10.1 Differential Motion

Our visual world is inherently dynamic. People, cars, dogs, etc. are (usually) moving. These may be gross motions, walking across the room, or smaller motions, scratching behind your ear. Our task is to estimate such image motions from two or more images taken at different instances in time.

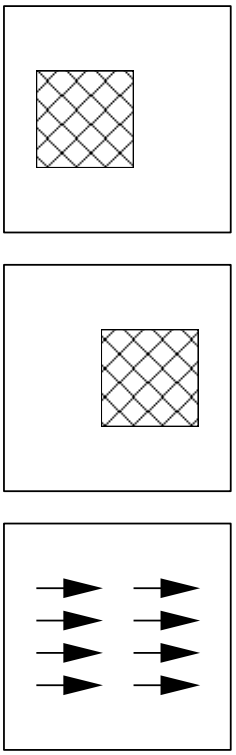

Figure 10.1 Flow field

With respect to notation, an image is denoted as $f(x, y)$ and an image sequence is denoted as $f(x(t), y(t), t)$, where $x(t)$ and $y(t)$ are the spatial parameters and t is the temporal parameter. For example, a sequence of N images taken in rapid succession may be represented as $f(x(t), y(t), t+i\Delta t)$ with $i \in [0, N-1]$, and Δt representing the amount of time between image capture (typically on the order of 1/30th of a second). Given such an image sequence, our task is to estimate the amount of motion at each point in the image. For a given instant in space and time, we require an estimate of motion (*velocity*) $\vec{v} = (\, v_x \quad v_y \,)$, where v_x and v_y denote the horizontal and vertical components of the velocity vector $\vec{v}$. Shown in Figure 10.1 are a pair of images taken at two moments in time as a textured square is translating uniformly across the image. Also shown is the corresponding estimate of motion often referred to as a *flow field*. The flow field consists of a velocity vector at each point in the image (shown of course are only a subset of these vectors).

In order to estimate motion, an assumption of *brightness constancy* is made. That is, it is assumed that as a small surface patch is moving, its brightness value remains unchanged. This constraint can be expressed with the following partial differential equation:

$$\frac{\partial f(x(t), y(t), t)}{\partial t} \;=\; 0. \tag{10.1}$$

This constraint holds for each point in space and time. Expanding this constraint according to the chain rule yields:

$$\frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial f}{\partial t} \;=\; 0, \tag{10.2}$$

where the partials of the spatial parameters x and y with respect to time correspond to the velocity components:

$$f_x v_x + f_y v_y + f_t \;=\; 0. \tag{10.3}$$

The subscripts on the function f denote partial derivatives. Note again that this constraint holds for each point in space and time but that for notational simplicity the spatial/temporal parameters are dropped. This transformed brightness constancy constraint is rewritten by packing together the partial derivatives and velocity components into row and column vectors.

$$\begin{aligned} \begin{pmatrix} f_x & f_y \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} + f_t &= 0 \\ \vec{f}_s^t \vec{v} + f_t &= 0. \end{aligned} \tag{10.4}$$

The space/time derivatives $\vec{f}_s$ and f_t are measured quantities, leaving us with a single constraint in two unknowns (the two components of the velocity vector, $\vec{v}$). The constraint can be solved by assuming that the motion is locally similar, and integrating this constraint over a local image neighborhood. A least-squares error function takes the form:

$$E(\vec{v}) = \left[\sum_{x,y} \vec{f}_s^t \vec{v} + \sum_{x,y} f_t \right]^2, \tag{10.5}$$

To solve for the motion this error function is first differentiated

$$\begin{aligned} \frac{\partial E(\vec{v})}{\partial \vec{v}} &= 2 \sum \vec{f}_s \left[\sum \vec{f}_s^t \vec{v} + \sum f_t \right] \\ &= 2 \sum \vec{f}_s \vec{f}_s^t \vec{v} + 2 \sum \vec{f}_s f_t. \end{aligned} \tag{10.6}$$

Setting equal to zero and recombining the terms into matrix form yields:

$$\begin{aligned} \begin{pmatrix} \sum f_x \\ \sum f_y \end{pmatrix} \begin{pmatrix} \sum f_x & \sum f_y \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} &= -\begin{pmatrix} \sum f_x f_t \\ \sum f_y f_t \end{pmatrix} \\ \begin{pmatrix} \sum f_x^2 & \sum f_x f_y \\ \sum f_x f_y & \sum f_y^2 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} &= -\begin{pmatrix} \sum f_x f_t \\ \sum f_y f_t \end{pmatrix} \\ M \vec{v} &= -\vec{b}. \end{aligned} \tag{10.7}$$

If the matrix M is invertible (full rank), then the velocity can be estimated by simply left multiplying by the inverse matrix:

$$\vec{v} = -M^{-1} \vec{b} \tag{10.8}$$

The critical question then is, when is the matrix M invertible? Generally speaking the matrix is rank deficient, and hence not invertible, when the intensity variation in a local image neighborhood varies only one-dimensionally (e.g., $f_x = 0$ or $f_y = 0$) or zero-dimensionally ($f_x = 0$ and $f_y = 0$). These singularities are sometimes referred to as the aperture and blank wall problem. The motion at such points simply can not be estimated.

Motion estimation then reduces to computing, for each point in space and time, the spatial/temporal derivatives f_x, f_y, and f_t. Of course the temporal derivative requires a minimum of two images, and is typically estimated from between two and seven images. The spatial/temporal derivatives are computed as follows. Given a temporal sequence of N images, the spatial derivatives are computed by first creating a temporally prefiltered image. The spatial derivative in the horizontal direction f_x is estimated by prefiltering this image in the vertical y direction and differentiating in x. Similarly, the spatial derivative in the vertical direction f_y is estimated by prefiltering in the horizontal x direction and differentiating in y. Finally, the temporal derivative is estimated by temporally differentiating the original N images, and prefiltering the result in both the x and y directions. The choice of filters depends on the image sequence length: an N tap pre/derivative filter pair is used for an image sequence of length N (See Section 7).

