Binocular stereo

- Binocular stereo is another vision module, which estimates the depth and orientation of surfaces. Humans have the ability to get depth from two eyes – hence the popularity of so-called 3D movies.
- This requires solving a *correspondence problem* between features in the two eyes which are caused/imaged by the same point in space. If correspondence can be performed, to estimate the *disparity* between points in the left and right images, then depth can be estimated by trigonometry. The correspondence problem is made easier by the *epipolar line constraint*, which means that corresponding points require only searching in a one-dimensional direction.
- But knowledge of the epipolar lines requires knowing the direction of gaze of the cameras (maybe done by feedback from muscles controlling the eyes, or by *calibration*). Note that partial occlusion can happen, when part of the scene is visible to one only eye. Da Vinci was the first to point out that this was a useful visual cue.
- We will discuss issues like calibrating cameras and estimating depth from disparity in later lectures. See Frisby and Stone for what is known about how humans perform these tasks.

Stereo figure

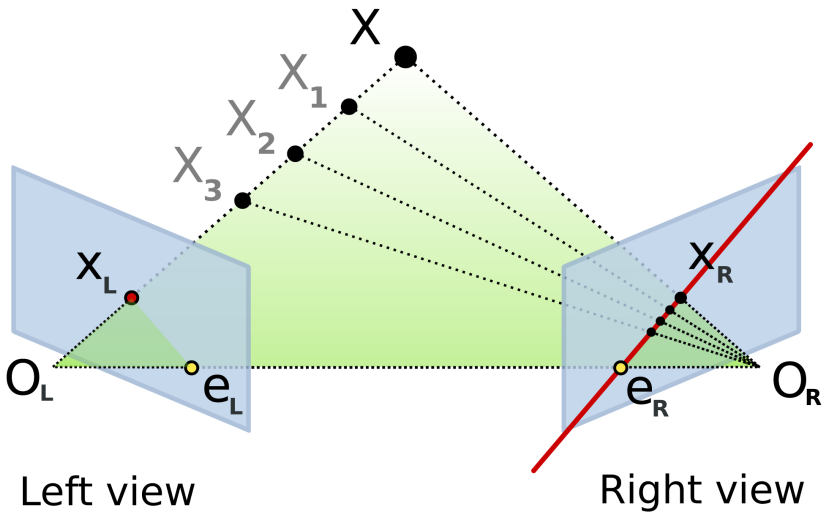

Figure 19: Stereopsis and epipolar lines. A point $\vec{x}$ in three-dimensional space gets projected onto positions x_L and x_R in the left and right eyes. This uses a pinhole camera model of each eye, where the eye is specified as a plane (in grey), and O_L and O_R represent the centers of projection. All points on the plane defined by O_L, O_R, and $\vec{x}$ get projected onto straight lines $\vec{e}_L$ and $\vec{e}_R$, the corresponding *epipolar lines*, in the two eyes, as shown by the projections of x_1, x_2, x_3 onto the right eye. If we alter the position of the point $\vec{x}$ in space, then we will get a family of corresponding epipolar lines. The *epipolar line constraint* states that points on an epipolar line in one eye can only be matched to a point in the other eye on the corresponding epipolar line.

Binocular stereo: Local and Low/Mid Level

- Is binocular stereo a high level task? I.e. does it require recognizing objects in each image and then matching them to solve the correspondence problem? *Not necessarily, Julesz random dot stereograms showed that humans could perceive depth even if the input images are random dot stimuli with no apparent structure*. There has been more than forty years study of study on low/mid level stereo algorithms.
- These low/mid level algorithms all have the same two ingredients: (I) A component that estimates the correspondence locally (see the rest of this lecture). (II) A component that imposes geometry assumptions about the shape of the object being viewed. Most commonly that the surface is piecewise smooth, but sometimes that the world consists of planar surfaces (Manhattan stereo).
- But other studies show that humans do exploit object specific cues when they do binocular stereo. For example, when humans look at an inverted face mask they perceive it as a normal face (despite binocular stereo cues indicating that they are looking at an inverted mask). Hence the human visual system uses a combination of low, mid, and high level cues when performing binocular stereo.
- Why is this a good idea? There are hazardous factors, see next slide, which cause stereo algorithms which use low and mid level cues to break down. In such cases, recognizing an object and using object specific knowledge is a sensible strategy. For example, to use knowledge of the 3D structure of a car to deal with specularity (a nuisance factor).

