Local models for binocular stereo (I)

- Linear filter models of receptive fields can also be used to perform local estimates of binocular stereo and motion. These models involve having filterbanks, or populations of filters, that are tuned to different properties of the stimuli, so that estimates of depth and motion can be extracted from the population (Zhaoping, 2014).
- Recall that we introduced binocular stereo earlier. Depth is estimated by triangulation provided we can solve the *correspondence problem* by finding which points in the left and right eyes correspond to the same point in three-dimensional space. This reduces to estimating the displacement, or *disparity*, between the images in the left and right eyes. In this section, we introduce the disparity energy model, which estimates disparity based on local properties of the image. Later we will discuss how nonlocal context can be used to improve disparity estimation.

Local models for binocular stereo (II)

- The disparity energy model is formulated using Gabor filters and has some claim to biological plausibility (Ohzawa et al., 1990; Qian, 1994). The model assumes that we have a large set of cells, receiving input from both images and tuned to different image frequencies and spatial phases.
- We give the presentation in one dimension, exploiting the epipolar line constraint. It assumes that the cell receives input from both left and right eyes with receptive fields $f_l(x) = \exp\{-x^2/(2\sigma^2)\}\cos(\omega x + \rho_l)$ and $f_r(x) = \exp\{-x^2/(2\sigma^2)\}\cos(\omega x + \rho_r)$. These are Gabors where the Gaussian has variance σ^2, tuned to frequency ω and with phases ρ_l, ρ_r. The linear response is:

$$r = \int dx\{f_l(x)I_l(x) + f_r(x)I_r(x)\} \tag{10}$$

- This filter is tuned to spatial frequency ω. The filter is most sensitive to the image component at this frequency. Hence we can represent the image (approximately) by $I(\vec{x}) = A\cos(\omega x + \theta)$, where A is the amplitude and θ is the phase.

Local models for binocular stereo (III)

- Suppose that the right image is a displaced version of the left image $I_r(x) = I_l(x + D(x))$, where $D(x)$ is the disparity. We assume that the disparity varies slowly so that we can approximate it locally as a constant D (over the size of the Gaussian, 2σ). To analyze the model, ignore the Gaussian when calculating r. This gives:

$$r_1 = A\{\cos(\theta - \rho_l) + \cos(\theta - \rho_r - \omega D)\} \tag{11}$$

  which can be re-expressed (using trigonometry identities):

$$r_1 = 2A\cos(\theta - \frac{\rho_l + \rho_r}{2} - \frac{\omega D}{2})\cos(\frac{\rho_l - \rho_r}{2} - \omega\frac{D}{2}) \tag{12}$$

- The response of the cell depends on the disparity but also on image properties (e.g., image phase θ). So we need a population of cells to detect disparity.

Lcoal models for binocular stereo (IV)

- To see this, suppose that we consider quadrature pairs of the two cells tuned to the same ω. Where one cell has phases ρ_l, ρ_r, and the other has phases ρ'_l, ρ'_r, where $(\rho_l - \rho_r) = (\rho'_l - \rho'_r)$ and $\rho'_l + \rho'_r = \rho_l + \rho_r - \pi$. Then the second cell has response
$r_2 = 2A\cos(\theta - \frac{\rho_l+\rho_r}{2} + \frac{\pi}{2} - \frac{\omega D}{2})\cos(\frac{\rho_l-\rho_r}{2} - \omega\frac{D}{2}) =$
$2A\sin(\theta - \frac{\rho_l+\rho_r}{2} - \frac{\omega D}{2})\cos(\frac{\rho_l-\rho_r}{2} - \frac{\omega D}{2})$. Hence if we square and add the responses of the two cells, we obtain:

$$r_1^2 + r_2^2 = A^2\cos^2(\frac{\rho_l - \rho_r}{2} - \omega\frac{D}{2}) \tag{13}$$

- This response depends only on the disparity D and the image frequency ω. It takes largest values when $\rho_l - \rho_r = \omega D$. Hence we can estimate D from a population of quadrature cells tuned to different phases ρ_l, ρ_r and frequencies ω.

