simplify the analysis, initially we will focus on the simplest special case where the flow is the same at all positions in the image or at least constant in some large region of the image. To further simplify matters, it will be assumed that special problems due to the fact that certain pixels are near the boundary of an image can be safely ignored.

1.5. Methods of Computing Optical Flow

There are three primary classes of methods for computing optical flow: gradient-based methods, frequency-domain methods, and correlation methods. Gradient-based and frequency-domain methods derive optical flow in two separate stages: first, information about the one-dimensional motion components of local edges or single spatial frequencies is obtained; then, the individual measurements within some neighborhood are combined into an estimate of optical flow. These two classes of methods are faced with similar noise issues and thus can be given a very similar mathematical analysis. Correlation techniques perform region-based matching and in general cannot be separated into one-dimensional and two-dimensional components. Thus, they will be given a separate analysis that is somewhat different, but not very different.

Gradient-based techniques [8, 22, 40, 44] compute the spatial and temporal derivatives of the intensity or functions of the intensity. These measurements define at individual points the component of flow perpendicular to edges, the normal flow. To derive these measurements the images are usually smoothed in space and time with low-pass filters and numerical differentiation is performed. Then the optical flow is computed from the local one-dimensional measurements in a neighborhood using assumptions about the smoothness [9, 21, 22, 27–29, 41] or an explicit model (such as polynomial) of the underlying flow field. The estimation amounts to solving an optimization problem minimizing some function of deviation from

the model; if the flow field is assumed to be constant, least squares or weighted least squares estimation is often used [25, 34], but other techniques such as total least squares [45], or robust techniques, can alternatively be used. If more elaborate smoothness assumptions are employed, iterative techniques must be used.

Energy-based techniques [1, 2, 4, 20] are based on the constraint that all the energy of a translating pattern lies in a plane through the origin in spatiotemporal frequency space. Usually, the energy for a number of spatial and temporal frequency triplets is extracted by means of spatiotemporal energy filters of different kinds. Computationally this results in taking a local Fourier or related transform and it requires some smoothing and interpolation either in space-time or in the frequency domain. Since energy can only be extracted within regions, the implicit assumption is that the flow field is constant within the range of the filters. The fitting of the plane to the estimated energy responses again amounts to an optimization problem which can be solved by linear estimation, but more often is addressed using total least squares estimation.

Another approach to flow estimation in frequency space is based on the assumption that phase is preserved [14]. In this case the phase response for spatiotemporal frequencies is computed using energy filters and then the spatial and temporal derivatives of the phase are estimated to obtain one-dimensional motion components.

Correlation techniques [6, 7, 26, 47] have mostly been used in the processing of stereo images where one component of the displacement is defined by the epipolar constraint and to establish sparse feature correspondence when far-apart views obtained by a moving camera are considered (discrete motion). They have also been used to derive dense correspondence fields and optical flow fields. Correlation techniques compare regions of usually large extent in the two images to find the displacement between the regions which provides the best match. A measure of similarity is computed between regions centered at discrete (pixel) locations and the exact displacement is then estimated by interpolation between the discrete positions. Similarity may be measured using cross-correlation, which may be normalized, or using a distance measure such as sum-of-squared-differences. It is then necessary to find the displacement that maximizes the correlation or minimizes the distance measure. By considering large matching regions it is implicitly assumed that the correspondence field is constant, and the aperture problem is circumvented in this way. There are also correlation techniques for flow which match small image regions and thus face the aperture problem. In this case local correlation surfaces are combined via smoothness constraints to estimate the optical flow field [3, 35].

2. GRADIENT-BASED AND FREQUENCY-DOMAIN METHODS

Gradient-based methods and frequency-domain methods use essentially the same constraint; the frequency-domain constraint can be obtained from the gradient constraint by taking a Fourier transform. For both methods the constraint is encoded as an overdetermined system of noisy linear equations. There are many techniques for solving such a system of equations. Many different approximations can be employed. Many different methods exist for handling the noise. But in any case, there will be a noise term that comes from an inaccurate estimate of the spatial derivatives and a noise term that comes from an inaccurate estimate of the temporal derivative. The bias will arise because we cannot obtain a good estimate of the ratio of these two noise terms.

We proceed by providing a general framework for the basic constraint equation which describes the one-dimensional motion components of single spatial frequencies (Section 2.1). We then describe three classes of estimation techniques for solving a system of these equations: ordinary least squares (Section 2.2) and total least squares (Section 2.3), and we also sketch some robust techniques (Section 2.4). A discussion of models that do not assume constant flow (Section 2.5) and a summary (Section 2.6) conclude the section.

2.1. The Constraint Equation

If I is the attribute value that is constant at corresponding points, then

$$\frac{DI}{Dt} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \tag{1}$$

where t represents time and x and y represent two different Cartesian coordinates. The letters u and v represent the x and y components of the flow $\mathbf{u}$, respectively. Of course, our observations of these derivatives are inevitably going to be highly noisy but that fact is incorporated in our noise model. The mathematical issue of whether these derivatives exist is not a real issue. Both in the real world and the image domain, we smooth or average a small amount before we do anything else.

We first consider the simplest statistical model. Equation (1) cannot be expected to be strictly valid. More reasonable would be

$$\frac{DI}{Dt} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = \epsilon. \tag{2}$$

Here ϵ is a noise variable that might be assumed to be zero-mean Gaussian.

There is another interpretation of this equation; one might assume that (1) is exactly true but one cannot observe the temporal derivative $\frac{\partial I}{\partial t}$ with perfect accuracy. Instead one can only observe $\frac{\partial I}{\partial t} + \epsilon$ where ϵ is some noise variable that might be zero-mean Gaussian. Later we assume there are also errors in the observation of the spatial derivatives, and these terms cause the main difficulties, but we ignore these terms for a while in order to simplify the analysis.

Equation (1) (or (2)) is really many equations. There is one equation for each point in the image, or one equation for each point where the data are reasonably accurate, and to indicate the dependence of ϵ on the point p of observation we will write ϵ_p in (2).

We assume the flow is constant over the region of interest. If the different ϵ_p's have the same statistical distribution and are independent, zero-mean Gaussian variables, then the maximum likelihood solution is obtained by using least squares: Find the u, v that minimize the expression $\sum_p \epsilon_p^2$.

Even if the ϵ_p's are not Gaussian, but the different ϵ_p's have equal variance and all the different noise variables are uncorrelated, then the least squares solution is the best linear unbiased estimate (BLUE) [24, 43]. We explain now what being BLUE means. Let unprimed letters represent estimates and use primes to indicate that actual quantities are being referred to. Being unbiased means that the expected value of $\mathbf{u} - \mathbf{u}'$ is zero. (The ordinary least squares estimate is not unbiased if there are errors in the measurement of the spatial as well as the temporal derivatives.) To say that the BLUE estimate is linear is to say that it is linear in the observed values of the temporal derivatives $\frac{\partial I}{\partial t}$. For an estimate to be

the best means it is the best with respect to some measure or metric. Goodness is measured by the expected value of $\|\mathbf{u}' - \mathbf{u}\|^2$. Here $\|\ \|$ means the length of a vector.

It is significant that the BLUE estimate need not be the best linear estimate, only the best linear unbiased estimate [23]. The classical James–Stein [23] best linear biased estimate actually is relevant only for estimation of an n-component quantity with $n > 2$. In that case, it makes sense to use a biased estimate (an estimate that is known to be an underestimate) in order to reduce the variance of the estimate and thereby reduce the expected size of the error.

Flow has only two components, but if the error in (1) is nonzero mean [17], while the different ϵ_p's are uncorrelated and identically distributed, then there is a three-component unknown; there are two components of flow and the third component is the mean value of ϵ_p, and we do have to estimate the mean value of ϵ_p in order to compute flow. If there is a change in the global ambient illumination, ϵ_p will not be zero-mean. Later we will relax the assumption that flow is constant, and then we will have more than two components to estimate, so James–Stein biased estimation is relevant.

2.1.1. A parenthetical remark about how to handle correlated error. Even ignoring the possibility that the best linear estimate is biased rather than BLUE, we see that there are other problems. If the various noises ϵ_p are not uncorrelated or the different noise variables have different variances, it is first necessary to apply a whitening transformation before computing the ordinary least squares solution. If we wish to solve a system of equations $L_i = 0$ with correlated error, then we must first multiply by some matrix Q such that $\sum_j Q_{ij} L_j = 0$ is a system of equations with uncorrelated errors of equal variance. Considering the set of expressions L_i as a vector, we need to solve $QL = 0$ using ordinary least squares. In order to actually apply the whitening transformation to (1), we need to know something about the statistics of the ϵ_p. In practice, it might be difficult to accurately estimate Q. We will ignore this problem.

2.1.2. Preprocessing by linear smoothing. A problem arises because we cannot even apply (1) unless we can make reasonable estimates of certain derivatives of I. Since, in any case, it is difficult to estimate pointwise spatial and temporal derivatives, it makes sense to apply some kind of linear operation before solving for the flow. In fact, let G be a linear operator and pretend that I, which is a function of one temporal variable and two image spatial variables, is defined on all of R^3, so that

$$GI(x, y, t) = \iiint G_{x,y,t}(a, b, c) I(a, b, c)\, da\, db\, dc. \tag{3}$$

Furthermore let us assume that G is a convolution, so that the coefficient $G_{x,y,t}(a, b, c)$ depends only on $(x - a, y - b, t - c)$. Then if flow is constant, from the fact that the derivation and convolution operators commute, we can conclude that

$$\frac{\partial GI}{\partial x} u + \frac{\partial GI}{\partial y} v + \frac{\partial GI}{\partial t} = 0. \tag{4}$$

Like (1), (4) is only approximately correct. Equation (4) normally makes little sense unless u, v are constant, but if the convolution is very local so that $G_{x,y,t}(a, b, c)$ is very small unless (x, y, t) and (a, b, c) are very close, then it makes sense to apply (4) at a point (x, y, t) even if flow is not constant everywhere but only approximately constant in the vicinity of (x, y, t).

If (1) has Gaussian error, so does (4). If the errors in (1) are uncorrelated, that need not be true of (4).

For an interesting special case of (4), let G be a Gaussian smoother. In fact, G might be an ordered set of smoothers. Equation (4) makes sense if $G_{x,y,t}(a, b, c)$ is a real vector rather than a real scalar. A two-component real vector can be reinterpreted as a complex scalar. This will allow the Gabor transform to fit into the schema of (4) and be relevant when we discuss frequency-based methods.

