

Motion

Psychophysics, Computational Theory and Physiological Basis

Christof Koch

January 29, 2008

The ability to perceive motion and moving objects is a critical visual modality. While some animals do with little color and some without binocular stereo, all animals with eyes have at least a rudimentary ability to perceive motion. The uses of visual motion are varied:

- detecting and tracking objects by relative motion (object-tracking).
- monocular cue for depth from movement (structure-from-motion).
- optical flow for navigation using qualitative features (focus-of-expansion, time-to-contact).
- image segmentation via motion discontinuities.

The importance of motion is underscored by a patient with a (fortunately very rare) *motion deficit* caused by bilateral lesions in the parietal-occipital region, leading to “stroboscopic vision” (Hess, Baker and Zihl, 1989)¹.

The psychophysical study of motion goes back to the Gestalt movement during the first part of the last century in Germany (Wertheimer, 1923; Koffka, 1935). They distinguished between continuous or true motion and discrete or apparent motion (β movement). However, this distinction is an artificial one, because any system with receptive fields of non-zero finite support can be fooled by appropriate apparent motion stimuli.

¹The patient had a bilateral loss in the superior temporal region (in the general area corresponding to area MT in the monkey and had severely impaired motion perception (e.g. she was unable to perceive motion above $6^\circ/sec$). That is, she can see a car but not motion of the car. Her residual motion vision corresponds to the “short-range” process. Striking was the patient’s normal performance on other perceptual tasks not involving motion (vernier acuity, temporal resolution, stereopsis, color discrimination).

Note a few peculiarities of motion perception: we can't see things that move too slowly (e.g. the sun) or too fast (e.g. a propeller) and we also see a sequence of stationary scenes as moving, *i.e.*, the basis of movies.

Optical flow is not such a popular topic in machine vision anymore. Thus, *Computer Vision: A Modern Approach* by Forsyth & Ponce is not that useful. Much better is Horn's textbook (Horn, 1986).

1 Short-Range versus Long-Range Motion

From the standpoint of both psychophysics (Braddick, 1974) and machine vision, two types of motion processes can be distinguished. *Short-range* or *intensity-based* motion uses the image intensity or some linear/nonlinear function of intensity to estimate motion at **every** location. Intensity-based schemes give rise to the *aperture* problem and are noise prone. *Long-range*, *token-* or *feature-based* motion schemes match high-level tokens in consecutive image frames (e.g. corners or small, rapidly moving blobs) at *sparse* locations in the image. They give rise to the *correspondence problem*. This is related to, but more difficult, than the correspondence problem in stereo since there is no epipolar line constraint. Matching only occurs at those locations where tokens have been identified. Long-range motion algorithms find application in high-level and well-specified applications where model-based approaches can be used (e.g. missile defense, tracking cars).

The short-range system can be seen in action in the two movies shown in class, adapted from the original Braddick (1974) experiments. In both, the background is decorrelated from one frame to the next and consists of random 5x5 blocks of all black/white pixels (50%). In both movies, a small, constant random square moves across the changing background. In one case, it moves one pixel and in the other it jumps by nine each frame. The latter does not result in any sense of motion.

The short- and long-range motion process duality is one means to conceptualize at least two ways in which our brains perceive motion.

Short-range motion	Long-range motion
Criterion of Segregation in Random-Dot Display	Criterion of Smooth Apparent Motion for Isolated Element
Spatial displacement must be 15' or less (Braddick, 1974).	Spatial displacement may be many degrees.
Interstimulus interval (i.s.i.) must be less than 80-100 msec (with 100 msec stimulus exposure)	i.s.i. may be up to at least 300 msec.
Segregation abolished by bright uniform field in i.s.i. (Braddick 1980).	Motion perceived whether i.s.i. is bright or dark.
Successive stimuli must be delivered to the same eye or to both eyes together (Braddick 1974), as must bright field for effective masking (Braddick 1980).	Successive stimuli may be delivered to the same or different eyes.
Pattern defined by chromatic but not luminance contrast is inadequate (Ramachandran and Gregory, 1978).	Stimuli may be defined by chromatic contrast alone (Ramachandran and Gregory, 1978).
Can be adapted by appropriate long-duration stimuli (motion aftereffect).	Split-brain patients can perform long-range motion across the midline (Ramachandran <i>et al.</i> , 1986).

Table 1. Determinants of Apparent Motion found with Two Perceptual Criteria. Modified from Braddick (1980).

2 First-, Second- and Third-Order Motion Processing

A different way (Chubb & Sperling, 1988; Lu & Sperling, 2001) to classify motion perception in humans and other animals is based on a first, second-

and third-order motion system that can extract motion from different classes of stimuli.

The *first order* system uses a primitive motion energy computation (akin to the Reichardt-Hassenstein/correlation model of insect motion or the spatio-temporal energy model discussed next) to extract *motion energy* (**not** velocity) from a moving luminance profile. This is also called Fourier motion. In the quicktime movie shown in class, we have

$$I(x, t) = \sin(\omega \cdot x + f \cdot t). \quad (1)$$

It assumes that the photoreceptor intensity, or some linearly filtered version of the intensity, is moving.