Local models for binocular stereo (V)

- A neural network for estimating D using a population of neurons consists of two steps. In step (1) we define a set of disparity cells tuned to disparities $\{D_i : i = 1, ..., N\}$. The disparity cell tuned to disparity D_i receives input $\cos^2(\frac{\rho_l - \rho_r}{2} - \omega\frac{D_i}{2})$ from each quadrature pair (ρ_l, ρ_r, ω) and sums these inputs together to compute a vote $v(D_i)$:

$$v(D_i) = \sum_{\rho_l, \rho_r, \omega} \cos^2(\frac{\rho_l - \rho_r}{2} - \omega\frac{D_i}{2}). \tag{14}$$

  Step (2) uses a winner-take-all network (Maass, 2000) to compute the disparity with the biggest vote by solving $\hat{D} = \arg\max_{i=1,...,N} v(D_i)$, so that $v(\hat{D}) \geq v(D_i)$ for $i = 1, ..., N$.

- There is plenty of evidence that the brain represents information by neural populations (Georgopoulos et al., 1983; McIlwain, 1991). There have also been several theoretical studies of how populations of neurons could encode knowledge and perform computations (Pouget et al., 2003; Ma et al., 2006).

Illustration of local model of binocular stereo

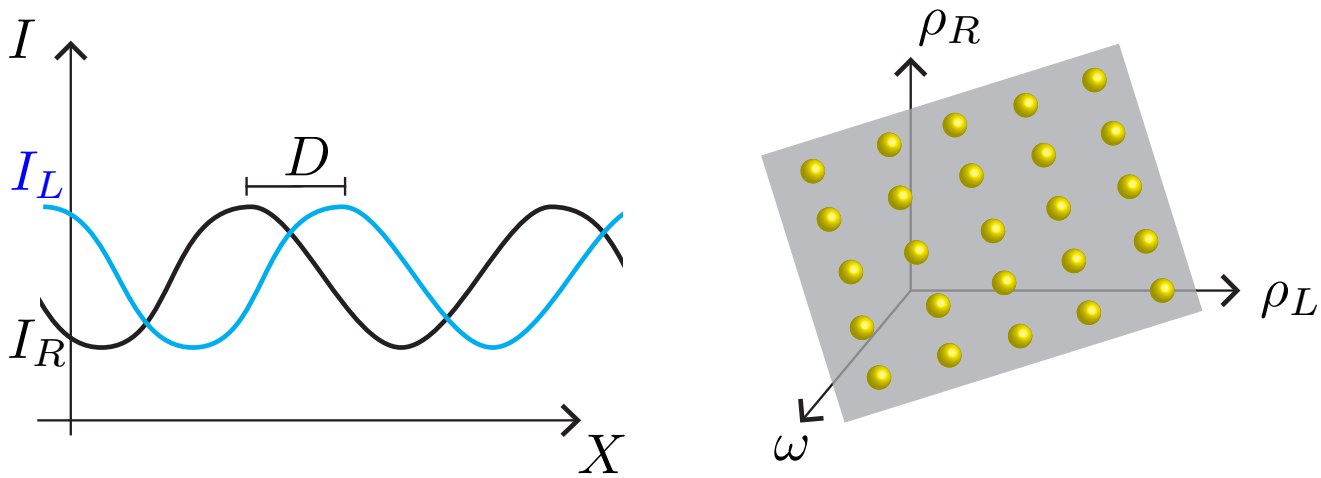

Figure 19: Left: The disparity D between the images in the two eyes corresponds to a change of phase if we approximate the intensities by sinusoids. Right: The local disparity D is encoded by the feature response of cells tuned to frequencies that obey $\rho_l - \rho_r = \omega D$.

Motion measurement: Spatio-temporal filters.

We now discuss how related models can be used to estimate motion for sequences of images. Spatiotemporal filters are biologically plausible ways to measure motion that agree with properties of cells in the visual cortex. The standard model suggests two classes of cells: the first comprises spatiotemporal filters that are sensitive to the directions of motion, while the second class combines outputs of these filters to estimate the motion itself (Adelson & Bergen, 1985; Grzywacz & Yuille, 1990; Schrater et al., 2000).

