

Computer Vision CITS4240

School of Computer Science & Software Engineering
The University of Western Australia

Motion

A lot of information can be extracted from time varying sequences of images, often more easily than from static images. For example, camouflaged objects are only easily seen when they move. Moreover, the relative sizes and position of objects are more easily determined when the objects move. Even simple image differencing provides an edge detector for the silhouettes of texture-free objects moving over any static background.

The analysis of visual motion divides into two stages:

- the measurement of the motion, and
- the use of motion data to segment the scene into distinct objects and to extract three dimensional information about the shape and motion of the objects.

There are two types of motion to consider: movement in the scene with a static camera, and movement of the camera, or ego motion. Since motion is relative anyway, these types of motion should be the same. However, this is not always the case, since if the scene moves relative to the illumination, shadow effects need to be dealt with. Also, specularities can cause relative motion within the scene. For this lecture, we will ignore all such complications.

The motion field

When an object moves in front of a camera, there is a corresponding change in the image. Thus, if a point p_o on a object moves with a velocity \mathbf{v}_o , then the imaged point p_i can be assigned a vector \mathbf{v}_i to indicate its movement on the image plane (see Figure 1) . The collection of all these vectors forms the *motion field*.

If we are only dealing with rigid body translations and rotations, then the motion field will be continuous except at the silhouette boundaries of objects (see Figure 2).

In the case of pure camera translation, the direction of motion is along the projection ray through that image point from which (or towards which) all motion vectors radiate. The point of divergence (or convergence) of all motion field vectors is called the *focus of*

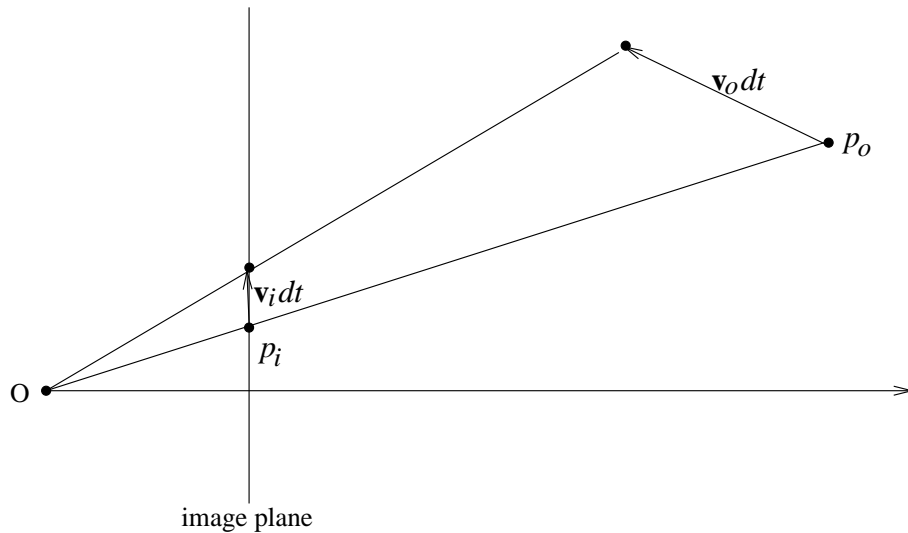


Figure 1: Object motion creates a motion field in the image.

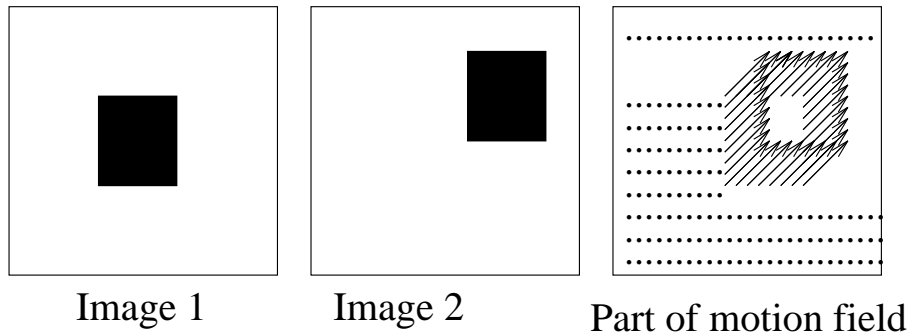


Figure 2: The motion field of a moving square.

expansion FOE (or *focus of contraction* FOC). Thus, in the case of divergence we have forward motion of the camera, and in the case of convergence, backwards motion.

If we take the axis of camera translation as the camera baseline in stereo, then every projection of a fixed scene point must translate along an epipolar line, and all such lines converge at the epipole, which is just the FOE.