Second-order motion uses motion energy to extract motion from moving texture-contrast modulations. In the example shown in class, the contrast of a stationary “carrier” (a random background) is modulated by a moving sinusoidal function:

$$I(x, t) = \text{noise}(x) \cdot \sin(\omega \cdot x + f \cdot t) \quad (2)$$

where $\text{noise}(x)$ is a uniformly distributed random variable between -0.5 and +0.5.

Here, the expected luminance over some neighbourhood is zero (on average) and a simple linear filter is blind to such motion. This type of motion is called *second-order* or *Fourier motion*. We show a second example, in which the random b/w signal at one pixel (or one vertical line of pixels) is flipped and this disturbance travels across the image.

Such a moving stimulus is not picked up by a first-order motion system. It can be picked up if the input is first passed through a temporal or spatial filter followed by a quadratic nonlinearity (a full-wave rectification would also do). In the case of the travelling disturbance, the absolute value of the temporal derivative could be used as the input to a first-order motion system.

Finally, a third type of motion analysis, called *third-order* or *feature tracking system* computes correspondences among features by using top-down attentional-driven processes as well as bottom-up components. This third-order motion process might correspond to the long-range motion system of Braddick. Lu and Sperling (2001) have constructed moving stimuli

that can't be detected by first- and second-order motion system but that human observers clearly perceive as moving.

Sperling's framework is more exact and rigorous than the short- and long-range dichotomy introduced by Braddick. Short-range motion stimuli could be picked up by either first- or second-order motion detectors while long-range motion would be based on selective attention.

3 Spatio-Temporal Energy Model

Van Santen and Sperling (1984), Watson and Ahumada (1985) and Adelson and Bergen (1985) all proposed motion models for human perception based on careful psychophysical measurements. The output can formally be expressed as a second-order interaction using the simplest type of nonlinearity possible, multiplication. Using psychophysics—which only has access to the final output of a system—none of these three models can be distinguished from each other. However, some of the internal stages differ; in particular, the spatio-temporal energy model has a stage responding only to motion in one direction, while the correlation model has no such stage.

The most popular of these models, the *spatio-temporal motion energy* model of Adelson and Bergen (1985), carries the idea of finding the orientation of a stimulus in the x-y plane over into the spatio-temporal domain, i.e. into the x-t (or the x-y-t space). They derive a series of spatio-temporal filters that detect a certain orientation in the x-t plane. The key observation is the following.

The motion of a particular point on a moving pattern $I(x, t) = I(x - vt)$, can be represented by a straight line of slope $1/v$ in a space-time diagram. The Fourier transform (in space and time) of I is then given by $\tilde{I}(\omega_x)\delta(\omega_x v + \omega_t)$ ² In other words, the entire support of this function lies on a line of slope v . The Fourier transform of I moving in the opposite direction is only different from zero along a line going through the origin and with slope $-v$.

One can now design filters that only have non-zero support along these diagonals to estimate both the direction as well as the magnitude of speed. This idea can be easily generalized to 2-D motion.

²Remember that the Fourier transform of a shifted function, $I(x + x_0)$, is $e^{-ix_0\omega}\tilde{I}$ and the transform of $e^{-ix_0\omega}$ is $\delta(x + x_0)$.

Because of mathematical convenience and because the receptive fields of simple cells can be rather well approximated by Gabor functions, spatio-temporal energy models use the following pair of spatio-temporal Gabor filters with even and odd phases tuned to leftward (-) and rightward (+) motion (other oriented filters could also be used)

$$S_{even}^{\pm}(x, t) = \frac{1}{2\pi\sigma_x\sigma_t} \times \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{t^2}{2\sigma_t^2}\right) \times \cos(\omega_x x \mp \omega_t t), \quad (3)$$

and

$$S_{odd}^{\pm}(x, t) = \frac{1}{2\pi\sigma_x\sigma_t} \times \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{t^2}{2\sigma_t^2}\right) \times \sin(\omega_x x \mp \omega_t t). \quad (4)$$

This generalizes in a trivial manner to 2-D space. These filters are parametrized by a particular spatial and temporal scales, σ_x and σ_t

These filters are not **separable in space-time**³.

Because the output of the motion signal should not vary with time for a constantly moving stimulus and should not depend on the exact position of the stimulus with respect to the filter, Adelson and Bergen use a standard engineering trick, that of squaring and adding two filters that are 90° out of phase (based on the equation: $\sin^2 \alpha + \cos^2 \alpha = 1$), to render the output independent of the phase. A pair of such filters are said to be in *quadrature*. This quadratic nonlinearity also removes the dependency of the output on the sign of the contrast of the stimulus (responding equally well to a black-to-white step as to a white-to-black step edge). This amounts to computing

$$E^+(x, t) = (S_{even}^+ * I(x, t))^2 + (S_{odd}^+ * I(x, t))^2, \quad (5)$$

for the unit responding to rightward motion and

$$E^-(x, t) = (S_{even}^- * I(x, t))^2 + (S_{odd}^- * I(x, t))^2, \quad (6)$$

for the unit for leftward motion. Such an uni-directional stage is a prediction of the spatio-temporal energy model.