2.1.3. Application of linear transformations to I. We might apply the Fourier or some other linear transform to $\frac{DI}{Dt}$ considered as a function of position in space-time. A Fourier transformation is a convolution by a set of exponential functions. It is especially sensible to apply the Fourier convolution rather than some other convolution because by Parseval's theorem, the quadratic norm is preserved under Fourier transformation. This means that if f is a complex function and f^* its complex conjugate and $\|f\|_2^2 = \iiint f(x, y, t) f^*(x, y, t)\, dx\, dy\, dt$ and $\mathcal{F}$ represents the operation of computing the Fourier transform, then $\|\mathcal{F} f\|_2 = \|f\|_2$. Thus it does not matter whether we compute the least squares solution in the frequency or the space domain.

In more detail, letting subscripts represent partial differentiation, we start with the equation $I_x u + I_y v + I_t = \epsilon$. Taking the three-dimensional Fourier transform (the three dimensions being time and the two spatial dimensions of the image) and letting $\omega_x, \omega_y, \omega_t$ represent the spatial and temporal frequencies, we obtain the equation $(\omega_x u + \omega_y v + \omega_t)\mathcal{F}I - \sqrt{(}-1)\mathcal{F}\epsilon = 0$. To minimize the L^2 norm of ϵ is the same as minimizing the L^2 norm of $\mathcal{F}\epsilon$ and this would be the same as minimizing the L^2 norm of $(\omega_x u + \omega_y v + \omega_t)\mathcal{F}I$. This means we need to find the flow u, v that defines a plane with points ω_x, ω_y, w (that is for every $\omega_x, \omega_y, \omega_t$ there is w for which $\omega_x u + \omega_y v + w = 0$) such that the L^2 norm of the product $(\omega_t - w)$ times $\mathcal{F}I$ is minimized. In other words we need to minimize a weighted sum of energies; for each triplet of frequencies $(\omega_x, \omega_y, \omega_t)$ that is off the plane $\omega_x u + \omega_y v + w = 0$ corresponding to the flow u, v we multiply the energy in that frequency by a weight that is the distance squared $(\omega_t - w)^2$.

Another possibility is that there is noise in the estimation of the derivatives, both spatial and temporal. Then we would have the equation $(\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t}) - N_x u - N_y v - N_t = 0$ where the N's represent noise. Given data about the observed I (in the equation above, I represents the observed I), we want to estimate the flow u, v and the noise N_x, N_y, N_t in such a way as to minimize $\|N_x\|_2^2 + \|N_y\|_2^2 + \|N_t\|_2^2$, where again the subscripts indicate the L^2 norm (this corresponds to total least squares estimation).

Taking the Fourier transform, we obtain $(\omega_x u + \omega_y v + \omega_t)\mathcal{F}I - \sqrt{-1}(\mathcal{F}N_x u + \mathcal{F}N_y v + \mathcal{F}N_t) = 0$. Here the noises in the Fourier domain are functions of frequency. We divide the noise by $\sqrt{-1}\mathcal{F}I$ to obtain appropriate noise variables m for which we can write $((\omega_x + m_x)u + (\omega_y + m_y)v + (\omega_t + m_t))\mathcal{F}I = 0$. We have to choose u, v, N to minimize $\|\mathcal{F}N_x\|_2^2 + \|\mathcal{F}N_y\|_2^2 + \|\mathcal{F}N_t\|_2^2$, or equivalently minimize $\int((|m_x|^2 + |m_y|^2 + |m_t|^2)|\mathcal{F}I|^2)$, the integral being taken over all triplets ω_x, ω_y, ω_t. That is the triplets $(\omega_x + m_x, \omega_y + m_y, \omega_t + m_t)$ lie on the plane that defines the flow and we measure the sum of the distances between these triplets and the triplets $(\omega_x, \omega_y, \omega_t)$ (the normal distances to the plane) times the energy at each frequency.

Alternatively, we might want to compute only local Fourier transforms. And inevitably, we lose some information because we do not have an infinite image or a continuous set of observations, but that just adds additional noise in certain frequencies and forces us to

deal with local transforms. Let us multiply I by a Gaussian G centered at point P or by a function G that it is equal to 0 at points far from P and equal to 1 near P and then take the Fourier transform. The effect on (1) is to give less weight to data at points far from P when computing the least squares solution. The noise is also multiplied by G. In the frequency domain, instead of multiplying the noise by G, we convolve the Fourier transform of the noise by the Fourier transform of G. But the Fourier transform of a Gaussian is also a Gaussian. In other words, we minimize an expression very much like that in the previous paragraph, but here we first multiply $\mathcal{F}I$ by a distance, then smooth the result by convolving with a Gaussian, and finally compute an L^2 norm which must be minimized.

In Appendix A we analyze relevant weighting functions for phase-based methods [14]. Such methods are based on the assumption of conservation of local phase which is estimated using the Gabor transform or some local Fourier transform.

In practice, Fourier or even Gabor transforms might be too hard to compute, so one computes some finite approximation, but still uses the idea of minimizing something involving the product of a distance function and an energy function.

In any case there is an equation of the form

$$A_i u + B_i v = C_i \tag{5}$$

for each index i where the indices usually represent points in space-time or frequencies in some kind of transformation space. Often (5) is solved by a kind of ordinary least squares. Alternatively it may be solved with total least squares, a method that allows for errors in the observations of the spatial derivatives of I, or some more robust method. It may first be necessary to apply a whitening transform in order to handle correlation between the errors of different equations, but for the most part we will ignore that possibility.

2.2. Errors in the Ordinary Least Squares Solution

Let us first analyze the simplest method of solving (5): ordinary least squares. In the following, unprimed letters are used to denote estimates, primed letters to denote actual values, and δ's to denote errors, where $\delta A = A - A'$, $\delta B = B - B'$, and $\delta C = C - C'$.

Let n be the number of indices i to which (5) applies. In order to explicitly represent the errors the equation is rewritten as

$$(A_i - \delta A_i)u + (B_i - \delta B_i)v = C_i - \delta C_i. \tag{6}$$

It is also convenient to explicitly represent the equation in matrix form

$$(E - \delta E)\mathbf{u} = \mathbf{C} - \delta\mathbf{C}. \tag{7}$$

Here E and δE are n by 2 matrices which incorporate the data in the A_i and B_i. The vector $\mathbf{u}$ denotes the flow whose components are u and v.

By definition the least squares solution is given by

$$\mathbf{u} = (E^t E)^{-1} E^t \mathbf{C}. \tag{8}$$

If there are no errors in the estimation of the coefficients A_i and B_i, then under the usual assumptions that the different δC_i are uncorrelated and have the same variance, least squares

gives an unbiased estimate and it is also simple to give confidence limits for the solution.

$$\rho = \frac{\|\mathbf{C} - E\mathbf{u})\|_2^2}{n-2} \tag{9}$$

is an unbiased estimator of the variance of the δC_i [15, 24] where $\| \ \|_2$ represents the quadratic norm defined on vectors (i.e., the square root of the sum of the squares of the values of the components of the vector) and $\mathbf{u}$ represents the ordinary least squares solution to (5).

If we use a weighted least squares solution instead of the ordinary least squares solution and the weights are positive, then if $((w_i)^{1/2})^2$ is the weight of equation i, the weighted least squares solution is the same as the ordinary least square solution that would be obtained if we put $w_i^{1/2}E_{ij}$ and $w_i^{1/2}C_i$ in place of E_{ij} and C_i, in (5). So even in this case it is possible to estimate the variance of the error in the flow estimate.

But, in fact, these error estimates should be modified to take into account the bias. There will be errors δA_i, δB_i and these errors will cause the least squares estimate of the flow to be biased.

It is well known in the statistics community that the usual effect of the errors δA_i, δB_i (i.e., errors in the matrix E) is to produce an underestimate of the magnitude of $\mathbf{u}$ [15, 18, 37], and the bias also affects the estimate of the direction.

In the following, for two somewhat different models, this bias is demonstrated. In both cases it is assumed that the errors δE_i and δC_i are independent, that there is no correlation between the spatial (δE_i) and the temporal (δC_i) noise and no correlation between the noise and the data. The difference lies in the assumptions about the conditional probability of the noise and the additional assumption of Gaussianness in one of the models.

In the first case (Section 2.2.1) it is assumed that the noise is symmetric around the actual values. That is, given E' and C', the distribution of the noise $\delta E = E - E'$ and $\delta C = C - C'$ is assumed to be symmetric, but not necessarily Gaussian. In this case there is a downward bias, but only if there is a sizeable number of measurements.

In the second case (Appendix B), what is assumed is symmetry of the noise around the estimated values; given the known data E and C, there is a Gaussian probability distribution for the errors δE and δC. In this case there is a downward bias for any number of measurements.

2.2.1. Bias for noise symmetric around the actual values. First we explain why, in the case of very few measurements, the bias is an overestimate of the magnitude of $\mathbf{u}$.

This can be seen by considering the simplest linear system, one equation with one unknown. $Eu = C$ where $E \neq 0$. The ordinary least squares solution for this equation is the same as the equation obtained by simply dividing C by E.

Let

$$u' = \frac{C'}{E'},$$

where primes represent actual values. The estimated solution is

$$\frac{C}{E} = \frac{C' + \delta C}{E' + \delta E}.$$

The expected value $E(u)$ of the value of the estimated solution u is the expected value of

$$\frac{C'}{E' + \delta E}.$$

δE (and δC) are assumed to be symmetric in the sense that, for any real number s, the probability that $\delta E = s$ is the same as the probability that $\delta E = -s$.

Now temporarily make the special assumption that $|\delta E| = s < |E'|$. Then the expected value of u is just the expected value of

$$\frac{C'}{2}\left(\frac{1}{E' + s} + \frac{1}{E' - s}\right) = \frac{C'E'}{(E')^2 - s^2} \tag{10}$$

which is greater than the absolute value of the actual solution $C'E'/E'^2$ if $s \neq 0$.

If there is a large number of equations, there is a simple argument (see e.g., [38]) that the ordinary least squares solution is downward biased. This argument is essentially an asymptotic argument. The least squares solution is the ratio

$$\mathbf{u} = (E^t E)^{-1} E^t \mathbf{C}. \tag{11}$$

We have

$$\mathbf{u} = (((E')^t + (\delta E)^t)(E' + \delta E))^{-1}(E' + \delta E)^t(\mathbf{C}' + \delta \mathbf{C}). \tag{12}$$

If there is correlation between the temporal noise δC and the spatial noise δE, this correlation can affect the expected value of $\mathbf{u}$. If, however, the expected values of δC and δE are zero, and δC and δE are independent and also independent of E' and C', then the expected value of the least square solution is just the expected value of

$$(((E')^t + (\delta E)^t)(E' + \delta E))^{-1} E'^t \mathbf{C}'. \tag{13}$$

But this expression can be rewritten as

$$((E')^t E' + (\delta E)^t \delta E + (E')^t \delta E + (\delta E)^t E')^{-1} E'^t \mathbf{C}'. \tag{14}$$

The argument is that if there are enough equations then this last expression can be closely approximated by

$$((E')^t E' + (\delta E)^t \delta E)^{-1} E'^t \mathbf{C}'. \tag{15}$$

This is because terms of the form $(\delta E)^t E'$ are likely to be small if there are many equations in the system we are solving using least squares. The elements of this product matrix are of the form $\sum_i \delta E_{ij} E_{ik}$. If there are enough equations, these sums should be small if the expected value of $\delta E_{ij} = 0$.