Optical flow

Optical flow is the apparent motion of brightness patterns in the image. Generally, optical flow corresponds to the motion field, but not always. For example, the motion field and optical flow of a rotating barber's pole are different, as illustrated in figure 3. In general, such cases are unusual, and for this lecture we will assume that optical flow corresponds to the motion field.

One problem we do have to worry about, however, is that we are only able to measure the component of optical flow that is in the direction of the intensity gradient. We are

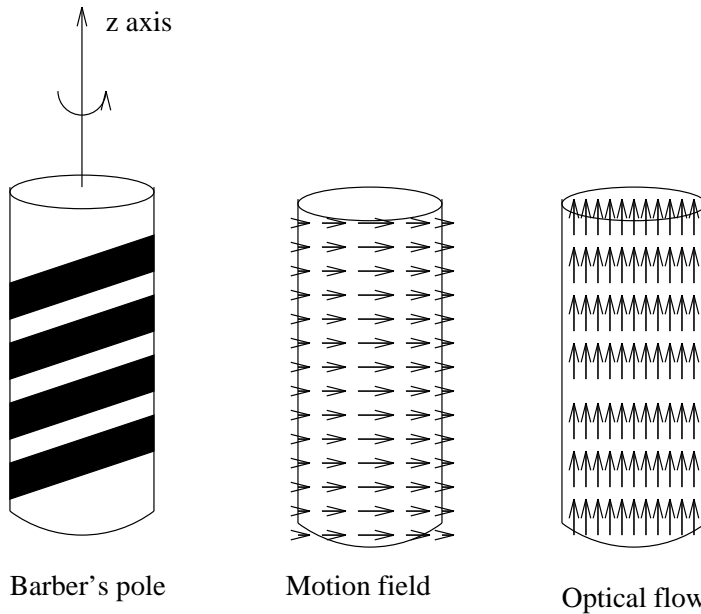


Figure 3: The motion field and optical flow of a barber's pole.

unable to measure the component tangential to the intensity gradient. This problem is illustrated in figure 4, and further developed below.

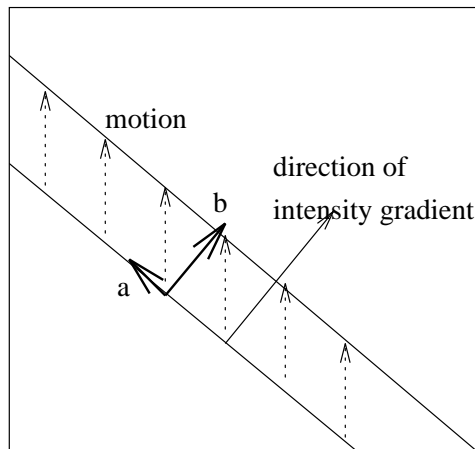


Figure 4: The aperture problem. We can only measure the component b.

Denote the intensity by $I(x, y, t)$. This is a function of three variables as we now have *spatiotemporal* variation in our signal. To see how I changes in time, we differentiate with respect to t :

$$\frac{dI}{dt} = \frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{\partial I}{\partial t}.$$

Now, we assume that the image intensity of each visible scene point is unchanging over time (for example, shadows and illuminations are *not* changing due to any motion), so we have

$$\frac{dI}{dt} = 0,$$

which implies

$$I_x u + I_y v + I_t = 0,$$

where the partial derivatives of I are denoted by subscripts, and u and v are the x and y components of the optical flow vector.

This last equation is called the *optical flow constraint equation* since it expresses a constraint on the components u and v of the optical flow.

The optical flow constraint equation can be rewritten as

$$(I_x, I_y) \cdot (u, v) = -I_t.$$

The equation above only provides information about the component of the optical flow in the direction of the image intensity gradient. That is,

$$\begin{pmatrix} u \\ v \end{pmatrix} = -\frac{I_t}{I_x^2 + I_y^2} \begin{pmatrix} I_x \\ I_y \end{pmatrix}. \quad (1)$$

We cannot, however, determine the component of the optical flow at right angles to this direction. This ambiguity is known as the *aperture problem*.

Generally the aperture problem is ignored/avoided by choosing to detect points in the image having significant intensity gradient in more than one direction ('corner points' or 'interest points' as used in stereo). These points do not suffer from the aperture problem and can be tracked to obtain the motion field.

The problems of motion analysis

The ultimate aim of analyzing motion in images is reconstruct the 3D motion and structure of the observed world. To do so, we must first tackle the *correspondence problem*:

Which pixels of an image correspond to which pixels of the next frame of the image sequence?