These units respond uniquely to either left- or right-ward motion. In the last stage of the model the two uni-directional outputs are subtracted from each other,

$$E(x, t) = E^+(x, t) - E^-(x, t). \quad (7)$$

³A function $f(x, t)$ is separable if it can be written as $f(x, t) = g(x) \times h(t)$. A function expressed as a sum of separable functions, i.e. $\sum g_i(x) \times h_i(t)$ is itself not separable.

A number of psychophysical observations argue for such an opponency stage. (1) It is not generally possible to see both leftward and rightward motion at the same place and time within the same frequency band. (2) Adaptation phenomena such as the motion aftereffect suggest that motion perception involves the balance between opposing leftward and rightward-motion signals. (3) It has been found that leftward and rightward-moving gratings can effectively cancel each other's detectability.

The global output of these models, meant to mimic the motion perception of humans, is very similar to the output of the Reichardt correlation model formulated for the fly. That's pretty neat!

3.1 Psychophysical Support

van Santen and Sperling (1984) present psychophysical evidence arguing that the motion system responds proportional to the product of neighbouring contrasts.

Lu and Sperling (2001) exploit two specific properties that hold for both the correlation-model in the fly and the spatio-temporal energy model of human: *pseudo-linearity*, that is, when a stimulus is composed of sinusoidals of different frequencies, the detector response to the sum is the sum of its individual responses,⁴ and *static displays are ignored*, that is, the output to a stationary pattern is zero. In general, adding a stationary background pattern—the *pedestal*—to a moving pattern does not change the output of the motion (pedestal immunity of motion energy detectors). This prediction has been tested in people (Lu and Sperling, 2001) and is found to be true. That it, although these pattern may look quite distinct, the accuracy of left-right discrimination for a moving stimulus and the same moving stimulus plus a stationary offset are the same.

A true speed detector, such as the gradient model in its abstract form, will either signal directly speed (analog coding) or will fire maximally at its “optimal” speed (place coding), independent of the structure of the object. In contrast, a energy motion model does not correctly signal the local motion in terms of its velocity. Instead, its output depends on both the velocity v as well as on its spatial frequency λ .

⁴The system is pseudo-linear since this property only holds for sinusoidal of different temporal frequencies.

Diener, Wirt, Dichgans and Brandt (1976) provide psychophysical support that human observers—asked to judge the speed of a grating they fixate (when they do not track)—perceive a quantity proportional to velocity over spatial wavelength, given by

$$M = 0.61 \frac{v}{\lambda} + v \cdot b \quad (8)$$

The $1/\lambda$ effect on perceived speed disappears during pursuit movements of the eye (in other words, the system that tracks smooth motion implements a different motion algorithm than the one used to estimate motion in the scene).

Much of the psychophysics carried out in support of the energy model always uses very small visual contrast values, i.e. below 5%. Higher contrast values will saturate various components of the visual pathway (e.g. we know from functional MRI in humans, that the BOLD response in area MT/MST, one of the cortical areas specialized for motion analysis, saturates above 2-3%) and thereby bring saturating types of nonlinearities into play. These higher-order nonlinearities contaminate the essential quadratic nonlinearity needed for this computation. Such an effect has been observed for the fly, where the opto-motor response is proportional to the square of the contrast for small values, but flattens out for higher values of contrast, as well as for the human motion system.

Again, these spatio-energy models do **not** explicitly compute motion, but rather some function which varies with the direction and the magnitude of the velocity.

4 The Motion Pathway

As discussed early on in class, the *magnocellular* pathway, originating with the parasol cells in the retina and projecting into the two magnocellular layers of the LGN, is sensitive to low spatial contrasts and to high temporal frequencies.

The geniculate magnocellular cells terminate in layer $4C\alpha$ of V1. The output of this layer passes to layer 4B and from there directly to the *middle temporal* area MT. There is also a second, indirect route to MT via parts of area V2 termed *thick stripes*.

About 25% of cells in V1 are direction selective, that is respond strongly to motion in one, the *preferred*, direction and very little or not at all to motion in the opposite or *null* direction. V1 is the first stage where such direction selective cells—colocalized primarily to layers 4C α and 4B—are found in any significant numbers.

Next to V1, no cortical area is as popular among electrophysiologists as the fifth visual area (V characterized by John Allman in the New World monkey and by Semir Zeki in the Old World monkey (Allman and Kass, 1971; Zeki, 1974).⁵ For all of its popularity, MT is a comparatively small part of cortical real estate, about 50 mm² in area in the macaque, that is, about 1/20-th the extent of V1. Area MT has been identified in all primates, including humans, on the basis of four criteria: it receives a direct input from striate cortex, it contains a large proportion of cells that respond strongly to stimuli moving in one direction but show little or even a suppressed response to motion in the opposite direction, it contains a complete representation of the contralateral visual hemifield and is it is heavily myelinated.