We next remind the reader of a partial order defined on real matrices that generalizes the usual ordering of the real numbers. We use this partial order to define relative size.

Write $M > 0$ if M is a positive definite matrix. Write $M > N$ if $M - N > 0$ where M and N are two matrices, and similarly for $M \geq N$.

The order defined in this way generalizes the usual ordering of the real numbers and has some of the same properties. Thus if $M_1 > M_2$ and $N > 0$, $M_1N > M_2N$ and $M_1^{-1} < M_2^{-1}$. For any non-null vector V such that the indicated multiplications make sense, $\|M_1V\|_2 > \|M_2V\|_2$, where again $\| \ \|_2$ is the usual quadratic norm. References to the matrices M, N saying M is of greater size than N can be reinterpreted as meaning that if M^tM, N^tN are both nonsingular then $M^tM > N^tN$.

We can derive the result that the matrix $((E')^tE' + (\delta E)^t\delta E)^{-1}$ is smaller than the matrix $((E')^tE')^{-1}$. Multiplying by $(E')^tC'$, we get the result that $((E')^tE' + (\delta E)^t\delta E)^{-1}(E')^tC'$ is smaller than the actual solution. Thus the expected value of $\|u\|_2$ is smaller than the actual $\|u'\|_2$.

This argument for the case of ordinary least squares can also be applied to weighted least squares provided there is a sufficient degree of cancellation of the noise.

We can also say something about the direction of the bias. We assume we are using a gradient-based method only in order to simplify the description. Also, temporarily assume that there are only two (nonparallel) gradient directions in an image. Then even if the two directions are not orthogonal, it is easy to analyze the two-dimensional least squares problem as two one-dimensional problems. Assume that E and δE are written in a not necessarily orthogonal coordinate system in which the directions of the two axes are the two observed gradient directions. Then the direction in which there are more data (i.e., the direction in which E^tE is largest) is also the direction in which there is the largest signal to noise ratio (i.e., the largest ratio of E^tE to $\delta E^t\delta E$) But the effect of noise $\epsilon > 0$ on $1/(x^2 + \epsilon)$ is smaller the greater x is. So there is less bias in the direction where there are more data and more bias (i.e., underestimation) in the direction where there are less data, and thus there is a bias in the direction of the estimated flow.

There are not actually only two gradient directions, but all that really matters is the E matrix. Using a nonorthogonal coordinate system (i.e., rotating the two axes of the original coordinate system by different amounts), we can make E' diagonal, which is equivalent to assuming there were only two actual gradient directions. Similarly, we can assume that E is diagonal if we use a nonorthogonal coordinate system. If there are enough equation indices in (5), the two actual and the two observed gradient directions will be almost the same.

The interesting conclusion that can be drawn at this point is that since many common methods of computing optical flow essentially use ordinary least squares, there are many methods that will produce consistently biased results, and no Gaussianness assumptions are needed to derive that conclusion. More interesting, the bias is often an underestimation, and is smaller in the direction of more spatial gradients and greater in the direction of fewer gradients. It might be a good idea to try to estimate the amount of bias and then correct for it or perhaps employ a method more accurate than least squares in the first place. But it is not that easy to correct for the bias. One common technique used to correct for the bias is total least squares, but as we shall now see, this method has its own problems.

2.3. Total Least Squares

The problematic bias arose because of error in the E matrix and thus in the E^tE matrix. Let us rewrite (5) in the form

$$A_iu + B_iv - C_i = \delta A_iu + \delta B_iv - \delta C_i. \tag{16}$$

Even if u and v are known, there is no way of telling from the A, B, C data how the noise is apportioned among δA, δB, δC.

This is the main difficulty with total least squares: We have to know the relative amount of noise in the spatial and temporal errors. If the two spatial and the temporal variances of the noise are the same, then total least squares is approximately unbiased. Simulations have shown that if the spatial noise is larger than the temporal there is an underestimation. Otherwise, there is an overestimation. As discussed below, information about the noise ratios is difficult to compute; it can be obtained only from the change in flow between different regions.

Assuming again that the δC_i, δA_i, and δB_i are independent of each other and also of the data A, B, C,

$$\sigma^2(\delta C_i + \delta A_i u + \delta B_i v) = \sigma^2(\delta C_i) + u^2\sigma^2(\delta A_i) + v^2\sigma^2(\delta B_i). \tag{17}$$

Here $\sigma^2(x)$ is a function representing the variance of quantity x.

Given the total squared noise $N_i^2 = u^2(\delta A_i)^2 + v^2(\delta B_i)^2 + (\delta C_i)^2$, the most likely allocation of noise apportions the total noise so that

$$(\delta A_i)^2 = \frac{\sigma^2(\delta A_i)}{u^2\sigma^2(\delta A_i) + v^2\sigma^2(\delta B_i) + \sigma^2(\delta C_i)} N_i^2, \tag{18}$$

$$(\delta B_i)^2 = \frac{\sigma^2(\delta B_i)}{u^2\sigma^2(\delta A_i) + v^2\sigma^2(\delta B_i) + \sigma^2(\delta C_i)} N_i^2, \tag{19}$$

and

$$(\delta C_i)^2 = \frac{\sigma^2(\delta C_i)}{u^2\sigma^2(\delta A_i) + v^2\sigma^2(\delta B_i) + \sigma^2(\delta C_i)} N_i^2. \tag{20}$$

We choose N_i^2 so as to minimize $\sum_i N_i^2$ under the assumptions that $N_i^2 = u^2(\delta A_i)^2 + v^2(\delta B_i)^2 + (\delta C_i)^2$ and that (18) through (20) are valid. This is called the total least squares solution.

The total least squares solution can also be obtained by choosing δA_i, δB_i, δC_i so as to minimize

$$\sum_i \frac{\delta C_i^2}{\sigma^2(\delta C_i)} + \frac{(\delta A_i)^2}{\sigma^2(\delta A_i)} + \frac{(\delta B_i)^2}{\sigma^2(\delta B_i)}. \tag{21}$$

For an extensive discussion of the total least squares solution and of other methods involving errors in the variables, see [42]. An application of total least squares to optical flow computation using gradient techniques is found in [45].

Under the assumption the variances $\sigma^2(\delta A_i)$, $\sigma^2(\delta B_i)$, $\sigma^2(\delta C_i)$ are known, interesting simulation results and thoretical analysis of total least squares can be found in [16, 19]. What has been discovered is that the total least squares solution is approximately unbiased, but there is a trade-off: less bias but more variance.

To solve the total least squares problem we need to solve the problem of minimizing the expression

$$\frac{\sigma^2(\delta C_i)}{\sigma^2(\delta C_i) + \sigma^2(\delta A_i)u^2 + \sigma^2(\delta B_i)v^2} \sum_i (A_i u + B_i v - C_i)^2. \tag{22}$$

Even though this solution is approximately unbiased, the solution cannot be computed unless we know the variances of δA_i, δB_i, δC_i or at least know the ratio of the variances of the spatial and temporal noises. It is difficult to obtain this ratio. Even if we knew the actual constant flow we could not use the A, B, C data to determine how much of the blame for the fact that $A_i u + B_i v - C_i \neq 0$ is due to spatial noise and how much is due to temporal noise without additional assumptions. It is therefore going to be necessary to assume something rather questionable in order to obtain the necessary ratio of the spatial and temporal noises. In other words, we have to augment the model we have used so far.

If the flow is not constant then the amount of noise is a function of the flow. The variance of the total noise $A_i u + B_i v - C_i$ is $u^2\sigma^2(\delta A_i) + v^2\sigma^2(\delta B_i) + \sigma^2(\delta C_i)$. If $\mathbf{u}$ varied as a function of position, then if we knew the flow everywhere and if the statistics of the error in the A_i, B_i, C_i were independent of position, we could solve for the necessary variances. If we could obtain a reasonable, crude estimate not just of the flow but of the difference in flow between different patches of the image, and this crude estimate were fairly reliable, we could obtain rough estimates of the necessary ratios and then use these rough estimates to compute a total least squares solution.

However, there are many obstacles before us if we wish to assume we can use the methods of the previous paragraph to compute an approximately unbiased solution to the flow estimation problem. In regions where different objects are observed, or even in regions where different parts of the same object are observed and the texture properties differ greatly in different subregions, it is implausible to assume homogeneity of error statistics. In order to solve for the needed variances, we need to obtain a good estimate of the variance of $A_i u + B_i v - C_i$. We could use the estimate $1/(n-2)\sum_i (A_j u + B_j v - C_j)^2$, where n is the number of indices j for which the flow is approximately equal to that at index i. But unless n is large, this variance estimate can be noisy. If we assume the noise is Gaussian and the flow is exactly constant, we can treat $\sum_j ((\frac{(A_j u + B_j v - C_j)}{\sigma})^2)$ as having a χ^2 distribution with $n-2$ degrees of freedom where σ^2 is the actual variance of the noise $A_j u + B_j v - C_j$. But we do not know σ^2 or u or v. The actual variance may differ from the variance estimated from the limited set of data available. Using what we know about χ^2 distributions, we see that the standard deviation of $1/(n-2)\sum_j (A_j u + B_j v - C_j)^2$ is large unless the number of equations n is large.

There are still other sources of difficulty. We need substantial differences in flow between different regions of the image in order to be able to solve for the unknown variances of δA_i, δB_i, δC_i. Otherwise, we cannot disentangle these different variances. But if there is a substantial difference in flow, there may also be a substantial difference in the noise statistics of the different regions, so it may be difficult to compute the variance ratios we need.

Other methods can be used to determine the variances we need in order to compute a total least squares solution. But they also depend on questionable assumptions and noisy estimates. For example, one might assume that the variances of δA, δB, δC are proportional to the variances of A, B, C or proportional to some other easily obtained statistics of the data.

2.3.1. Confidence limits. If we can somehow obtain reasonable estimates of the variances, it is not too difficult to obtain a rough estimate of the variance of the error of the total least squares optical flow estimate. We have to apply a certain optimization condition in order to compute the total least squares solution. This involves solving a certain nonlinear equation. The nonlinear equation can be approximated by a linear equation. We know how

to estimate the variance of the error of a linear equation that has a unique solution. If $\mathbf{a} = K\mathbf{b}$ for a known matrix K, then if we know the statistics of $\mathbf{b}$, we know the statistics of $\mathbf{a}$. If we know the variance–covariance matrix of $\mathbf{b}$, we know the variance–covariance matrix of $\mathbf{a}$. What we need is a linear approximation to the nonlinear constraint that defines the total least squares solution; then we can write $\mathbf{u} = K\mathbf{C}$ for some matrix K, and this approximation must remain approximately correct (using the same matrix K) even if a small amount of noise is added to the data E, C. Or for simplicity, we can just use the ordinary least squares solution for the purpose of estimating the variance of the error in the flow estimate. The total least squares solution and the ordinary least squares solution will be different. But the variance estimate is rather crude anyway, and if the total least squares and ordinary least squares solutions are very different, most probably neither one is trustworthy.