Motion measurement: Figures

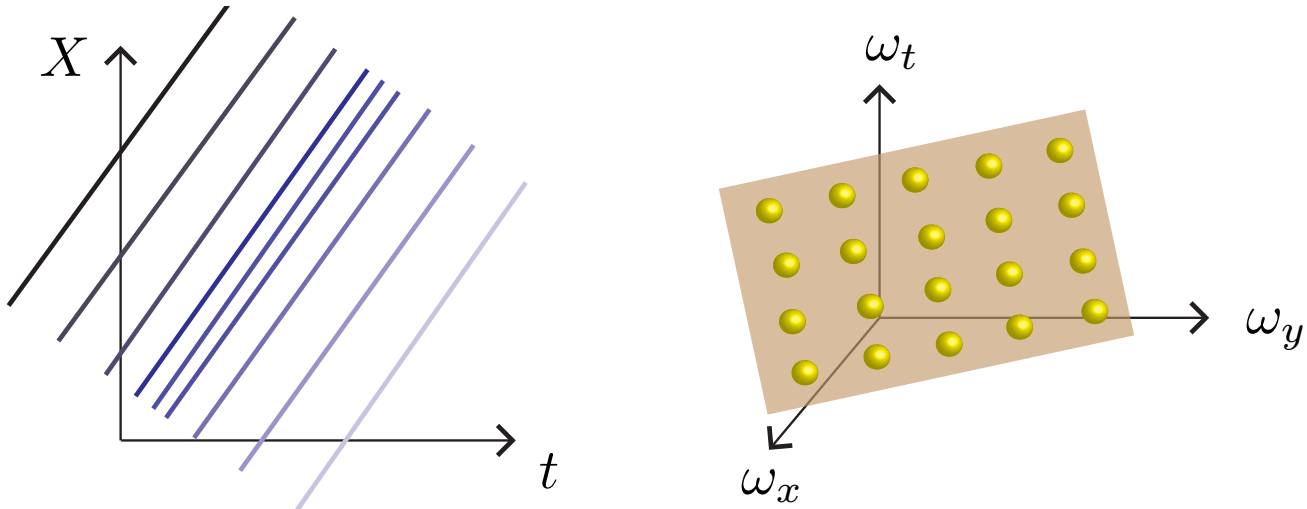

Figure 20: Left: This figure shows the space-time illustration of a signal traveling with constant velocity $I(X,t) = F(X - tv)$. This means that the intensity $I(X,t)$ is constant on the lines $X - tv = constant$. Right: A stimuli moving with velocity $\vec{v}$ will activate spatiotemporal filters $\vec{\omega}, \omega_t$, which lie on the plane $\vec{v} \cdot \vec{\omega} + \omega_t = 0$. Hence the velocity can be estimated from the population of activity of the filters.

Motion measurement (I)

- Measuring the motion velocity assumes that locally, the intensity can be modeled as a linear translating pattern:

$$I(\vec{x}, t) = F(\vec{x} - \vec{v}t). \tag{15}$$

- Differentiating with respect to $\vec{x}$ and t (using $\vec{\nabla} I = \vec{\nabla} F$ and $\frac{\partial I}{\partial t} = -\vec{v} \cdot \vec{\nabla} F$) gives the *optical flow equation*:

$$\vec{v} \cdot \vec{\nabla} I + \frac{\partial I}{\partial t} = 0. \tag{16}$$

- This enables us to estimate one component of the motion $\vec{v}$ but suffers from the aperture problem and so is ambiguous.

Motion measurement (II)

- The ambiguity can be resolved by a population of filters $\{G^{\mu}(\vec{x}, t) : \mu = 1, ..., M\}$ indexed by μ (e.g., Gaussians). These filters introduce local context:

$$G^{\mu} * I(\vec{x}, t) = \int G^{\mu}(\vec{x} - \vec{y}, t - s) I(\vec{y}, s) ds d\vec{y}. \tag{17}$$

  Each filter gives a constraint on the velocity:

$$\vec{v} \cdot \vec{\nabla} G^{\mu} * I + \frac{\partial G^{\mu} * I}{\partial t} = 0. \tag{18}$$

- We get an estimate of the velocity $\vec{v}$ by minimizing the cost:

$$E(\vec{v}) = \sum_{\mu=1}^{M} (\vec{v} \cdot \vec{\nabla} G^{\mu} * I + \frac{\partial G^{\mu} * I}{\partial t})^2.$$

- This minimization can be done using a similar neural network to that used for estimating disparity for stereo in the previous section.

