Lecture: Motion Perception

- The barberpole illusion shoes that the perception of motion is not straightforward. The barberpoles rotate to the right, but the perception of motion is vertically upwards. This is because there may not be enough information to determine the motion unambiguously.
- Consider a moving bar. We can observe the motion in the direction perpendicular to the bar. But we cannot observe the motion along the bar. So the local observation is consistent with many possible motions.
- This is called the *aperture problem*. At the endpoints of the bar the motion will seem to be unambiguous. But the observations at the endpoints have to propagate to the other points on the bar. How is this done? How far can information at unambiguous points be propagated?
- Consider a rotating ellipse. This has no ambiguous points. It is perceived either as: (i) a non-rigid rotating ellipse, or (ii) a rigid circle rotating in 3D. But, surprisingly, it is not seen as a rigidly rotating ellipse unless the aspect ratio is very big (because now it appears to have endpoints).
- Nakayama performed a series of experiments which studied the effects of occluding the endpoints of the bars (so that the ends are not visible), or having isolated dots move near the bar, and other variants. This showed that perception is very subtle and that the motion of the bar could be *captured* by the motion of the dots. Many of these effects could be modelled by the motion coherence theory (see Yuille and Grzywacz handout later in the course) which claims that motion is perceived by combining the local measurements with a prior that the motion is slow

Short-range and long-range motion

- The human eye receives input images as a continuous stream in time $I(\vec{x}, t)$, where t is continuous. Typical videos have 24 frames per second (recent movies, e.g., the Hobbit, tried 48 frames per second but moviegoers complained that this looked weird). Curiously dogs may not be able to see motion at 24 frames per second, but humans can. Humans can even perceive motion with far fewer frames per second.
- Short-range motion is defined to be situations where the image frames are so close together that it is difficult to distinguish them from continuous frames. This will means that the intensity is differentiable (as we will discuss later). In such cases we have the aperture problem.
- Long-range motion occurs there is a significant difference between neighboring image frames. In this case the images are not differentiable. Instead we have to solve a correspondence problem between features in the two images. This is similar to the binocular stereo correspondence problem, except it does not have an epipolar line constraint.
- For short-range motion, we assume that $I(\vec{x}, t) = F(\vec{x} - \vec{v}t)$ where $\vec{v}$ is the motion and we assume that the motion is locally rigid in the image plane with intensity $F(.)$. Differentiating $I(\vec{x}, t)$ with respect to $\vec{x}$ and t gives: $\vec{\nabla} I(\vec{x}, t) = \vec{\nabla} F(\vec{x} - \vec{v}t)$ and $\frac{\partial I(\vec{x},t)}{\partial t} = -\vec{v} \cdot \vec{\nabla} F(\vec{x} - \vec{v}t)$. This yields the *optical flow equation* $\vec{v} \cdot \vec{\nabla} I(\vec{s}, t) + \frac{\partial I(\vec{x},t)}{\partial t} = 0$. In other words, we can directly measure the motion component in the direction of the image gradient, but we do not know the perpendicular component.

Long-range motion

- Minimal mapping theory (Ullman) formulates long-range motion as finding the correspondence between points/dots $\{\vec{y}_a\}$ in the first image to points $\{\vec{x}_i\}$ in the second (all the dots are indistinguishable).
- There is a correspondence variable $\{V_{ai}\}$ so that $V_{ai} = 1$ if the point at $\vec{y}_a$ in the first image is matched to the point $\vec{x}_i$ in the second image. $V_{ai} = 0$ otherwise, and we impose conditions that all dots in the first image must be matched to exactly one dot in the second image (can be relaxed if the number of dots in the two images is different). This is an example of a matching problem. Solving it requires making assumptions about the smoothness of the motion, which we will discuss later in the course.
- Barlow and Tripathy studied human perception of long-range motion. They used an experimental setup where the first image consisted of a set of random dots. The second image was generated by taking a subset of the dots in the first image and moving them horizontally by a fixed amount (randomly chosen) and filling up the rest of the image with randomly placed dots.
- They then tested human performance at several visual tasks, e.g., judging whether the dots moved to the right or to the left, and they compared performance to an *ideal observer model* that knew how the two images had been generated (statistically). Not surprisingly, humans performed much worse (by many orders of magnitude to the ideal observer). This was not a fair test. It can be shown, see later in the course, that human perception is similar to that of vision algorithms that solve the