2.4. Robust Techniques

Many of the essentially linear methods for obtaining flow estimates suffer from similar problems. In the form we have presented them so far, they are not very robust. A few outliers can greatly affect the computed result. Several robust methods have been developed to alleviate the problem [32]. It is difficult to analyze complex nonlinear methods. But a large number of these methods are subject to the same biases and inaccuracies as linear methods.

We next analyze what happens when we try to robustify the classical way of obtaining the value of (constant) flow using ordinary least squares. The ordinary least squares solution can be rewritten in a way that is very illuminating. This analysis also applies to total least squares. We just have to appropriately approximate the given total least squares problem by an ordinary least squares problem as discussed in Section 2.3.

As mentioned above, the least squares solution is the pair u, v that minimizes $\sum_i (A_i u + B_i v - C_i)^2$; hence we require

$$\sum_i A_i (A_i u + B_i v - C_i) = 0 \tag{23}$$

and

$$\sum_i B_i (A_i u + B_i v - C_i) = 0. \tag{24}$$

Thus

$$\sum_i \left(A_i^2 u + A_i B_i v - A_i C_i\right) = 0 \tag{25}$$

and

$$\sum_i \left(A_i B_i u + B_i^2 v - B_i C_i\right) = 0. \tag{26}$$

So assuming a unique solution exists, by Cramer's rule, $u = \frac{N}{D}$ is a ratio of two determinants.

$$N = \sum_i A_i C_i \sum_i B_i^2 - \sum_i A_i B_i \sum_i B_i C_i \tag{27}$$

$$= \sum_i \sum_j A_i C_i B_j^2 - A_i B_i B_j C_j \tag{28}$$

while

$$D = \sum_i \sum_j A_i^2 B_j^2 - A_i B_i A_j B_j. \tag{29}$$

Both N and D can be rewritten in ways that are more illuminating:

$$N = \sum_i \sum_j (A_i B_j)(C_i B_j - C_j B_i) \tag{30}$$

$$= \sum_i \sum_{j>i} (A_i B_j - A_j B_i)(C_i B_j - C_j B_i) \tag{31}$$

$$= \sum_i \sum_{j>i} (A_i B_j - A_j B_i)^2 \frac{C_i B_j - C_j B_i}{A_i B_j - A_j B_i} \tag{32}$$

and

$$D = \sum_i \sum_{j>i} (A_i B_j - A_j B_i)(A_i B_j - B_i A_j) \tag{33}$$

$$= \sum_i \sum_{j>i} (A_i B_j - A_j B_i)^2. \tag{34}$$

But the solution to the pair of equations

$$A_i u + B_i v - C_i = 0; \qquad A_j u + B_j v - C_j = 0 \tag{35}$$

is

$$u = \frac{C_i B_j - C_j B_i}{A_i B_j - A_j B_i}; \qquad v = \frac{C_i A_j - C_j A_i}{A_i B_j - A_j B_i}. \tag{36}$$

Hence the least square solution $\frac{N}{D}$ can be reinterpreted as a weighted average. For any pair of equations indexed by $i \neq j$, letting $D_{ij} = A_i B_j - A_j B_i$ and $N_{ij} = C_i A_j - C_j A_i$, we can solve the pair of equations for u and obtain the solution $u_{ij} = N_{ij}/D_{ij}$ provided the denominator is nonzero. So u is a weighted average of the u_{ij}'s with weights D_{ij}^2.

We have only given the equation for u_{ij} but a very similar equation for v_{ij} can be given. The result established here also applies to weighted least squares (even if there are negative weights): just replace the equation $E_i \mathbf{u} = C_i$ by $w_i E_i = w_i C_i$ where w_i is the square root of the weight of the ith equation; if the weight is negative, w_i will be imaginary.

Robust methods somehow combine the local $\mathbf{u}_{ij}$'s. One can think of many robust methods. For example, one can compute the minimum-volume ellipse containing most of the weight of the (u_{ij}, v_{ij}). Or, instead of considering all $i \neq j$ to solve for $\mathbf{u}$, one can use some sample, for example, by picking the pairs with large D_{ij}^2's.

It is difficult to provide a general analysis of the conditions for bias in robust methods. We know that the least squares solution $\mathbf{u}$, which is a weighted average of the local $\mathbf{u}_{ij}$'s is biased. Many robust methods can also be understood as averages of the $\mathbf{u}_{ij}$'s using different weights. The point is that to choose the right weights, which would let us avoid the bias, the statistics of the noise have to be known. Thus by the same argument that the noise parameters often cannot be estimated well, robust methods like the ones discussed above are biased.

2.5. *Nonconstant Flow*

Reference has already been made to the possibility that the flow is not constant. It was assumed that the flow is locally constant, but even that is not plausible. It is more likely that the flow is a simple function of position, perhaps approximately linear or approximately quadratic. If depth is constant, flow is an approximately quadratic function of position. More generally, the flow $\mathbf{u}$ can be decomposed into a linear combination of basis flows $\mathbf{w}_i$ where each $\mathbf{w}_i$ is a known function of position and $\mathbf{u} = \sum_i u_i \mathbf{w}_i$ for some unknown coefficients u_i. We can still use the fact that $\frac{DI}{Dt} = 0$ to obtain a linear equation $\sum_j E_{ji} u_j - C_i = 0$ for the unknown u_j. We still have the possibility of applying a local smoothing operator before employing the principles that corresponding points have the same attribute value. We can still take Fourier transforms. We still have the same problems with bias in the ordinary least squares solution, and we can still obtain a rough estimate of variance using this solution.

A problem with this discussion is that if the flow is not constant, the result of applying a smoothing operator to $A_i u$ is not the same as u times the result of applying the smoothing operator to A_i. But if the basis vectors $\mathbf{w}_j$ are known, we can compute in advance, for any vector $\mathbf{z}_j$ of coefficients, the effect of applying the smoothing operator to $\mathbf{w}_j \cdot \mathbf{z}_j$.

In our framework, we cannot easily handle sharp discontinuities in the flow field if the locations of these discontinuities are not known. But we can model more than simple linear decomposition of the flow.

One possibility is to define plausible a priori models of how equation error varies with the index of an equation. If the indices i represent points where flow is observed, and we wish to obtain the value of the flow in the vicinity of some point p, it is plausible that the further a point p_1 is from p, the less likely it is that p and p_1 have the same flow. Thus instead of computing a least squares solution to $A_i u + B_i v - C_i = 0$, we should compute the solution to $k_i(A_i u + B_i v - C_i) = 0$ where k_i is a weight dependent on the distance between p and the point p_1 that is indexed by i. This procedure will, in effect, give us a variety of smoothed flow values. Different weighting functions can be used for smoothing over different size regions. We will still get consistent underestimates using the ordinary least squares solutions under certain conditions. In order to get reasonably reliable flow estimates, we will want the data from many different equations to substantially influence the computed solution, but that will produce a tendency for $\delta E^t E'$ to be small compared to $E'^t E'$ and thus give us underestimates.

Other linear methods can be treated within our framework. It might be better to more explicitly model the random point-to-point variations in flow. Then the equation $A_i u + B_i v - C_i = 0$ fails to be exactly true not only because of noise in the data A_i, B_i, C_i but also because of "noise" in the u, v. That means that the flow $\mathbf{u}$ consists of a constant regional flow $\mathbf{u}_r$ to which is added noise $\mathbf{u}_n$ which is zero-mean. This noise is assumed to be independent of the noise in the A_i, B_i, C_i and it is also assumed that the $\mathbf{u}_n$'s of distinct points are independent and identically distributed. The amount of variance in $A_i u + B_i v - C_i$ due to random point-to-point variations in flow is $A_i'^2 \sigma_2^2 + B_i'^2 \sigma_2^2(v)$. Here $\sigma_2^2(u)$, $\sigma_2^2(v)$ represent the variances of the point-to-point flow variations.

Assume for the sake of simplicity that there is no noise in the A or the B data. Then under the assumption that the noise is Gaussian, since the weights should be inverse to the variances, the maximum likelihood solution for $\mathbf{u}$ is the ordinary least squares solution of

$$\frac{\sigma_2^2(C_i)}{\sigma_2^2(C_i) + A_i^2 \sigma_2^2(u) + B_i^2 \sigma_2^2(v)}(A_i u + B_i v - C_i) = 0. \tag{37}$$

In order to solve this, it is necessary to know the variance of the noise in C_i and the magnitude of the variance of the noise in the flow components u, v. Rough estimates of these quantities can be made if there is a rough estimate of the value of the flow and hence a rough estimate of how the size of $A_i u + B_i v - C_i$ varies as a function of A_i and B_i. A problem is the noisiness of this method of inferring the size of the variances in u, v from the data. Even if we have good estimates of the variances, we still have the same problem of bias as before when we neglect $\delta A_i \delta B_i \neq 0$.

2.6. General Remarks on Gradient and Frequency-space Methods

We have seen how difficult it is to estimate the noise parameters using the limited information available to us. This is true for several possible models of the noise. The result is inevitable bias.

If we had a larg amount of data for which the noise parameters were fixed, it would be easy to closely approximate the noise parameters. But the noise parameters do not stay fixed long enough. Sensor characteristics may stay fixed, but there are many other sources of noise besides sensor noise. Different lighting conditions, different physical properties of the objects being viewed, and different orientations of the viewer in 3D space all result in different amounts of noise. Aside from all these factors, in order to estimate derivatives or to compute Fourier transforms, we need to interpolate. The accuracy of interpolation can depend in complex ways on the pattern of intensities in the image.

3. CORRELATION METHODS

We next discuss a model for correlation methods which also gives bias. Classical correlation methods find the $\mathbf{u}$ that maximizes the correlation between $I(p, t)$ and $I(p + \mathbf{u}, t + 1)$ where this correlation is computed over a large fragment of the image. The point p is an arbitrary point in 2D image space and t represents time. If there is constant flow equal to $\mathbf{u}$ and corresponding points have the same intensity, then this correlation should be perfect.

To simplify matters, we assume that the measure of correlation is additive, so that the correlation is just $\sum_{p,t} g(I(p, t), I(p + \mathbf{u}, t + 1))$ for some function g. For example, the correlation might be measured by the covariance if we could safely ignore the fact that the errors in the estimates of the value of I at nearby points are not independent. This assumption is comparable to the assumption in gradient-based methods that the errors in (5) at different i's are independent. For simplicity we will assume that g is a quadratic function.