The V1 to MT connection originates with neurons in layer 4B and, to a lesser extent, in the upper part of layer 6. This forward projection is therefore dominated by magnocellular projections. Like most connections in cortex, this one is reciprocal, with the feedback pathway from MT into V1 showing less specificity. MT receives additional input from the thick, cytochrome oxidase rich stripes of V2, containing many direction selective cells.

While one could conclude from this that all of MT input derives, directly or indirectly, from V1, this is not the case. Inactivating V1 by cooling—an operation that is reversible—or by surgically ablation, causes MT cells to respond more sluggish and at a lower rate, yet without losing their receptive field properties (such as direction selectivity). All MT responses can be eliminated if both V1 and the superior colliculus are removed. This supports the notion that MT, as do other areas, receives a secondary visual input that bypasses the LGN by going from the retina to the superior colliculus and from there to a part of the thalamus known as the pulvinar that projects directly into MT and other visual cortical areas.

⁵For historical reasons, MT was the label used for the area in the new world monkey, while V5 was reserved for the old world monkey. However, today both terms are used interchangeable in much of the literature, with the human homologue sometimes referred to as V5/MT.

More than 80% of MT neurons prefer stimuli moving in a particular direction, with the average cell firing more than ten times stronger to motion in its preferred direction than to motion in the opposite direction. And cells retain this selectivity over a considerable range of speeds, stimulus size and position.

The moving stimulus can be a bar, a grating, or a cloud of dots. Neurons care little whether the animal moves its eye over a stationary stimulus or whether the stimulus moves. Neurons in MT respond rapidly to an appropriate stimulus, typically within 50-60 msec, compatible with the heavy myelination of this area that ensures that action potentials travel at high speed along the axons.

MT provides a major output to the dorsal stream, in particular to areas MST (medial superior temporal) and VIP (ventral intraparietal).

Inactivating MT cells temporarily by injecting a chemical substance into this part of the brain increases the threshold for detecting weak motion signals in behaving monkeys without increasing the acuity, color or stereo thresholds (Newsome and Pare, 1988). This provides *prima facie* evidence for the critical role of MT in primate motion perception.

Historically, the cortical area in humans taken to be the homologue of monkey MT has been identified using stains that emphasize axonal myelination in cortical structures. More recently, an antibody that selectively binds to a protein associated with neurons in the magnocellular pathway have been used. Such anatomical techniques have a major drawback in that they can only be applied to *post mortem* material.

The advent of brain imaging has changed this situation dramatically. Positron emission tomography was first used to locate MT by Zeki, Frackowiak and their colleagues in London on the basis of its differential activation by moving clouds of dots compared to stationary ones. PET is quite sensitive to changes in blood flow dynamics in response to neuronal activity but requires injection of a radioactive bolus, hindering its repetitive use in humans. Functional MRI has no such limitation and today MT is quickly and routinely localized on the basis of the different hemodynamic responses between low contrast concentric rings that are slowly expanding and a stationary version of this stimulus or using clouds of moving stimuli. I'll show a little movie from Melissa Saenz in my laboratory that Demonstrates this procedure.

5 The Velocity Field

Let us now return to the computational problem of motion.

Here we must distinguish between the *optical flow field* and the underlying *velocity field*. The latter is a purely geometrical concept, while the former one depends on our assumption about the world, light sources etc. Let us study this difference in the case of an observer (with her optical axis in the direction of the z-axis) moving through a rigid (static) environment. This problem is known as *passive navigation*. A world point $\mathbf{R} = (X, Y, Z)^t$ maps onto the point $\mathbf{r} = (x, y, 1)^t$ in the image plane (the image plane is at a fixed distance $|\hat{\mathbf{z}}|$ from the origin, where $\hat{\mathbf{z}}$ is the unit vector in the Z direction). Here $Z > 0$ for all points in front of the imaging system and $\hat{\mathbf{z}} \cdot \mathbf{r} = 1$. We consider $|\hat{\mathbf{z}}|$ to be the focal length with $|\hat{\mathbf{z}}| = 1$.

The points \mathbf{r} and \mathbf{R} are related via the *perspective projection* equation. For this geometry we have:

$$\frac{\mathbf{r}}{|\hat{\mathbf{z}}|} = \frac{\mathbf{R}}{Z} \Rightarrow \mathbf{r} = \frac{\mathbf{R}}{Z} = \frac{\mathbf{R}}{\mathbf{R} \cdot \hat{\mathbf{z}}}. \quad (9)$$