Motion measurement (III)

We have a set of cells tuned to different velocities $\{\vec{v}_i : i = 1, ..., N\}$. The cell tuned to velocity $\vec{v}_i$ receives input $(\vec{v} \cdot \vec{\nabla} G^\mu * I + \frac{\partial G^\mu * I}{\partial t})^2$ from each filter μ and sums the responses to obtain $E(\vec{v}_i)$. Then we use a variant of winner-take-all to compute $\hat{\vec{v}} = \arg\min_{i=1,...,N} E(\vec{v}_i)$.

Motion measurement: The need for spatial and temporal context

This approach assumes that there is enough local information to resolve the motion ambiguity which may not be the case. For example, for the stimuli in figure 12.7 in the chapter, we can only locally estimate one component of the motion because of the aperture problem. To resolve this ambiguity, we need to use more spatial or temporal context.

Motion measurement: Spatial and temporal context (I)

An alternative way to analyze this problem is by applying Ffourier analysis to equation (15):

$$\hat{I}(\vec{\omega}, \omega_t) = \frac{1}{2\pi} \int \int \int \exp\{i(\vec{\omega} \cdot \vec{x} + \omega_t t)\} I(\vec{x}, t) d\vec{x} dt$$

$$\hat{I}(\vec{\omega}, \omega_t) = \frac{1}{2\pi} \int \int \int \exp\{i(\vec{v} \cdot \vec{x} + \omega_t) t\} \exp\{i\vec{\omega} \cdot (\vec{x} - \vec{v}t)\} F(\vec{x} - \vec{v}t) d\vec{x} dt$$

$$\hat{I}(\vec{\omega}, \omega_t) = \frac{1}{2\pi} \int \exp\{i(\vec{v} \cdot \vec{\omega} + \omega_t) t\} dt \int \int \exp\{i\vec{\omega} \cdot \vec{\bar{x}}\} F(\vec{\bar{x}}) d\vec{\bar{x}}$$

$$\hat{I}(\vec{\omega}, \omega_t) = \delta(\vec{v} \cdot \vec{\omega} + \omega_t) \hat{F}(\vec{\omega})$$

where $\vec{\bar{x}} = \vec{x} - \vec{v}t$ is a change of variables in the integral.

Motion measurement: Spatial and temporal context (II)

This shows that if we have filters $\exp\{i(\vec{x}\vec{\omega} + \omega_t t)\}$ tuned to spatiotemporal frequencies $\vec{\omega}, \omega_t$, then the only filters that respond are those whose frequencies obey the equation $\vec{v} \cdot \vec{\omega} + \omega_t = 0$ and hence lie on a plane in frequency space. Hence we can determine $\vec{v}$ from a population of filters by observing which filters are activated and finding the best fit plane.

Motion measurement – Non-Fourier

- In practice, we cannot use filters tuned to frequency because these are not bounded in space and time. But it can be shown (Grzywacz & Yuille, 1990) that if the filters are spatio-temporal Gabors, then the most active filters are those whose spatiotemporal tuning is centered on the plane $\vec{v} \cdot \vec{\omega} + \omega_t = 0$. Hence the plane in frequency space can be estimated from a population of spatiotemporal filters and the velocity locally estimated.
- This gives a two stage model of motion estimation, in which the first population of neurons (i.e., filters) are each sensitive to the spatiotemporal frequency of the input image but not directly to the motion. The second population of neurons extract the motion information from the first population, and hence these neurons are tuned directly to motion. This is consistent with experimental findings (Adelson & Bergen, 1985),(Grzywacz & Yuille, 1990), (Schrater et al., 2000). Similar models arise in related work on the fly and beetle visual systems (Hassenstein & Reichardt. 1956; Borst & Euler, 2011).