Nothing changes very much if instead of assuming that corresponding points have the same I, we allow for slowly changing I and requiring that I at time $t + 1$ be a linear function of I at time t. Then we also have to pick appropriate coefficients for the linear function and compare the actual I at a point Q corresponding to point P with the predicated I. The prediction is based on a linear predictor with unknown coefficients. We can solve for the unknown coefficients using least squares.

There are rather annoying artifacts due to gridding. If in computing the correlation we only sum over (p, t) on the grid, the correlation would be affected by how near the points $(p + \mathbf{u}, t + 1)$ are to points on the grid. This is because in order to compute the values of I at points off the grid we need to interpolate. The interpolation entails a kind of smoothing that cleans up some of the noise and hence increases the correlation; we need to do some

smoothing to evaluate I at points half-way between two grid points, but much less smoothing to evaluate I near the grid points. In general, interpolation involves different amounts of noise-smoothing at different points, but we ignore this issue. We assume either that we are only working with points on a grid or that we have designed an interpolation scheme and a correlation measure that do not suffer from anomalies due to gridding, we deliberately add random noise to the interpolated intensity values to counteract the problem, or we compute correlations using only points on the grid (and if we want to obtain subpixel accuracy estimates of flow, we interpolate the correlations).

We analyze the bias only in the case of constant flow. The same kind of analysis could be applied more generally, and roughly the same result would be obtained, but the notation would have to be more complex. We will also not explicitly handle the case where what must match at corresponding points is not I, but some linear function of I.

The essence of our claim is that if the noise values at different points are correlated so that the closer two points are in space-time the greater the correlation, then there will be a bias toward underestimation. The same bias will arise if we assume that the observed I is obtained by smoothing the actual values of I in some space-time neighborhood and then adding noise.

Let $\mathbf{u}'$ be the actual flow, and let the time interval between the images be one unit. To simplify the analysis we consider only one space dimension.

We assume that the observed I is equal to the actual I' plus δI, a noise term. The δI's at different points are not uncorrelated. We define a Euclidean distance metric in space-time and assume that the smaller the distance between two points, the greater the correlation of the δI's found at the two points. Then if corresponding points have the same intensity, the difference between the intensity observed at $(x + u + \delta C, 1)$ and the intensity observed at $(x, 0)$ is the sum of two terms. One term is the difference between the actual $I'(x, 0)$ and $I'(x + \delta x, 0)$ (note that $I'(x + \delta x + u, 1) = I'(x + \delta x, 0)$. The other term is the difference between the δI at $(x + u + \delta x, 1)$ and the δI at $(x, 0)$. Assuming that the noise is independent of the actual I',

$$\sum (I(x + u + \delta x, 1) - I(x, 0))^2$$
$$= \sum (I'(x, 0) - I'(x + \delta C, 0))^2 + \sum (\delta I(x + u + \delta x, 1) - \delta I(x, 0))^2 + \phi. \quad (38)$$

Here $\phi = 2\sum (I'(x, 0) - I'(x + \delta x, 0))(\delta I(x + u + \delta x, 1) - \delta I(x, 0)))$ and has expected value zero. Thus it can be ignored when analyzing bias, because the noise is independent of the actual I', and ϕ is the summation of many terms which will tend to cancel each other out so that ϕ will tend to be small compared to the other two terms. If we compute the correlation over a large enough region so that boundary effects can be ignored, and we pick a value s, the first term on the right-hand side of the above equation will be the same if we change the value of δx from $s > 0$ to $-s$, but the second term will be smaller because of the correlation pattern of the noise. This will result in bias. A preference for $u - s$ over $u + s$ will arise.

If there are many sharp gradients in I' in the x-direction, the first term will be fairly large unless δx is small, so the bias will be less significant.

The above was only a 1D motion analysis. But the only additional complexity in the 2D case is the notation; there will still be the same kind of bias. A slight generalization of the argument shows that in the 2D case, if there is bias, it is smaller in the direction in which there is more data; the result is a bias in the orientation of the flow.

3.1. The Effect of Smoothing

What if our model of noise does not apply and the local estimates are noisy but unbiased? Even so, local correlation estimates of flow are likely to have a lot of error, and they need to be smoothed to give sensible results. In fact, the most frequently used correlation methods (such as those discussed in [3, 35]) compute this correlation over small areas of support and then apply smoothing to obtain the optical flow. These methods, because of the smoothing, will be biased.

The simplest kind of smoothing assumes constant flow and uses least squares to smooth the local estimates; indeed, it computes a weighted average. More sophisticated smoothers perform a kind of regularization and suffer from the biases discussed in Section 2.5; they still do not avoid bias.

4. AN EXPLANATION OF OPTICAL ILLUSIONS

4.1. The Model

We use a gradient-based method; simple least squares estimation; and additive, identically, independently distributed, symmetric noise. This is the model studied in Section 2.2.1; an asymptotic proof of the bias was given there.

To have a notation for the estimated flow which allows us to give a detailed explanation and which also shows the bias for a smaller number of measurements, we develop the least squares solution in a Taylor expansion.

In (5), let the variables A, B, C be the spatial and temporal derivatives of the image intensity function. Then this equation is the optical constraint equation. The estimated values (A_i, B_i, C_i) consist of the actual values (A_i', B_i', C_i') and the additive noise $(\delta A_i, \delta B_i, \delta C_i)$. The expected values of the first-order terms are zero and the expected values of the second-order terms are given by the covariance matrix

$$E((E\mathbf{C})^t(E\mathbf{C})) = \begin{pmatrix} \sigma_s^2 & & \\ & \sigma_s^2 & \\ & & \sigma_t^2 \end{pmatrix}.$$

The expected values of higher order terms are assumed to be negligible.

In Appendix C the expected value of the estimated flow (18) is developed in a second-order Taylor expansion at zero noise; it converges in probability to

$$\operatorname*{plim}_{n\to\infty} E(\mathbf{u}) = \mathbf{u}' - n\sigma_s^2 M'^{-1}\mathbf{u}', \tag{39}$$

where $M' = E'^t E'$, the matrix of exact spatial gradient values, and n is the number of measurements. Formula (39) is well known [33] and could also be derived from (15).

This formulation allows for easy interpretation of the effects of the gradient distribution on the bias of the computed flow, as all the information is encoded in the matrix M'. In the case of a uniform distribution of the image gradients in the region where the flow is computed, M' (and therefore M'^{-1}) are multiples of the identity matrix, leading to a bias solely in the length of the computed optical flow; there is an underestimation. In a region where there is a unique gradient vector, M' will be of rank 1; this is the aperture problem. In the general case the bias can be understood by analyzing the eigenvectors of M'. As M' is a

real symmetric matrix, its two eigenvectors are orthogonal to each other. M'^{-1} has the same eigenvectors as M' and inverse eigenvalues. The direction of the eigenvector corresponding to the larger eigenvalue of M'^{-1} is dominated by the normal to the major orientation of the image gradients, and the product of M'^{-1} with vector $\mathbf{u}'$ is most strongly influenced by this orientation. Thus there is more underestimation in the direction of fewer measurements and less underestimation in the direction of more measurements. The estimated flow therefore is biased downward in size and biased toward the major direction of the gradients (that is, toward the eigenvector corresponding to the larger eigenvalue of M').

4.2. Dissection

Figure 2 displays the expected values of the noise terms for the gradient distribution that occurs in one of the regions of the Ouchi illusion shown in Fig. 1 with blocks four times longer than they are wide. The image gradients are in two orthogonal directions with four times as many measurements in one direction as in the other. The plots show the change in the bias as the angle between the gradients and the true flow direction varies. The angle θ is measured between the positive x-axis and the direction of more gradients; the other gradient direction is at an angle $\theta + \pi/2$ with the positive x-axis (see Fig. 2a). Figures 2b and 2c show the expected errors in length and angle. The plots are based on the exact second-order Taylor expansion given in Appendix C.

For such a gradient distribution the bias can be understood rather easily. The eigenvectors of M'^{-1} are in the directions of the two gradient measurements with the larger eigenvalue

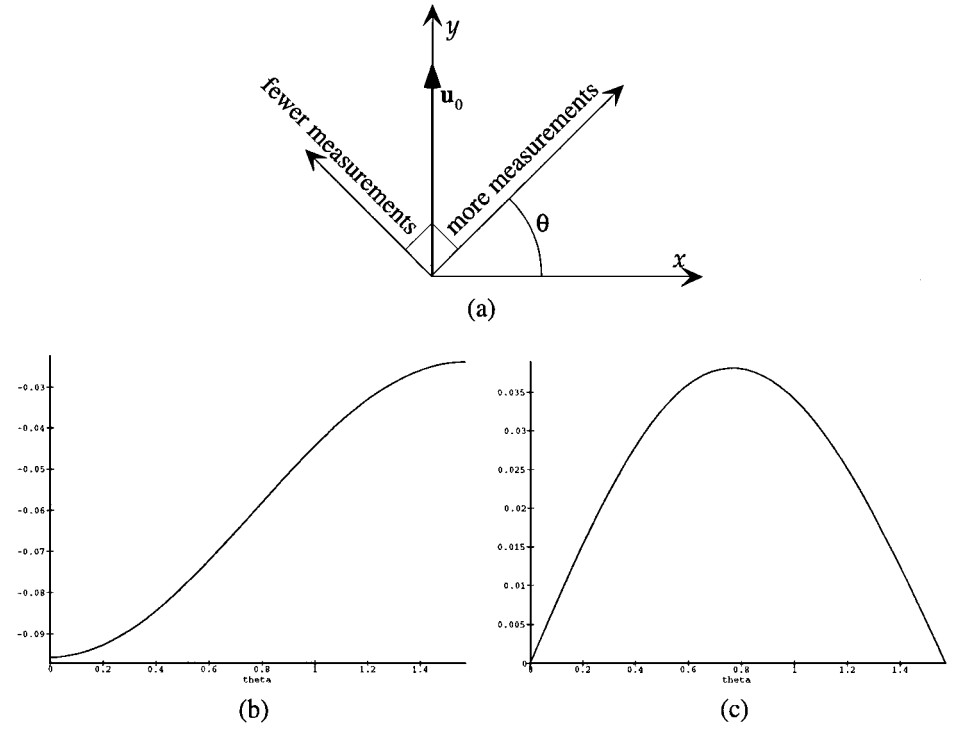

FIG. 2. (a) Sixteen measurements are in the direction making angle θ with the positive x-axis and four measurements are in the direction $\theta + \pi/2$. The optical flow is along the positive y-axis and of length 1. (b) Expected error in length. (c) Expected error in angle measured in radians between the expected flow and the actual flow. The error has $\sigma_s = 0.15$.

corresponding to fewer gradients. As $\mathbf{u}' = (0, 1)$, the noise term in (39) leads to a bias in length as shown by the curve in Fig. 2b, which has its minimum at 0 and its maximum at $\pi/2$ (that is, when $\mathbf{u}'$ is aligned with the major gradient direction). The error in angle is greatest for $\pi/4$ (that is, when $\mathbf{u}'$ is exactly between the two eigenvectors of M'^{-1}) and it is 0 for 0 and $\pi/2$ (Fig. 2c). Overall, this means the bias is largest when the major gradient direction is normal to the flow and is nearly eliminated when it is aligned with the flow (that is, in the Ouchi pattern, when the long edge of the block is perpendicular to the motion). The bias for angles θ between $\pi/2$ and π is obtained from the above plots by reflecting the curves in $\pi/2$ and changing the sign of the error in the angle.