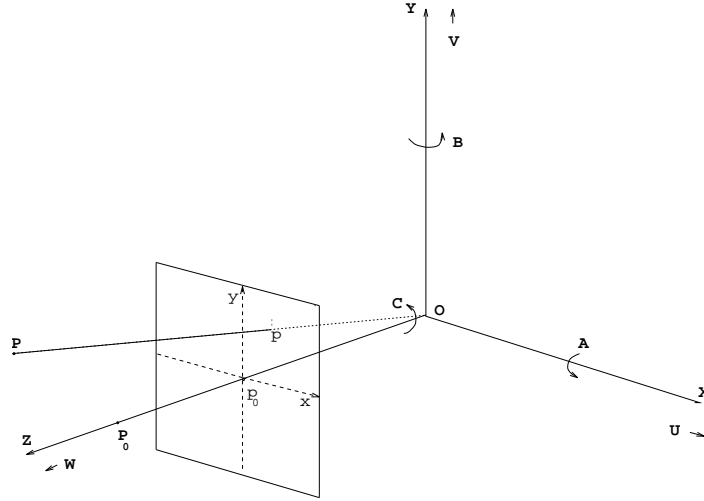
In other words, $x = X/Z$ and $y = Y/Z$. Thus, under this perspective (or ideal pinhole camera) projection, the coordinates in the image plane are scaled by the Z coordinate. Notice that if the variations in depth among points \mathbf{R} in the world are small relative to the total distance involved, the perspective projection turns into a *orthographic projection*, with $x \approx \alpha X$ and $y \approx \alpha Y$.

Let us now assume that the observer moves with instantaneous translational velocity $\mathbf{t} = (U, V, W)^t$ and instantaneous rotational velocity $\omega = (A, B, C)^t$ around the origin, corresponding to (pitch, yaw, tilt). Then the 3-D velocity of the point R with respect to the XYZ coordinate system is

$$\boxed{\dot{\mathbf{R}} = \frac{d\mathbf{R}}{dt} = -\mathbf{t} - \omega \times \mathbf{R},} \quad (10)$$

where \times denotes the vector-product and $\dot{\mathbf{R}}$ denotes the derivative of the vector \mathbf{R} with respect to the variable t . Or, in components,

$$\frac{d\mathbf{R}}{dt} = \begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} = \begin{pmatrix} -U - BZ + CY \\ -V - CX + AZ \\ -W - AY + BX \end{pmatrix}. \quad (11)$$



The motion of the world point \mathbf{R} results in motion of the corresponding image point in the opposite direction. The value of this 2-D motion or velocity field is

$$\frac{d\mathbf{x}}{dt} = \frac{d}{dt} \frac{\mathbf{R}}{\mathbf{R} \cdot \hat{\mathbf{z}}} = \frac{\dot{\mathbf{R}}(\mathbf{R} \cdot \hat{\mathbf{z}}) - (\dot{\mathbf{R}} \cdot \hat{\mathbf{z}})\mathbf{R}}{(\mathbf{R} \cdot \hat{\mathbf{z}})^2}, \quad (12)$$

(remember that $\hat{\mathbf{z}}$ is a constant vector: $d\hat{\mathbf{z}}/dt = 0$). In compound form we can write for the 2-D velocity field induced in the image plane

$$\dot{\mathbf{r}} = \frac{d\mathbf{x}}{dt} = \begin{pmatrix} \dot{x} \\ \dot{y} \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{-U+xW}{Z} + Axy - B(x^2 + 1) + Cy \\ \frac{-V+yW}{Z} - Bxy + A(y^2 + 1) - Cx \\ 0 \end{pmatrix}, \quad (13)$$

a result first obtained by Longuet-Higgins and Prazdny (1980). We can write these equations in the form of

$$\dot{\mathbf{r}} = \dot{\mathbf{r}}_{trans} + \dot{\mathbf{r}}_{rot}, \quad (14)$$

where *trans* and *rot* are the component of the velocity field due to translation and rotation. Note that the 2-D velocity field in the image plane $\dot{\mathbf{r}}$ is a purely geometrical concept; we have not talked at all about image intensities.

One interesting question is whether it is possible, given this 2-D velocity field, to recover the 6 parameters of motion, that make up \mathbf{t} and ω ? The

general answer is no. There remains always a scaling factor. A particular optical flow can either be generated by the surface $S_1(x, y, z)$ moving with the motion vector $\mathbf{m}_1 = \mathbf{t} + \omega$ or by the “dilated” surface $S_2(kx, ky, kz)$ moving with $\mathbf{m}_2 = k\mathbf{t} + \omega$. In other words, only the direction of the translation, but not its amplitude can be specified.

5.1 Pure Translation

Let us consider some special cases. How does the velocity field look in the case of pure translation, i.e. $\omega = (0, 0, 0)$? We then have two linear equations at each point:

$$\dot{x} = \frac{-U + xW}{Z} \quad (15)$$

$$\dot{y} = \frac{-V + yW}{Z}. \quad (16)$$

At the location $(x = U/W, y = V/W)$ the velocity field will be zero. This point is called *focus of expansion* (FOE). It points in the direction of motion (notice, once again, the scaling factor W). From psychophysical experiments (Warren and Hannon, 1988), it is known that human observers can determine the FOE—and thus their heading—within 1° , even in the face of eye movements (and, therefore, rotatory movements in the optical flow). Under physiological conditions (that is for reasonable eye movements of $1^\circ/\text{sec}$ or higher) humans require extra-retinal information to perceive heading accurately (Royden, Banks and Crowell, 1992).

In the case of translation, the velocity field varies inversely with depth Z . Therefore distant objects (e.g., sun) will not appear to move on the retina. If we know U and W we can then recover the depth.