There are two main approaches:

- Convert the motion problem to a stereo problem and find the correspondence between a number of (sparse) feature points (e.g. corners) in the image at time t to the image at time $t + \delta t$.
- Compute the optical flow and use its geometrical properties to deduce three dimensional information about the scene and the motion. This approach requires the optical flow to be computed at regular pixel grid (e.g. every alternative or every 4th pixel) and some smoothness constraint of the optical flows over neighbouring pixels would be enforced.

The first approach, which leads to sparse 3D structure, is known as the *matching methods*. The second approach, which leads to dense 3D structure, is known as the *differential methods*.

Structure from motion using the matching methods

The determination of structure from motion using a number of image corners is effectively equivalent to stereo with a single camera. The problem is that first one has to deduce the three dimensional motion of the camera between the time intervals t and $t + \delta t$, assuming that the intrinsic parameters of the camera are known. Once this is achieved we can then solve for the three dimensional positions of the matched points using the standard stereo equations.

If both the view and scene geometry are unknown but the scene structure remains rigid between views, then it is possible to deduce the viewing geometry (up to a scale factor) and hence to solve for the scene structure. In such a scenario, all we have is a number of matched points in two images. The viewing geometry is specified by six parameters, namely the translation and rotation parameters of the camera motion.

Each observation of a point in the two images gives us four pieces of information, the row and column pixel coordinates in the two images. However, we also introduce three unknowns for each observation, namely the 3D coordinates of the observed point.

Thus, if n points are observed, we have $6 + 3n$ unknowns, and $4n$ observations. However, it is only possible to deduce camera translation up to a scale factor, as is illustrated in Figure 5. So we really only have $5 + 3n$ solvable unknowns. From this we can deduce that we must have at least $n = 5$ observations of matched points to solve the system. In practice many more points than 5 are used to reduce the influence of noise.

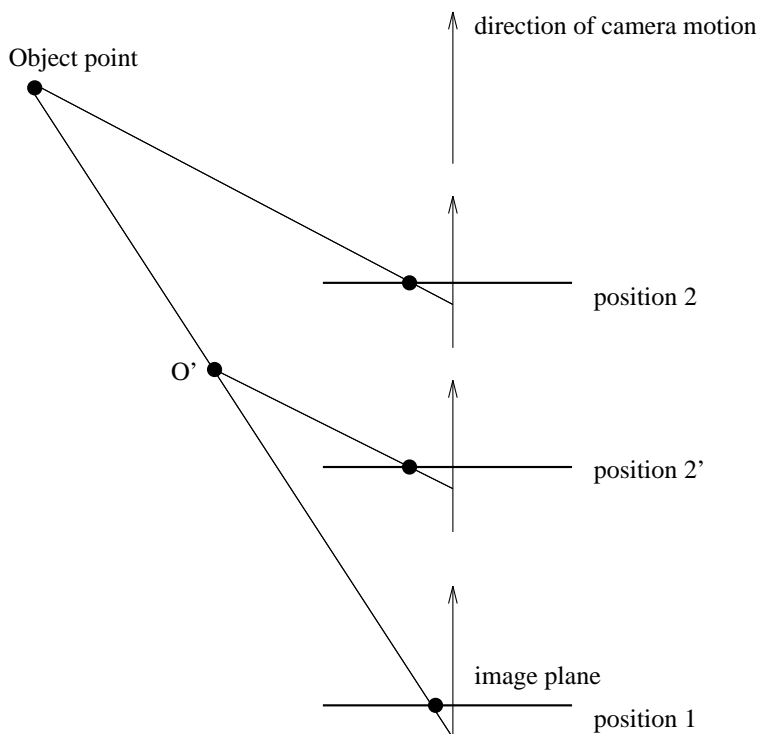


Figure 5: Camera translation can only be solved up to a scale factor. As the image point moves, this can correspond to a point far away and a large camera movement, or a point close by and a small camera movement.

Structure from motion using the differential methods

These methods are also known as structure from motion using optical flow. From our analysis that yields equation (1), it is clear that an additional constraint is required in order to solve the optical flow problem completely. Usually the motion field varies smoothly in most parts of the images. We shall try to therefore minimize a measure of *departure from smoothness*:

$$e_s = \iint ((u_x^2 + u_y^2) + (v_x^2 + v_y^2)) dx dy, \quad (2)$$

while the error in the optical flow constraint equation,

$$e_c = \iint (I_x u + I_y v + I_t)^2 dx dy, \quad (3)$$

should also be small.

Overall, we want to minimize $e_s + \lambda e_c$, where λ is a parameter that weights the error in the image motion equation relative to the departure from smoothness.