10.2 Differential Stereo

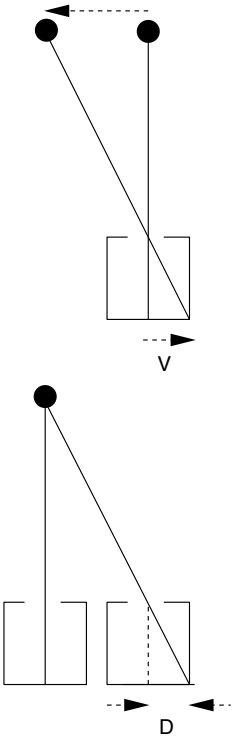

Figure 10.2 Motion and Stereo

Motion estimation involves determining, from a single stationary camera, how much an object moves over time (its velocity). Stereo estimation involves determining the displacement *disparity* of a stationary object as it is imaged onto a pair of spatially offset cameras. As illustrated in Figure 10.2, these problems are virtually identical: velocity ($\vec{v}$) $\equiv$ disparity (Δ). Motion and stereo estimation are often considered as separate problems. Motion is thought of in a continuous (differential) framework, while stereo, with its discrete pair of images, is thought of in terms of a discrete matching problem. This dichotomy is unnecessary: stereo estimation can be cast within a differential framework.

Stereo estimation typically involves a pair of cameras spatially offset in the horizontal direction such that their optical axis remain parallel (Figure 10.2). Denoting an image as $f(x, y)$, the image that is formed by translating the camera in a purely horizontal direction is given by $f(x + \Delta(x, y), y)$. If a point in the world (X, Y, Z) is imaged to the image position (x, y), then the shift $\Delta(x, y)$ is inversely proportional to the distance Z (i.e., nearby objects have large disparities, relative to distant objects). Given this, a *stereo pair* of images is denoted as:

$$f_L(x + \delta(x, y), y) \qquad \text{and} \qquad f_R(x - \delta(x, y), y), \tag{10.9}$$

where the disparity $\Delta = 2\delta$. Our task is to determine, for each point in the image, the disparity (δ) between the left and right images. That is, to find the shift that brings the stereo pair back

into register. To this end, we write a quadratic error function to be minimized:

$$E(\delta(x,y)) = [f_L(x+\delta(x,y),y) - f_R(x-\delta(x,y),y)]^2. \quad (10.10)$$

In this form, solving for δ is non-trivial. We may simplify things by expressing the image pair in terms of their truncated first-order Taylor series expansion:

$$f(x+\delta(x,y),y) \;=\; f(x,y) + \delta(x,y) f_x(x,y), \quad (10.11)$$

where $f_x(x,y)$ denotes the partial derivative of f with respect to x. With this first-order approximation, the error function to be minimized takes the form:

$$\begin{aligned} E(\delta) &= [(f_L + \delta(f_L)_x) - (f_R - \delta(f_R)_x)]^2 \\ &= [(f_L - f_R) + \delta(f_L + f_R)_x]^2, \end{aligned} \quad (10.12)$$

where for notational convenience, the spatial parameters have been dropped. Differentiating, setting the result equal to zero and solving for δ yields:

$$\begin{aligned} \frac{dE(\delta)}{d\delta} &= 2(f_L + f_R)_x[(f_L - f_R) + \delta(f_L + f_R)_x] \\ &= 0 \\ \delta &= -\frac{f_L - f_R}{(f_L + f_R)_x} \end{aligned} \quad (10.13)$$

Stereo estimation then reduces to computing, for each point in the image, spatial derivatives and the difference between the left and right stereo pair (a crude derivative with respect to viewpoint).

Why, if motion and stereo estimation are similar, do the mathematical formulations look so different? Upon closer inspection they are in fact quite similar. The above formulation amounts to a constrained version of motion estimation. In particular, because of the strictly horizontal shift of the camera pair, the disparity was constrained along the horizontal direction. If we reconsider the motion estimation formulation assuming motion only along the horizontal direction, then the similarity of the formulations becomes evident. Recall that in motion estimation the brightness constancy assumption led to the following constraint:

$$\frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial f}{\partial t} \;=\; 0, \quad (10.14)$$

Constraining the motion along the vertical y direction to be zero yields:

$$\frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial t} \;=\; 0, \quad (10.15)$$

where the partial derivative of the spatial parameter x with respect to time correspond to the motion (speed) in the horizontal direction:

$$f_x v_x + f_t = 0. \tag{10.16}$$

Unlike before, this leads to a single constraint with a single unknown which can be solved for directly:

$$v_x = -\frac{f_t}{f_x}. \tag{10.17}$$

This solution now looks very similar to the solution for differential stereo in Equation 10.13. In both solutions the numerator is a derivative, in one case with respect to time (motion) and in the other with respect to viewpoint (stereo). Also in both solutions, the denominator is a spatial derivative. In the stereo case, the denominator consists of the spatial derivative of the sum of the left and right image pair. This may seem odd, but recall that differentiation of a multi-dimensional function requires differentiating along the desired dimension and prefiltering along all other dimensions (in this case the viewpoint dimension).

In both the differential motion and stereo formulations there exists singularities when the denominator (spatial derivative) is zero. As with the earlier motion estimation this can be partially alleviated by integrating the disparities over a local image neighborhood. However, if the spatial derivative is zero over a large area, corresponding to a surface in the world with no texture, then disparities at these points simply can not be estimated.