Let’s look at an even more specialized case, where we assume that the observer is translating along her line of sight (i.e., $U = 0, V = 0$). We then have

$$\dot{\mathbf{r}} = \begin{pmatrix} xW/Z \\ yW/Z \\ 0 \end{pmatrix}, \quad (17)$$

or, $\dot{x}/x = \dot{y}/y = W/Z$. The variable Z/W is the time τ it takes for an object moving at constant velocity W to cross the distance Z ; $W/Z = 1/\tau$. The

interval τ —known as the *time-to-contact*, *collision* or *crash*—is an important behavioral variable used by flies, birds and humans (Gray and Regan, 1998)⁶:

$$\frac{\dot{r}}{r} = \frac{1}{\tau}. \quad (18)$$

For motion towards a fronto-parallel plane, this equation can be used to estimate the time to impact without making any assumptions about the size or nature of the approaching object and without requiring stereo. One only needs to measure the relative rate of optical looming.

A very robust algorithm to compute time-to-contact exploits the *Divergence theorem*, a fact pointed out by Poggio, Verri and Torre (1991). For any area A and its closed contour C , we can always write the following expression

$$\int_A \nabla \cdot \mathbf{V}(x, y) dx dy = \int_C \mathbf{V} \cdot \mathbf{n} ds. \quad (19)$$

Or, in words, the divergence of the velocity field over an area A is identical to the velocity component perpendicular to the contour (\mathbf{n} is the normal vector to the contour).

Let us apply this to the case where the motion detectors are arranged radially in a circle of radius R . In the general case of an arbitrary translation (U, V, W) of a planar surface parallel to the image plane at a distance Z_0 (notice that this includes the case when the observer is not directly heading for the stationary system) the resulting velocity field is linear and $\nabla \cdot \dot{\mathbf{r}}$ is always constant, independent of the exact location of the area:

$$\int_C \dot{\mathbf{r}} \cdot \mathbf{n} ds = 2\pi R^2 \frac{W}{Z_0} = \frac{2\pi R^2}{\tau}. \quad (20)$$

What this means is that a simple integration of the output of a number of 1-D motion detectors arranged along a circular contour directly yields the time-to-contact, **without** computing any spatial or temporal derivatives. Furthermore, because we are integrating over the output of a number of elementary motion detectors, each individual measurement can be noisy. This seems to be a very robust operation.

⁶The accuracy with which top athletes can judge the time of arrival of an approaching ball is a remarkable ± 2 msec.

5.2 Pure Rotation

Let us look at the velocity field in another special case, namely rotation around the vertical axis only: $\mathbf{t} = (0, 0, 0)$ and $\boldsymbol{\omega} = (0, B, 0)$ (yaw).

$$\dot{\mathbf{r}} = \begin{pmatrix} -B(x^2 + 1) \\ -Bxy \\ 0 \end{pmatrix}. \quad (21)$$

The velocity field does not have a focus-of-expansion, nor does it contain any information about depth, since Z never appears anywhere. Thus, the velocity of points only depend on their location (x, y) on the imaging plane, but not at all on their distance.

5.3 Passive Navigation

For passive navigation (where the world is assumed to be stationary or move much less than the moving observer), the system needs to recover five degrees of freedom (tilt, pan and yaw plus two degrees of freedom for translation) while the number of data point available scales with the size of the image. Even if these measurements are noisy, in general it is possible to recover the parameters associated with self motion from images, even without non-visual (e.g. vestibular, GPS) information.

6 The Optical Flow

So far, we have not mentioned the changing image intensities. This is the only data available at the level of the retina, of course. Let us look at the optical flow, induced by the time-varying image intensities falling onto the retina or camera. The standard approach (Fenneman and Thomspon, 1979; Horn and Schunck, 1981; Nagel, 1982) assumes that a local patch of the image does not change as the observer moves (or as the image moves relative to the observer). That is,

$$\boxed{\frac{dI(x,y,t)}{dt} = 0.} \quad (22)$$

This is not to say that the image remains the same throughout, just that locally the total derivative doesn't change. This *total temporal derivative*

can be thought of as the temporal derivative along the trajectory of the point (x, y) onto the image plane. This can be re-written as the *brightness change constraint equation* using the change rule

$$\frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{\partial I}{\partial t} = I_x u + I_y v + I_t = 0. \quad (23)$$

The 2-D *optical flow field* (or optical flow for short) is identified with $(dx/dt, dy/dt) = (u, v)$.

The image brightness constancy equation is a sort of conservation law, strictly only true for rigid translation of a Lambertian surface in the image plane (Verri and Poggio, 1987). This is easy to see in the 1-D case: assume $I(x, t) = I(x - ut)$. Then

$$u = -\frac{I_t}{I_x}. \quad (24)$$

In two dimensions, this implies that only the component of the velocity perpendicular to the gradient ∇I can be determined. It is equal to

$$-\frac{I_t}{|\nabla I|}. \quad (25)$$

Motion algorithms such as this one that explicitly compute velocity based on the ratio of **temporal** and **spatial derivatives** (first or higher-order derivatives) are known as *gradient* algorithms.