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

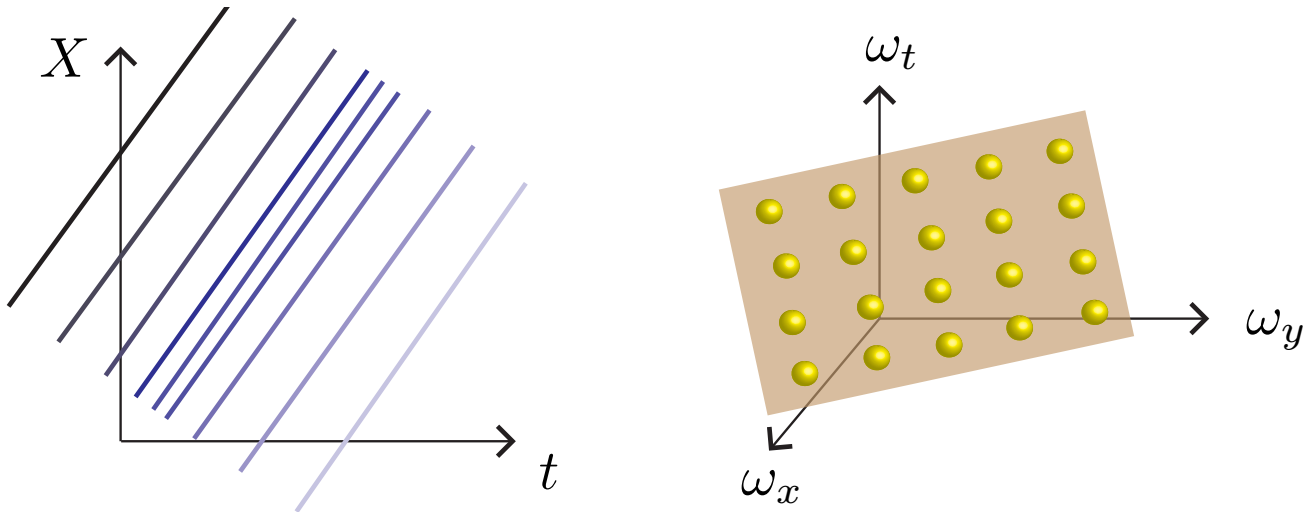

Figure 22: Left: This figure shows the space-time illustration of a signal traveling with constant velocity $I(X,t) = F(X - tv)$. This means that the intensity $I(X,t)$ is constant on the lines $X - tv = constant$. Right: A stimuli moving with velocity $\vec{v}$ will activate spatiotemporal filters $\vec{\omega}, \omega_t$, 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 – Relaxing the Fourier Assumption

- 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).

Motion measurement – Summary

- As for binocular stereo, most motion algorithms combine local measurement cues with assumptions about properties of the motion field. For example, that the motion field is smooth or that it is due to a rigidly rotation object. Some motion algorithms also assume that motion flow is temporally coherent (e.g., the direction of motion changes little over time).
- But very few algorithms use the types of spatio-temporal features presented here. There are two reasons: (i) Most motion algorithms use two image time frames only (for practicality), (II) Most motion algorithms use non-linear filters. Increasing learnt by neural networks (from datasets with ground truth motion known).
- Motion algorithms share the same hazardous factors with binocular stereo algorithms. They produce poor results, for example, if you observe a rigidly rotating specular object (we will discuss specularity in more detail later in the course). As for binocular stereo, humans may deal with these difficult cases by recognizing and using knowledge about objects.