Binocular stereo: Hazard Factors

(a) Specularity (b) Texturelessness (c) Transparency (d) Disparity jumps

Figure 20: Binocular stereo algorithms tend to fail in regions where there is transparency, specularity, lack of texture and very thin objects. These image regions are called hazardous regions. For example, a street light is a thin object and covers a small region of an image, but missing it could be a disaster for autonomous driving. Cars are highly specular, particularly when it is raining, and so could defeat most standard stereo algorithms. Flat textureless regions also give stereo algorithms no local cues to estimate depth but, in the special case, where the surfaces are planar then algorithms like Manhattan stereo may be effective. Transparency is another problem since we want stereo algorithms to work even when we view objects through glass (increasingly common in modern buildings).

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

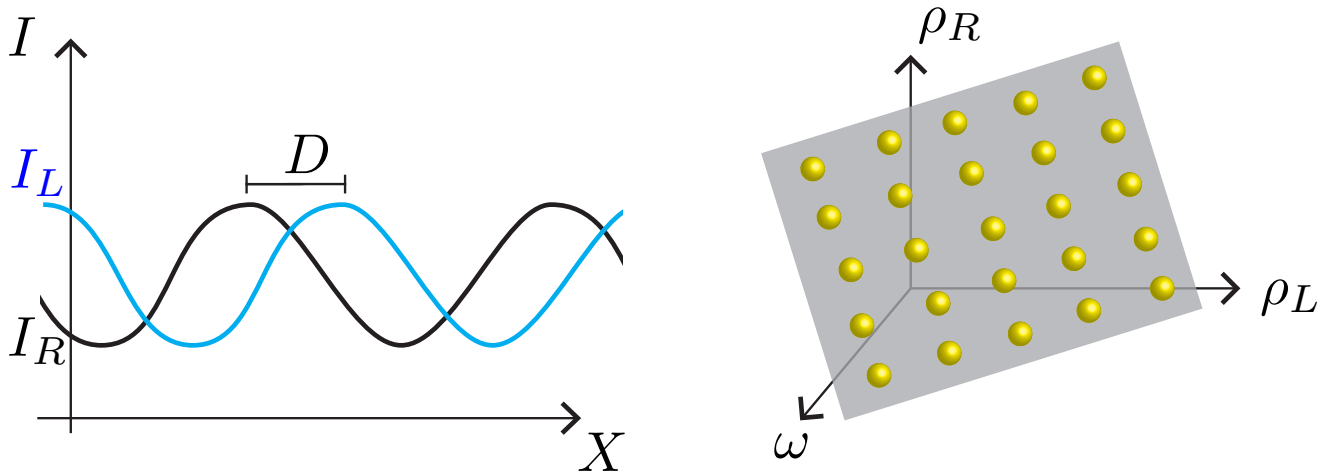

Figure 21: Left: The disparity D between the images in the two eyes corresponds to a change of phase if we approximate the intensities by sinusoids. Right: The local disparity D is encoded by the feature response of cells tuned to frequencies that obey $\rho_l - \rho_r = \omega D$.

Low Level Cues for binocular stereo

- This type of stereo model can be loosely related to properties of neurons in the visual cortex. But there are no very effective stereo algorithms that use these types of cues. Computer vision models tend to use more complex image features (and smoothness assumptions, see later in the course).
- Currently most stereo algorithms use deep networks to learn corresponding features. These perform well but can be sensitive to hazard factors. Perhaps the brain using these types of non-linear filters, instead of the linear filter described here? Recall, that linear filters are an approximate model of real neurons (and maybe not a good one).
- Nevertheless, these linear models of stereo illustrate the types of local cues that all stereo models assume (though problematic for highly specular objects).
- We will return to binocular stereo later in the course when we discuss models that impose geometric assumptions such as piecewise smoothness.