$I_x u + I_y v + I_t = 0$ is a single linear equation in two unknowns and therefore does not have a unique solution. If the optical flow $u = u(x, y)$ and $v = v(x, y)$ is measured at n locations in the visual field, we are faced with n linear equations in $2n$ unknowns. This *ill-posed* problem (infinite many solutions) is also known as the *strong aperture problem*: only the component of motion normal to the local contrast can be recovered:

$$(u, v) = -I_t \cdot \nabla \mathbf{I} / (\nabla I)^2. \quad (26)$$

In practice, the brightness of a patch rarely remains exactly the same. The object (or the camera) might rotate, revealing new parts of the object, a specularity might occur etc. However, as long as the change in brightness at an image point due to the motion within the interval dt is much larger than the change in brightness due to other effects, such as a change in viewing conditions or illumination, the image constraint equation is not an unreasonable one.

It is important to emphasize the difference between the optical flow field (u, v) and the velocity or motion field (\dot{x}, \dot{y}) . Under some conditions, one may be zero while the other one is different from zero almost everywhere and vice versa. Verri and Poggio (1987) showed exactly how and under what conditions these two quantities differ. In brief, $dI/dt = 0$ is only satisfied (using a perspective projection geometry), if a Lambertian surface translates rigidly through 3-D space. In general, the difference between (u, v) and (\dot{x}, \dot{y}) becomes smaller as ∇I becomes larger. Points at which ∇I is large are characterized by sharp changes in intensity—edges/features—that usually correspond to important physical events on surfaces, such as boundaries, orientation discontinuities, and especially surface markings. In other words, the difference between what one wants, i.e. the 2-D velocity field, and what one can get, i.e. the 2-D motion field, is small at these locations. Conversely, the difference becomes large if the surface markings are weak (no or little spatial contrast) or when the changes of brightness due to changes in viewing direction or illumination are rapid (e.g., when an object enters a sharp shadow or passes through a specular reflection).

These are many ways to obtain an estimate of the velocity field from optical flow. The image constraint equation $dI/dt = 0$ represents **only one** way to estimate the optical flow (for a discussion see Verri and Poggio, 1987 or Poggio, Yang and Torre, 1989). Its main merit is its simplicity.

In general, computing motion can be conceptualized as involving two separate stages:

- Local evaluation of the optical flow using a corresponding motion detector and
- Regularization of these noisy (at all locations), possibly sparse (at locations where $\nabla I \approx 0$) and possibly non-unique (for some definitions of optical flow) measurements.

Let us now turn to this second, motion integration/regularization step.

7 Computing Optical Flow Using Local Constancy Assumption

One manner in which the optical flow can be solved for is to assume that the optical flow is constant within a small neighbourhood of the point being considered (say, over a 5×5 image patch). This popular machine vision approach is due to Lucas and Kanade (1981). To see how this works, let us turn again to the brightness constancy assumption:

$$\frac{d}{dt}I(x, y, t) = \nabla I \cdot \mathbf{v} + \frac{\partial I}{\partial t} = 0. \quad (27)$$

If the local image gradient, ∇I is big enough, we can turn this scalar equation into a vector equation by multiplying on the left by ∇I^t :

$$\nabla I^t \nabla I \cdot \mathbf{v} = -\nabla I^t I_t. \quad (28)$$

We integrate this over a small window $W(x_0, y_0)$:

$$\int_{W(x_0, y_0)} \nabla I^t \nabla I \cdot \mathbf{v} dx dy = - \int_{W(x_0, y_0)} \nabla I^t I_t dx dy. \quad (29)$$

If we assume that the velocity \mathbf{v} is constant in each point of the window W containing $n \times n$ pixels, we end up having n^2 equations in two unknowns

$$M\mathbf{v} = e. \quad (30)$$

This overconstrained problem can be solved using a least-square approach.

At image locations with a strong contrast, the square matrix M can be inverted and we can estimate the local optical flow here

$$\mathbf{v} = -(\nabla I^t \cdot \nabla I)^{-1} \nabla I^t I_t. \quad (31)$$

This is the *de facto* standard to compute optical flow (Barron, Fleet and Beauchemin, 1994; Baker & Matthews, 2004).

One problem common to such differential techniques is that they break down when the displacement across frames is bigger than a few pixels, that is, bigger than the blurring length of the image. One way to overcome this inconvenience is to apply these techniques in a spatial *coarse-to-fine* manner according to the following steps:

- build a pyramid of images by smoothing and subsampling the original image
- select features at a coarse level
- compute the optical flow here (or track features)
- propagate this displacement to the next finer resolution using the displacement ($\mathbf{v} \times dt$) calculated at the previous stage as an initial guess for this finer stage.