Minimizing an integral of the form

$$\iint F(u, v, u_x, u_y, v_x, v_y) dx dy$$

is a problem in the calculus of variation. The corresponding Euler equations are (see the detailed explanation given in Appendix A.6 of [1])

$$\begin{aligned} F_u - \frac{\partial}{\partial x} F_{u_x} - \frac{\partial}{\partial x} F_{u_y} &= 0 \\ F_v - \frac{\partial}{\partial x} F_{v_x} - \frac{\partial}{\partial x} F_{v_y} &= 0 \end{aligned} \quad (4)$$

In our case,

$$F(u, v, u_x, u_y, v_x, v_y) = ((u_x^2 + u_y^2) + (v_x^2 + v_y^2)) + \lambda (I_x u + I_y v + I_t)^2,$$

so the Euler equations give

$$\begin{aligned} \nabla^2 u &= \lambda (I_x u + I_y v + I_t) I_x \\ \nabla^2 v &= \lambda (I_x u + I_y v + I_t) I_y, \end{aligned} \quad (5)$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

is the Laplacian operator. The equations given in (5) can be solved using iterative methods.

For the discrete case, we can measure a departure from smoothness by

$$S_{ij} = \frac{1}{4} \{ (u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2 + (v_{i+1,j} - v_{i,j})^2 + (v_{i,j+1} - v_{i,j})^2 \},$$

while any error in the optical flow constraint equation is given by

$$C_{ij} = (I_x u_{i,j} + I_y v_{i,j} + I_t)^2.$$

The term to be minimized would be

$$E = \sum_i \sum_j S_{ij} + \lambda C_{ij},$$

where λ is a regularisation constant.

We differentiate E with respect to u_{kl} and v_{kl} to get

$$\frac{\partial E}{\partial u_{kl}} = 2(u_{kl} - \bar{u}_{kl}) + 2\lambda(I_x u_{kl} + I_y v_{kl} + I_t)I_x,$$

and

$$\frac{\partial E}{\partial v_{kl}} = 2(v_{kl} - \bar{v}_{kl}) + 2\lambda(I_x u_{kl} + I_y v_{kl} + I_t)I_y,$$

where \bar{u} and \bar{v} are local averages of u and v .

We equate these expressions to zero to get

$$(1 + \lambda I_x^2)u_{kl} + \lambda I_x I_y v_{kl} = \bar{u}_{kl} - \lambda I_x I_t$$

and

$$(1 + \lambda I_y^2)v_{kl} + \lambda I_x I_y u_{kl} = \bar{v}_{kl} - \lambda I_y I_t.$$

This gives us two equations in the two unknowns u_{kl} and v_{kl} . These can be solved directly and suggest the iterative scheme

$$u_{kl}^{n+1} = \bar{u}_{kl}^n - \frac{I_x \bar{u}_{kl}^n + I_y \bar{v}_{kl}^n + I_t}{1 + \lambda(I_x^2 + I_y^2)} I_x$$

and

$$v_{kl}^{n+1} = \bar{v}_{kl}^n - \frac{I_x \bar{u}_{kl}^n + I_y \bar{v}_{kl}^n + I_t}{1 + \lambda(I_x^2 + I_y^2)} I_y.$$

In other words, the new value of (u, v) at a point is equal to the average of the surrounding values minus an adjustment in the direction of the brightness gradient.

Note that this scheme also requires estimates of the values I_x , I_y and I_t . These can be computed by taking local averages in neighbourhoods about the grid point (i, j, k) .

Which one of these approaches is more suitable? The answer to this question would depend on the type of images you have. If many prominent image corners are present in your images and if these corners can be detected and matched then the first approach would be more suitable. As noise is often a problem with the optical flow approach, some prior smoothing to the images is often required. Depth discontinuities are also a problem for optical flow computation.

References

- [1] Berthold Klaus Paul Horn. *Robot Vision*. Chapter 12, MIT Press, 1986.
- [2] J. J. Koenderink and A. J. van Doorn, “Invariant Properties of the Motion Parallax Field due to the Movement of Rigid Bodies Relative to an Observer”, *Optica ACTA*, vol 22. No 9, pp 773-791, 1975
- [3] J. J. Koenderink and A. J. van Doorn, “How an Ambulant Observer can Construct a Model of the Environment From the Geometrical Structure of the Visual Inflow”, *Kybernetik* pp 224-247, 1978.
- [4] Milan Sonka, Vaclav Hlavac, Roger Boyle, *Image Processing, Analysis, and Machine Vision*. Chapter 15, Brooks/Cole, 1999.
- [5] Emanuele Trucco and Alessandro Verri, *Introductory Techniques for 3-D Computer Vision*, Chapter 8, Prentice Hall, 1998.