Let us now use these graphs to discuss the Ouchi illusion. In the Ouchi pattern, the relative angles between the real motion and the predominant gradient direction differ in the inset and the surround, so the regional velocity estimates are biased in different ways. When, instead of freely viewing the pattern of Fig. 1, the page is moved in different directions, we observe that the illusory motion of the inset is mostly a sliding motion orthogonal to the longer edges of the rectangle and in the direction whose angle with the motion of the paper is less than 90°. Using Fig. 2, it can be verified that for all angles the difference between the error vector in the inset and the error vector in the surrounding area (or, equally, the estimated flow vectors) projected on the dominant gradient direction of the inset is in this direction. For example, when the motion is along the first meridian (to the right and up), the error in the inset is found in the graph at angle $\theta = \pi/4$ and in the surround at $\theta = 3\pi/4$. The two error vectors are of the same length, each toward the gradients of the longer edges, and the projection of the resulting difference vector is to the right. If the motion of the paper is to the right, the difference in error vectors is due to length, resulting in a perceived motion to the right. If the motion of the paper is upward, the difference vector is downward; its projection on the major gradient direction of the inset is zero and thus hardly any illusory motion is perceived. Figure 3 shows, for a set of true motions, the biases in the perceived motion.

We assume that in addition to computing flow, the visual system also performs segmentation, which is why a clear relative motion of the inset is seen. When experiencing the Ouchi illusion under free viewing conditions, the triggering motion is due to eye movements, which can be approximated through random, fronto-parallel translations. Since the difference in the bias vectors of the inset and surround has a significant projection on the dominant gradient direction of the inset for a large range of angles (that is, directions of eye movements), the illusion is easily experienced.

The Ouchi pattern is an ideal setting for demonstrating the bias. First, the gradient distribution in the pattern is such that the bias is highly pronounced. Second, the 3D motion of the observer relative to the pattern (which is either due to random eye movement or the jiggling motion of the paper) changes rapidly. This makes temporal integration of measurements very difficult, and thus the system cannot acquire enough data to learn the noise parameters.

In [13] it has been proposed that the bias in flow estimation due to errors in the spatial and temporal image derivatives also accounts for the findings of some studies using variations of the original Ouchi patterns and for a number of studies on the perception of moving plaids, in particular studies which reported a misperception in the estimated velocity of the plaid.

The erroneous estimation of image velocity in plaids has been given another explanation based on Bayesian modeling [46]. This explanation is based on the assumption that there is an a priori preference for small flow values. It is easily understood that this preference

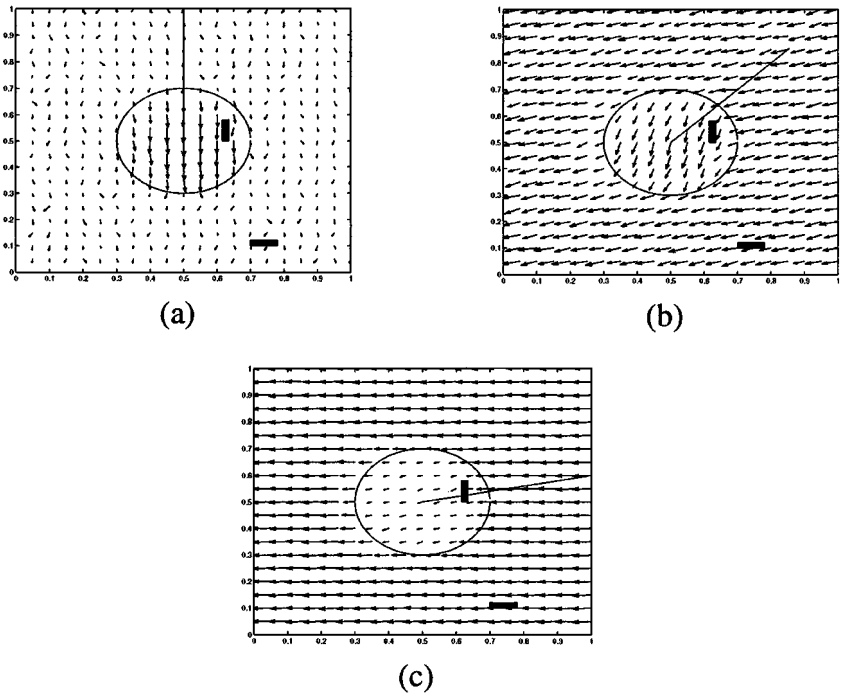

FIG. 3. The regional motion error vector field. The vectors shown are the differences between the true motion and the calculated motion. To derive the sliding motion, compute the difference between the error in the inset and the error in the surround, and project the resulting vector on the dominant gradient direction in the inset. The line from the center is the direction of the true motion. The noise is Gaussian and the spatial gradient magnitude is 1. In (a) and (b), $\sigma_s = \sigma_t = 0.1$; in (c), $\sigma_s = \sigma_t = 0.2$.

results in an increase in the a posteriori probability of small flow values and thus in a bias toward underestimation. Thus, in the Bayesian model the bias is in effect assumed, whereas in our model it is not.[2] It is true that most quantities in nature are more often small than large. Spatial derivatives of intensity, temporal derivatives of intensity, curvature, and many other visual quantities tend to be small more often than they are large. This is the basic justification of the smoothness assumptions that we often use. But why should a system prefer to estimate small flow values? Even if large flow values are rare, it is especially important to quickly detect them when they occur. This leads us to doubt that a Bayesian model which ignores the utility of flow information properly reflects the biological visual system.

5. COMPUTATIONAL MODELS OF MOTION PROCESSING

The current view which dominates modeling in both computer vision and biological vision is that the computation of optical flow is accomplished prior to any other computations involving image motion. First the optical flow is computed on the basis of 2D image information only; then it is used to compute 3D space and time interpretations, such as 3D motion estimation, segmentation, and shape estimation.

[2] Non-Bayesian methods also implicitly assume priors. Least-squares methods assume that a priori all solutions have equal probability.

We have seen, however, that estimation of optical flow entails computational problems. The estimation of optical flow requires that the data from each image region be aggregated, and this makes it inseparable from the detection of discontinuities (which are due to objects at different depths or differently moving scene elements). Without knowing the locations of discontinuities, it is hard to estimate flow there, but in order to detect the discontinuities, information about optical flow within their neighborhoods is needed. Noise in the estimates makes the problem even more difficult. As shown in the preceding analysis, for statistical reasons it is very difficult to obtain accurate optical flow estimates even within areas of smoothly changing flow. Theoretically, to achieve good flow estimates, very accurate estimates of the noise parameters are needed; but in order to estimate the noise well, motion information from large neighborhoods has to be integrated, and this requires detailed models of the flow field. The only way to obtain such models is from additional information about the dynamic scene; this includes knowledge about the discontinuities, the shapes of the visible surfaces, and the 3D motion.

This suggests that instead of following a two-step approach, which separates optical flow estimation from scene interpretation, new models should be developed that combine these processes. To obtain such models we might use a priori 3D constraints to improve our estimates of the 2D flow. These might be constraints on surface shape or 3D motion parameters. For example, we might seek the 2D flow that is consistent with the motion being rigid in some region and that minimizes some measure of curvature or some function of the derivatives of depth [5]. Or we might obtain useful a posteriori information about the depths or 3D motions of the objects in the scene using cues other than flow. Direct translation of these 3D constraints into constraints on the flow (or its derivatives) might be easy. If not, we could work explicitly with the 3D information.

Any such computational model has to consider the information exchange between the different processes. For example, we can envision an architecture which carries out the computations in a feedback loop. First, we estimate the image velocity by combining normal flow measurements. These estimates do not necessarily have to be quantitative, but could take the form of qualitative descriptions of local flow field patches or bounds on flow values. The flow computed in this way is used to obtain partial depth estimates and perform discontinuity detection; at the same time, an estimate of 3D motion can be derived. The computed 3D information can then be fed back and utilized together with the image measurements to obtain better flow estimation, discontinuity localization, and improved 3D motion and structure estimation.

However, even when we use the best computations, we cannot guarantee that optical flow will be estimated accurately all the time, and this has to be taken into account in visual navigation. Most computational models assume generically computed flow which is used for obtaining accurate 3D motion and for metric shape estimation. Consideration of the computational difficulties of this approach calls for a more purposive approach. Depending on the particular computation of 3D information, different representations of flow may be useful. For example, instead of attempting accurate egomotion estimation from optical flow, the approximate directions of translation and rotation can easily be obtained from patterns of the sign of the normal flow [10]. Instead of reconstructing the scene in view, it is computationally more feasible to derive less powerful shape representations sufficient for particular tasks, for example representations which only describe the qualitative shapes of scene patches, or we can obtain an ordering of surface patches with respect to their depth values [11, 12]. Also, instead of attempting segmentation directly

from image measurements, segmentation may be performed only in conjunction with other tasks.

6. CONCLUSIONS

This paper has analyzed the statistics of optical flow estimation. Noise in the data poses serious problems for the estimation of flow. The reason is that noise affects both the spatial and temporal components of the image measurements (that is, both the direction and the length of one-dimensional velocity measurements). To estimate flow well, the noise parameters need to be estimated accurately. In many situations this is impossible because the parameters are not static, but change with the viewing and lighting conditions, often too rapidly to collect enough data to obtain good estimates.

An unfortunate consequence of the unknowability of the noise parameters is bias in the flow estimates. Many flow estimation techniques have been analyzed here, including gradient-based, energy-based, and correlation methods. It was found that most techniques produce consistent bias; the estimates tend to be smaller in length and closer in direction to the dominant gradient direction in the patch than the actual values. A bias of this form also provides an explanation for the illusory motion perceived by humans when viewing the Ouchi pattern and for erroneous estimates in the perception of plaid motions.

Although it has long been known that the estimation of optical flow is a very difficult problem (and if formulated in the classic way, an ill-posed one), this paper for the first time points out one of its inherent computational problems. The point of our study has been to argue for a reevaluation of the role of flow estimation in 3D motion processing. Optical flow estimation should not be carried out in isolation but in conjunction with the higher-level processes of 3D motion and scene interpretation. This way of looking at the "motion pathway" [48] might stimulate new research on structure from motion.

