Probabilities on Graphs: Undirected Graphs and Spatial Context

Alan Yuille and Dan Kersten

May 6, 2015

Lecture 12.5

- This lecture describes how groups of neurons can perform edge detection, edge grouping, stereo, and motion.
- We also introduce weak methods for cue combination.
- This lecture includes the exercises: (12.5.1) Hopfield network for binocular stereo, and (12.5.2) Cue combination.

An Example: Foreground Background Segmentation

- Now combine the prior and likelihood to estimate the foreground and background. Formulate the problem as Bayes estimation with conditional distributions $P(f(I(x))|s)$ and priors $P(s)$ for $s \in \{0, 1\}$. The posterior distribution $P(s|f(I(x)))$ can be expressed in the form:

$$P(s|f(I(x))) = \frac{1}{Z} \exp\{s(\log \frac{P(f(I(x))|s=1)}{P(f(I(x))|s=0)} + \log \frac{P(s=1)}{P(s=0)})\},$$

  where Z is a normalization constant (chosen so that $\sum_{s=0}^{1} P(s|f(I(x))) = 1)$.
- This shows that the posterior distribution for the presence of pixels being foreground can be expressed in the same form. The only difference is that the input is a nonlinear function of the image instead of the image itself.
- This claim can be justified by expressing $P(f(I(x))|s) = \{P(f(I(x))|s=1)\}^{s}\{P(f(I(x))|s=0)\}^{1-s}$, $P(s) = \{P(s=1)\}^{s}\{P(s=0)\}^{1-s}$, then substituting these into the posterior $P(s|f(I(x))) = P(f(I(x))|s)P(s)/P(f(I(x)))$.

Probability models with context

- Now apply the model to foreground/background classification and modify it to include spatial context. Intuitively, neighboring pixels in the image are likely to be either all background or all foreground. This is a form of prior knowledge that can be learned by analyzing natural images.
- We specify neurons by spatial position $\vec{x}$ instead of index i. As above, we have distributions $P(f(I(\vec{x}))|s)$ for the features $f(I(\vec{x}))$ at position $\vec{x}$ conditioned on whether this is part of the foreground object $s(\vec{x}) = 1$, or not, $s(\vec{x}) = 0$. We use the notation $\vec{S}$ to be the set of the states of all neurons $\{s(\vec{x})\}$. We also specify a prior distribution:

$$P(\vec{S}) = \frac{1}{Z} \exp\{-\gamma \sum_{\vec{x}} \sum_{\vec{y} \in N(\vec{x})} \{s(\vec{x}) - s(\vec{y})\}^2\},$$

where γ is a constant. This prior uses a neighborhood $N(\vec{x})$, which specifies those spatial positions that directly interact with $\vec{x}$ in the model. In graphical terms, the positions $\vec{x}$ are the nodes $\mathcal{V}$ of a graph $\mathcal{G}$, and the edges $\mathcal{E}$ specify which nodes are connected.

Posterior distribution (I)

- These distributions $P(f(\vec{I})|\vec{S})$ and $P(\vec{S})$ can be combined to get the posterior distribution $P(\vec{S}|f(\vec{I}))$, which is of form:

  $$P(\vec{S}|f(\vec{I})) = \frac{1}{Z_p} \exp\{-E(\vec{S})\},$$

  where

  $$E(\vec{S}) = -\sum_{\vec{x}} s(\vec{x}) \log \frac{P(f(I(\vec{x}))|s=1)}{P(f(I(\vec{x}))|s=0)} + \sum_{\vec{x}} \sum_{\vec{y} \in N(\vec{x})} \gamma \{s(\vec{x}) - s(\vec{y})\}^2.$$

- The first term of $E(\vec{S})$ gives the local cues for foreground or background (the log-likelihood ratios of the features), while the second term adds the local context. This context encourages neighboring positions to be either all foreground or all background. Note that this method of specifying a distribution $P(\vec{S})$ in terms of a function $E(\vec{S})$ will keep reoccurring throughout this section.

Posterior distribution (II)

- This model specifies the posterior distribution for foreground-background classification using spatial context, and as we will show, similar methods can be applied to other visual tasks. But there remains the issue of how to estimate the most probable states, i.e., computing the Bayes estimator.

$$\hat{\vec{S}} = \arg\max P(\vec{S}|f(\vec{I})).$$

- In the next two sections we discuss neurally plausible algorithms that can do this. There are two types: (1) stochastic models that are natural extensions of the probabilistic neural models discussed earlier, which in the statistics literature are called *Gibbs* samplers (Liu, 2008), and (2) neural network models that are based on simplified biophysics of neurons but that can also, in certain cases, be related to *mean field approximations* to the stochastic models.

The line process model (I)

- Our first example is the classic *line process* model (Geman & Geman, 1984; Blake & Zisserman, 2003; Mumford & Shah, 1989), which was developed as a way to segment images. It has explicit *line process* variables that "break" images into regions where the intensity is piecewise smooth. Our presentation follows the work of Koch et al. (1986), who translated it into neural circuits.
- The model takes intensity values $\vec{I}$ as input, and outputs smoothed intensity values. But this smoothness is broken at places where the intensity changes are too high. The model has continuous variables $\vec{J}$ representing the intensity, and binary-valued variables $\vec{l}$ for the line processes (or edges). The model is formulated as performing *maximum a posteriori* (MAP) estimation. The algorithm for estimating MAP is a neural network model that can be derived from the original Markov model (Geman & Geman, 1984) by mean field theory (Geiger & Yuille, 1991). Note that in this model, the variables do not have to represent intensity. Instead they can represent texture, depth, or any other property that is spatially smooth except at sharp discontinuities.