8 References

- Adelson, E.H. and Bergen, J.R. Spatio-temporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**: 284-299 (1985).
- Baker S & Matthews I, Lucas-Kanade 20 years on: A unifying framework. *Intl. J. Comp. Vision* **56**: 221-25 (2004).
- Barron JL, Fleet, DJ & Beauchemin SS, Performance of optical flow techniques. *Intl. J. Comp. Vision* **12**: 43-77 (1994).
- Braddick, O. J. A short range process in apparent motion. *Vision Res.* **14**: 519-527 (1974).
- Braddick, O.J. Low-level and high-level processes in apparent motion. *Phil. Trans. R. Soc. Lond. B* **290**: 137-151 (1980).
- Chubb, C. and Sperling, G. Drift-balanced random stimuli: a general for studying non-Fourier motion perception. *J. Opt. Soc. Am. A* **5**: 1986-2007 (1988).
- DeAngelis, G.C., Ohzawa, I. and Freeman, R. D. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. *J. Neurophysiol.* **69**: 1118-1135 (1993).
- Diener, H.C., Wist, E.R., Dichgans, J. and Brandt, Th. The spatial frequency effect on perceived velocity. *Vision Res.* **16**: 169-176 (1975).
- Emerson, R.C., Bergen, J.R. and Adelson, E.H. Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Res.* **2**: 203-218 (1992).
- Fennema, C.L. and Thompson, W.R. Velocity determination in scenes containing several moving objects. *Comp. Graphics Image Proc.* **9**: 301-315 (1979).
- Gray, R. and Regan, D. Accuracy of estimating time to collision using binocular and monocular information. *Vision Res.* **38**: 499-512 (1998).
- Hassenstein, B. and Reichardt, W. Systemtheoretische Analyse der Zeit, Reihenfolgen, und Vorzeichenbewertung bei der Bewegungsperezeption des Rüsselkäfers *Chlorophanus*. *Z. Naturforsch.* **11b**: 513-524 (1956).
- Hess, R.H., Baker, C.L.Jr. and Zihl, J. The "Motion-blind" patient: low-level spatial and temporal filters. *J. Neurosci.* **9**: 1628-1640 (1989).
- Horn, BKP *Robot Vision*. MIT Press (1986).

Horn, BKP & Schunck BG, Determining optical flow. *Artif. Intell.* **17**: 185-203 (1981).

Koffka, K. *Principles of Gestalt Psychology*. Harcourt, Brace & World: New York (1935).

Longuet-Higgins, H.C. and Pradny, K. The interpretation of a moving retinal image. *Proc. R. Soc. Lond. B* **208**: 385-397 (1980).

Lu ZL & Sperling G Three-systems theory of human visual motion perception: review and update. *J. Opt. Soc. Am. A* **18(9)** 2331-2370 (2001).

Lucas, B. and Kanade, T. An iterative image registration technique with an application to stereo vision. *Proc. 7-th Intl. Joint Conf. Artif. Intell.*, Vancouver, pp. 674-679 (1981).

McKee, S.P. and Watamaniuk, S.N.J. The psychophysics of motion perception. In: *Visual detection of motion*, Smith, A.T. and Snowden, R.J., eds., pp. 85-114. Academic Press: San Diego, CA (1994).

Nagel, H.-H. On change detection and displacement vector estimation in image sequences. *Pattern Rec. Letters* **1**: 55-59 (1982).

Nakayama, K. and Silverman, G.H. The aperture problem: I. Perception of non-rigidity and motion direction in translating sinusoidal lines. *Vision Res.* **28**: 739-746 (1988).

Neuhaus, W. Experimentelle Untersuchung der Scheinbewegung. *Archiv gesamte psychologie* **75**: 315-418 (1930).

Poggio, T., Yang, W. and Torre, V. Optical flow: computational properties and networks, biological and analog. In: *The computing neuron*, Durbin, R., Miall, C. and Mitchison, G., eds., Addison-Wesley: Reading, MA (1989).

Poggio, T., Verri, A. and Torre, V. Green theorems and qualitative properties of the optical flow. *MIT AI Lab. Memo No. 1289* MIT, Cambridge MA (1991).

Ramachandran, V.S., Cronin-Golomb, A. and Myers, J.J. Perception of apparent motion by commissurotomy patients. *Nature* **320**: 358-359 (1986).

Ramachandran, V.S. and Gregory, R.L. Does color provide an input to human motion perception? *Nature* **275**: 55-56 (1978).

Royden, C. S. and Banks, M. S. and Crowell, J. A. The perception of heading during eye movements. *Nature* **360**: 583-585 (1992).

Spillmann, L. and Werner, J.S. *Visual Perception: The Neurophysiological Founda-*

tions. Academic Press (1990).

van Santen, J.P.H. and Sperling, G. Temporal covariance model of motion perception. *J. Opt. Soc. Am. A* **1**: 451-473 (1984).

van Santen, J.P.H. and Sperling, G. Elaborated Reichardt detectors. *J. Opt. Soc. Am. A* **2**: 300-320 (1985).

Verri, A. and Poggio, T. Motion field and optical flow: qualitative properties. *IEEE PAMI* **11**: 490-498 (1989).

Warren, W.H. and Hannon, D.J. Direction of self-motion is perceived from optical flow. *Nature* **336**: 162-164 (1988).

Watson, A.B. and Ahumada, A.J. Model of human visual-motion sensing. *J. Opt. Soc. Am. A* **2**: 322-341 (1985).

Wertheimer, M. Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung* **4**: 301-350 (1923).

Zeki, S.M. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *J. Physiol.* **236**: 549-573 (1974).