APPENDIX A

Conservation of Phase

Assuming constant local phase, let $G_1 I$, $G_2 I$ represent the real and complex parts of some local Fourier transform. Define the ratio $G_1 I / G_2 I$ component-wise so that

$$\frac{G_1 I}{G_2 I}(\omega_x, \omega_y, \omega_t, x, y, t) \equiv \frac{G_1 I(\omega_x, \omega_y, \omega_t, x, y, t)}{G_2 I(\omega_x, \omega_y, \omega_t, x, y, t)}.$$

It is not assumed that $\frac{DI}{Dt} = 0$ but rather that

$$\frac{D\frac{G_1 I}{G_2 I}}{Dt} = 0. \tag{A.1}$$

Using the usual rule for differentiation of a quotient, we obtain

$$\frac{G_2 I \frac{DG_1 I}{Dt} - G_1 I \frac{DG_2 I}{Dt}}{(G_2 I)^2} = 0, \tag{A.2}$$

Where $(G_2 I)^2$ is a product that is defined component-wise.

Expressions such as $DG_jI/(Dt)$ can be rewritten in terms of partial derivatives:

$$\frac{DG_jI}{Dt} = \frac{\partial G_jI}{\partial x}u + \frac{\partial G_jI}{\partial y}v + \frac{\partial G_jI}{\partial t}.$$

Hence (A.2) can be rewritten in the form given by (5).

The only remaining problem is the noise analysis of (A.2). We apply (A.2) to one particular sextuplet of independent variables $(\omega_x, \omega_y, \omega_t, x, y, t)$ and analyze the noise there. Assume that G_1I, G_2I are known with reasonable accuracy, but that the derivatives are hard to estimate. The derivatives of both G_1I and G_2I are noisy. Let the variance in the noise of $DG_jI/(Dt)$ be $\sigma^2_{G_j}$. Let us also assume that the DG_1I/Dt noise and the $DG_2I/(Dt)$ noise are not correlated with each other. Then the variance in the left-hand side of (A.2) is $(\sigma^2_{G_1}/(G_2I)^2) + ((G_1I)^2\sigma^2_{G_2}/(G_2I)^4)$. As a crude approximation, one might say that the variance in the left-hand side of (A.2) is inversely proportional to the square of the amplitude (i.e., directly proportional to $((G_1I)^2 + (G_2I)^2)^{-1}$). (Assume $\sigma^2_{G_1} = \sigma^2_{G_2}$. Then the noise is directly proportional to a quotient whose denominator is $(G_2I)^4$ and whose numerator is just the squared amplitude of the G-transform.)

One might apply (A.2) to estimate the flow in the vicinity of some point (x_0, y_0, t_0). In that case we fix the spatial parameters so that $x = x_0, y = y_0, t = t_0$ but $\omega_x, \omega_y, \omega_t$ can vary through all possible frequencies. It makes sense to assume that the errors in the observation of DG_jI/Dt at the different frequencies are independent. It also makes sense to assume that $\sigma^2_{G_1}, \sigma^2_{G_2}$ is independent of frequency. So in computing the weighted least squares solution to (5), the weights need to be inversely proportional to the variances [24] and thus directly proportional to the squared amplitude.

APPENDIX B

Bias if the Noise is Gaussian

Here we present a nonasymptotic argument for the statement that ordinary least squares estimates tend to be underestimates. We assume that the different noises δA_i, δB_i, δC_i are independent and identically distributed. We also need to assume that the probability distribution of the noise given the known data is Gaussian.

First consider the one-dimensional case, so that all the $B_i = 0$. To simplify matters further, assume that $\delta C_i = 0$ for all i and $A_i = 1$ for all i. Then the estimated value of u is just the ordinary average:

$$u = \frac{\sum_i C_i}{\sum_i 1}. \tag{B.1}$$

If the actual value of flow in the x-direction is u', then

$$A'_i = \frac{C_i}{u'}, \tag{B.2}$$

so that

$$A'_i - A_i = \frac{C_i - u'}{u'}. \tag{B.3}$$

We want to say something about the probability that the actual flow in the x-direction has some value u' given that A and C have the values they do. It will be convenient to argue not directly in terms of probability, but instead in terms of an energy function. If Pr is the probability of some event, then by definition the energy T is given by the relation $Pr = k_1 e^{-k_2 T}$ where k_1, k_2 are constants of no interest to us.

We know that the energy can be written as

$$\sum_i \left(\frac{C_i - u'}{u'} \right)^2 \right) \tag{B.4}$$

$$= \sum_i \frac{(C_i - u)^2 + (u - u')^2 + 2(u - u')(C_i - u)}{(u')^2}. \tag{B.5}$$

But $\sum_i (C_i - u) = 0$ by definition of the average. So the energy function considered as a function of u' is a monotonically increasing function of $((T + (u - u')^2)/((u')^2))$ where $T = \frac{1}{n} \sum_i (C_i - u)^2$ and T does not depend on u'. Here n is the number of indices i over which we are summing. So if coordinates are chosen such that $u > 0$, then for $s > 0$, the energy function is less if $u' = u + s$ than if $u' = u - s$, and thus it is more likely that $u' - u = s$ than that $u' - u = -s$. In other words, the estimate is more likely to be too small than too large.

The same argument works if we remove the constraint that $A_i = 1$. In this case the energy function is

$$\sum_i \left(\frac{C_i - A_i u'}{u'} \right)^2 = \sum_i \left(\frac{(C_i - A_i u)^2 + (A_i u - A_i u')^2 + 2A_i (u - u')(C_i - A_i u)}{(u')^2} \right). \tag{B.6}$$

By definition of the least squares solution, $\sum_i A_i C_i = \sum_i A_i A_i u$. The energy function is now a monotonically increasing function of $((T + \tau^2 (u - u')^2)/((u')^2))$ where T and τ are expressions that do not depend on u', and we can draw the same conclusion as before.

Now consider the problem in two dimensions. Choose coordinates so that $v = 0$, and assume that $\delta C = \delta B = 0$. Then we must have

$$u' = \frac{C_i - B_i v'}{A_i'} \tag{B.7}$$

and

$$A_i' = \frac{C_i - B_i v'}{u'}. \tag{B.8}$$

Thus the energy function is

$$\sum_i \left(\frac{C_i - B_i v' - A_i u'}{u'} \right)^2$$

$$= \sum_i \frac{(C_i - B_i v' - A_i u)^2 + A_i^2 (u - u')^2 + 2A_i (u - u')(C_i - B_i v' - A_i u)}{(u')^2} \tag{B.9}$$

Taking into account the fact that $\sum_i A_i C_i = \sum_i A_i^2 u$, we can see that the energy function can be written in the form $(T + \tau^2 s^2 + Q v' s)/((u')^2)$ where $s = u - u'$ and T, τ, Q are

not dependent on u', v'. So let us fix the value of $|s|$ and the value of $|v'|$ and again choose coordinates so that $u > 0$ and take advantage of the fact that the expression for the energy is simple.

The energy function is a fraction. There are two possible values for the numerator: $T + \tau^2 s^2 + |Qv's|$ and $T + \tau^2 s^2 - |Qv's|$. Let us call these two values $q_1 > q_2$. There are also two possibilities for the denominator—call them $r_1 > r_2 \geq 0$. If $s < 0$, the energy can be q_1/r_1 or q_2/r_1. If we replace the denominator r_1 by r_2, we obtain the possibilities that arise when $s > 0$, but if we replace the denominator r_1 by r_2, we increase the energy; therefore, it is less probable that s is positive than that it is negative. This argument assumed the values of $|s|$, $|v'|$ to be fixed, but it does not matter what they are equal to. Hence more likely than not s is negative and we have an underestimate.

Let us generalize still further by allowing $\delta B_i \neq 0$. Then we can let the δB_i be arbitrary. We have

$$A_i' = \frac{C_i - B_i' v'}{u'}. \tag{B.10}$$

Thus the energy function is

$$\sum_i \nu(\delta B_i)^2 + \left(\frac{C_i - B_i' v' - A_i u'}{u'}\right)^2 = \sum_i \nu(\delta B_i)^2 + \sum_i \frac{(C_i - B_i' v' - A_i u)^2 + A_i^2 (u - u')^2 + 2A_i(u - u')(C_i - B_i' v' - A_i u)}{(u')^2}. \tag{B.11}$$

Here ν is a weight that depends on the variance of the noise in the y-direction. Choose three quantities, k_1, k_2, and K_3. The first two quantities are scalars and we require $|s| = |k_1|$ while $|v'| = |k_2|$. Here, as before, $s = u' - u$. K_3 is a vector and we require that either $\delta B = K_3$ or $\delta B = -K_3$. An argument similar to that given in the next to last paragraph shows that it is less likely that $s > 0$ than that $s < 0$, and this argument depends in no essential way on the particular values k_1, k_2, K_3 chosen.

The final generalization allows there to be noise in the C_i. We have

$$A_i' = \frac{C_i' - B_i' v'}{u'}. \tag{B.12}$$

Thus the energy function is

$$\sum_i \nu_1(\delta B_i)^2 + \nu_2(\delta C_i)^2 + \left(\frac{C_i' - B_i' v' - A_i u'}{u'}\right)^2 = \sum_i \nu_1(\delta B_i)^2 + \sum_i \nu_2(\delta C_i)^2 + \sum_i \frac{(C_i' - B_i' v' - A_i u)^2 + A_i^2 (u - u')^2 + 2A_i(u - u')(C_i' - B_i' v' - A_i u)}{(u')^2}. \tag{B.13}$$

Here ν_1, ν_2 are weights. Again use the fact that $\sum_i A_i C_i = \sum_i A_i^2 u$. We see that the energy function can be written in the form

$$R + \frac{T + \tau^2 s^2 + Qv's + \left(\sum_i A_i \delta C_i\right)s}{(u')^2}. \tag{B.14}$$

Here R, T, τ, Q do not depend on u', v'. Again an argument that temporarily fixes the values of certain quantities can be used. Namely, we need to fix the values of $|s|$, $|v'|$, and find noise vectors K_3, K_4 and require that either $\delta B = K_3$ or $\delta B = -K_3$, and similarly either $\delta C = K_4$ or $\delta C = -K_4$. Again, we can show that underestimation is more probable than overestimation.

APPENDIX C

Expected Value of the Least Squares Flow Solution

The expected value $E(\mathbf{u})$ of the least squares solution is given by

$$E(\mathbf{u}) = E((E^t E)^{-1}(E^t \mathbf{C})).$$

As the noise is considered independent and zero-mean, all the first order terms and the second order terms in the temporal noise vanish, and thus the expansion at point noise $N = 0$ (i.e., $\delta A_i = \delta B_i = \delta C_i = 0$) can be written as

$$E(\mathbf{u}) = \mathbf{u}' + \sum_i \left(\frac{\partial^2 \mathbf{u}}{\partial \delta A_i^2}\bigg|_{N=0} \frac{E\left(\delta A_i^2\right)}{2} + \frac{\partial^2 \mathbf{u}}{\partial \delta B_i^2}\bigg|_{N=0} \frac{E\left(\delta B_i^2\right)}{2} \right).$$

For notational simplicity, we define

$$\begin{aligned} M &= E^t E \quad \text{and} \quad & \mathbf{b} &= E^t \mathbf{C} \\ M' &= E'^t E' & \mathbf{b}' &= E'^t \mathbf{C}'. \end{aligned}$$

Using the fact that for an arbitrary matrix Q

$$\frac{-\partial Q^{-1}}{\partial x} = Q^{-1} \frac{\partial Q}{\partial x} Q^{-1}$$

We find the first order and second order derivatives to be

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial \delta A_i} &= -M^{-1} \begin{bmatrix} 2A_i & B_i \\ B_i & 0 \end{bmatrix} M^{-1}\mathbf{b} + M^{-1} \begin{bmatrix} C_i \\ 0 \end{bmatrix} \\ \frac{\partial^2 \mathbf{u}}{\partial \delta A_i^2} &= 2M^{-1} \begin{bmatrix} 2A_i & B_i \\ B_i & 0 \end{bmatrix} M^{-1} \begin{bmatrix} 2A_i & B_i \\ B_i & 0 \end{bmatrix} M^{-1}\mathbf{b} \\ &\quad - M^{-1} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} M^{-1}\mathbf{b} - 2M^{-1} \begin{bmatrix} 2A_i & B_i \\ B_i & 0 \end{bmatrix} M^{-1} \begin{bmatrix} C_i \\ 0 \end{bmatrix} \end{aligned}$$

and similarly, we have symmetric expressions for

$$\frac{\partial \mathbf{u}}{\partial \delta B_i} \quad \text{and} \quad \frac{\partial^2 \mathbf{u}}{\partial \delta B_i^2}.$$

Since we assume $E(\delta A_i^2) = E(\delta B_i^2)$, the expansion can thus be simplified to

$$E(\mathbf{u}) = \mathbf{u}' - n\underline{\underline{M'^{-1}\mathbf{u}'}}\sigma_s^2 + \sum_i \left\{ M'^{-1}\left(\begin{bmatrix} 2A_i' & B_i' \\ B_i' & 0 \end{bmatrix} M'^{-1} \begin{bmatrix} 2A_i' & B_i' \\ B_i' & 0 \end{bmatrix} + \begin{bmatrix} 0 & A_i' \\ A_i' & 2B_i' \end{bmatrix} M'^{-1} \begin{bmatrix} 0 & A_i' \\ A_i' & 2B_i' \end{bmatrix} \right) \mathbf{u} - M'^{-1}\left(\begin{bmatrix} 2A_i' & B_i' \\ B_i' & 0 \end{bmatrix} M'^{-1} \begin{bmatrix} C_i' \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & A_i' \\ A_i' & 2B_i' \end{bmatrix} M'^{-1} \begin{bmatrix} 0 \\ C_i' \end{bmatrix} \right) \right\} \sigma_s^2,$$

where we have underlined the term that diminishes proportionally to $\frac{1}{n}$ (where n is the number of measurements being combined in a region). The sum of these n terms will give a consistent, statistically constant response. The rest of the terms diminish proportionally to $1/n^2$. Informal experiments show that the sum of these terms becomes negligible for $n > 5$, a number clearly smaller than the number of terms likely to be combined in any real system.

ACKNOWLEDGMENT

The support of this research by the Office of Naval Research under Contract N00014-95-1-0521 is gratefully acknowledged, as is the help of Sara Larson in preparing this paper.

REFERENCES

1. E. H. Adelson and J. R. Bergen, Spatiotemporal energy models for the perception of motion, *J. Opt. Soc. Amer. A* **2**, 1985, 284–299.
2. E. H. Adelson and J. R. Bergen, The extraction of spatiotemporal energy in human and machine vision, in *Proc. IEEE Workshop on Visual Motion, 1986*, pp. 151–156.
3. P. Anandan, A computational framework and an algorithm for the measurement of visual motion, *Internat. J. Comput. Vision* **2**, 1989, 283–310.
4. J. Bigün, G. Granlund, and J. Wibelund, Multidimensional orientation estimation with applications to texture analysis and optical flow, *IEEE Trans. Pattern Anal. Mach. Intelligence* **13**, 1991, 775–790.
5. T. Brodský, C. Fermüller, and Y. Aloimonos, Simultaneous estimation of viewing geometry and structure, in *Proc. European Conference on Computer Vision, 1998*, pp. 342–358.
6. H. Bulthoff, J. Little, and T. Poggio, A parallel algorithm for real-time computation of optical flow, *Nature* **337**, 1989, 549–553.
7. P. J. Burt, C. Yen, and X. Xu, Multi-resolution flow through motion analysis, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1983*, pp. 246–252.
8. B. F. Buxton and H. Buxton, Computation of optic flow from the motion of edge features in image sequences, *Image Vision Comput.* **2**, 1984, 59–75.
9. W. Enkelmann, Investigations of multigrid algorithms for estimation of optical flow fields in image sequences, *Comput. Vision Graphics, Image Process.* **43**, 1988, 150–177.
10. C. Fermüller and Y. Aloimonos, Direct perception of three-dimensional motion from patterns of visual motion, *Science* **270**, 1995, 1973–1976.
11. C. Fermüller and Y. Aloimonos, Representations for active vision, in *Proc. International Joint Conference on Artificial Intelligence, 1995*.
12. C. Fermüller, L. Cheong, and Y. Aloimonos, Visual space distortion, *Biol. Cybernet.* **77**, 1997, 323–337.
13. C. Fermüller, R. Pless, and Y. Aloimonos, The Ouchi illusion as an artifact of biased flow estimation, *Vision Res.* **40**, 2000, 77–96.

14. D. J. Fleet and A. D. Jepson, Computation of component velocity from local phase information, *Internat. J. Comput. Vision* **5**, 1990, 77–104.
15. W. Fuller, "*Measurement Error Models*," Wiley, New York, 1987.
16. W. Fuller, Estimated true values for errors-in-variables models, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling* (S. van Huffel, Ed.), SIAM, Philadelphia, 1997.
17. M. Gennert and S. Negahdaripour, "Relaxing the Brightness Constancy Assumption in Computing Optical Flow," Al Memo 975, MIT, June 1987.
18. L. J. Gleser, Estimation in a multivariate "errors in variables" regression model: Large sample results, *Ann. Statist.* **9**, 1981, 24–44.
19. G. H. Golub and C. F. van Loan, An analysis of the total least squares problem, *SIAM. J. Numer. Anal*. **17**, 1980, 883–893.
20. D. Heeger, Optical flow using spatiotemporal filters, *Internat. J. Comput. Vision* **1**, 1988, 279–302.
21. E. Hildreth, Computations underlying the measurement of visual motion, *Artificial Intelligence* **23**, 1984, 309–354.
22. B. K. P. Horn and B. G. Schunk, Determining optical flow, *Artificial Intelligence* **17**, 1981, 185–203.
23. W. James and C. Stein, Estimation with quadratic loss, in *Proc. 4th Berkeley Symp. Math. Statist, Prob*. Vol. 1, pp. 361–379, University of California Press, Berkeley, CA, 1960.
24. E. H. Lehmann, *Theory of Point Estimation*. Wiley, New York, 1983.
25. B. Lucas and T. Kanade, An lterative image registration technique with an application to stereo vision, in *Proc. DARPA image Understanding Workshop 1981*, pp. 121–130.
26. H. Moravec, Towards automatic visual obstacle avoidance, in *Proc. International Joint Conference on Artificial Intelligence, 1977*, pp. 584–585.
27. H.-H. Nagel, Displacement vectors derived from second order intensity variations in image sequences, *Comput. Vision Graphics Image Process.* **21**, 1983, 85–117.
28. H.-H. Nagel, On the estimation of optical flow: Relations between different approaches and some new results, *Artificial Intelligence* **33**, 1987, 459–483.
29. H.-H. Nagel and W. Enkelmann, An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences, *IEEE Trans. Pattern Anal. Mach. Intelligence* **8**, 1986, 565–593.
30. H.-H. Nagel and M. Haag, Bias-corrected optical flow estimation for road vehicle tracking, in *Proc. International Conference on Computer Vision, Bombay, India, 1998*, pp. 1006–1011.
31. H. Ouchi. *Japanese and Geometrical Art*, Dover, New York, 1977.
32. P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
33. P. Schmidt, *Econometrics*, Dekker, New York, 1976.
34. E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, Probability distributions of optical flow, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Maui, Hawaii, 1991*, pp. 310–315.
35. A. Singh, *Optic Flow Computation: A Unified Perspective*, IEEE Computer Society Press, Los Alamitos, CA, 1992.
36. L. Spillmann, U. Tulunay-Keesey, and J. Olson, Apparent floating motion in normal and stabilized vision, *Investigative Ophthal. Visual Sci. Supple*. **34**, 1993, 1031.
37. G. W. Stewart, Stochastic pertubation theory. *SIAM Rev*. **32**, 1990, 576–610.
38. G. W. Stewart, Errors in variables for numerical analysis, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling* (S. Van Huffel, Ed.), SIAM, Philadelphia, 1997.
39. R. Szeliski, Bayesian modeling of uncertainty in low-level vision, *Internat. J. Comput. Vision*, **5**, 1990, 271–301.
40. W. B. Thompson and S. T. Barnard, Lower level estimation and interpretation of visual motion, *IEEE Comput*. August 1987, 20–28.
41. S. Uras, F. Girosi, and V. Torre, A computational approach to motion perception, *Biol. Cybernet*. **60**, 1988, 79–87.
42. S. van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.

43. V. G. Voinov and M. S. Nikulin, *Unbiased Estimators and Their Applications. Vol. 2, Multivariate Case*, Kluwer, Dordrecht, Norwell MA, 1993.
44. A. M. Waxman, J. Wu., and F. Bergholm, Convected activation profiles and receptive fields for real measurements of short range visual motion, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1988*, pp. 717–723.
45. J. Weber and J. Malik, Robust computation of optical flow in a multi-scale differential framework, *Internat. J. Comput. Vision* **14**, 1995, 67–81.
46. Y. Weiss and E. H. Adelson, "Slow and Smooth, a Bayesian Theory for the Combination of Local Motion Signals in Human Visions." AI Memo 1616, MIT, 1998.
47. R. Y. Wong and E. L. Hall, Sequential hierarchical scene matching, *IEEE Trans. Comput*. **27**, 1978, 359–366.
48. S. M. Zeki, *A Vision of the Brain*, Blackwell Scientific, London, 1993.