The line process model (II)

- For simplicity we present the weak membrane model in one dimension. The input is $\vec{I} = \{I(x) : x \in \mathcal{D}\}$; the estimated, or smoothed, image is $\vec{J} = \{J(x) : x \in \mathcal{D}\}$; and the line processes are denoted by $\vec{l} = \{l(x) : x \in \mathcal{D}\}$, where $l(x) \in \{0, 1\}$.
- The model is specified by a posterior probability distribution:

$$P(\vec{J}, \vec{l} | \vec{I}) = \frac{1}{Z} \exp\{-E[\vec{J}, \vec{l} : \vec{I}]/T\},$$

where

$$E[\vec{J}, \vec{l} : \vec{I}] = \sum_x (I(x) - J(x))^2 + A \sum_x (J(x+1) - J(x))^2 (1 - l(x)) + B \sum_x l(x).$$

The line process model (III)

The first term ensures that the estimated intensity $J(x)$ is close to the input intensity $I(x)$. The second encourages the estimated intensity $J(x)$ to be spatially smooth (e.g., $J(x) \approx J(x+1)$), unless a line process is activated by setting $I(x) = 1$. The third pays a penalty for activating a line process. The result encourages the estimated intensity to be piecewise smooth unless the input $I(x)$ changes significantly, in which case a line process is switched on and the smoothness is broken. The parameter T is the variance of the probability distribution and has a default value $T = 1$.

The line process model illustration

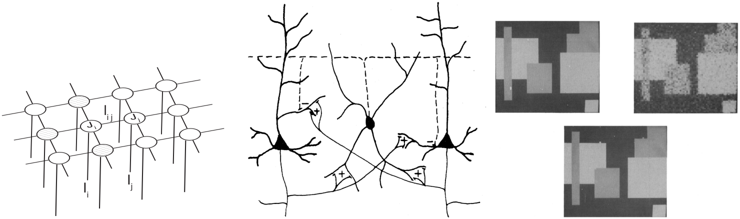

Figure 1: A representation of the line process model (left) compared to a real neural network (center). On the right, the original image (upper left), the image corrupted with noise (upper right), and the image estimated using the line process model (bottom).

The line process model and neural circuits (I)

- This model can be implemented by a neural circuit (Koch et al., 1986). The connections between these neurons is shown in the previous figure. To implement this model Koch et al., (1986) proposed a neural net model that is equivalent to doing mean field theory on the weak membrane MRF (as discussed earlier) by replacing the binary-valued line process variables $l(x)$ by continuous variables $q(x) \in [0, 1]$ (corresponding roughly to the probability that the line process is switched on).
- This gives an algorithm that updates the regional variables $\vec{J}$ and the line variables $\vec{q}$ in a coupled manner. It is helpful, as before, to introduce a new variable $\vec{u}$ which relates by $q(x) = \frac{1}{1+\exp\{-u(x)/T\}}$ and $u(x) = T \log \frac{q(x)}{1-q(x)}$.

The line process model and neural circuits (II)

$$\frac{dJ(x)}{dt} = -2(J(x) - I(x))$$

$$= -2A\{(1-q(x))(J(x) - J(x+1)) + (1-q(x-1))(J(x) - J(x-1))\}, \quad (1)$$

$$\frac{dq(x)}{dt} = \frac{1}{T} q(x)(1-q(x))\{A(J(x+1) - J(x))^2 - B - T\log\frac{q(x)}{1-q(x)}\}, \quad (2)$$

$$\frac{du(x)}{dt} = -u(x) + A(J(x+1) - J(x))^2 - B. \quad (3)$$

The update rule for the estimated intensity $\vec{J}$ behaves like nonlinear diffusion, which smooths the intensity while keeping it similar to input $\vec{I}$. The diffusion is modulated by the strength of the edges $\vec{q}$. The update for the lines $\vec{q}$ is driven by the differences between the estimated intensity; if this is small, then the lines are not activated.

The line process model and neural circuits (III)

This algorithm has a Lyapunov function $L(\vec{J}, \vec{q})$ (derived using mean field theory methods) and so will converge to a fixed point, with

$$L(\vec{J}, \vec{q}) = \sum_x (I(x) - J(x))^2 + A \sum_x (J(x+1) - J(x))^2 (1 - q(x)) + B \sum_x q(x) \\ + T \sum_x \{q(x) \log q(x) + (1 - q(x)) \log(1 - q(x))\}. \quad (4)$$

Relations to electrophysiology (I)

- There is some evidence that a generalization of this models roughly matches the electrophysiological findings for those types of stimuli. The generalization is performed by replacing the intensity variables $I(x), J(x)$ by a filterbank of Gabor filters so that the weak membrane model enforces edges at places where the texture properties change (Lee et al., 1992). The experiments, and their relation to the weak membrane models are reviewed in (Lee & Yuille, 2006). The initial responses of the neurons, for the first 80 msec, are consistent with the linear filter models described earlier. But after 80 msec, the activity of the neurons changes and appears to take spatial context into account.
- While the weak membrane model is broadly consistent with the perceptual phenomena of segmentation and "filling in," the types of filling in, their dynamics, and the neural representations of contours and surface are complicated (von der Heydt, 2002; Komatsu, 2006). Exactly how contour and surface information is represented and processed in cortex is an active topic of research (Grossberg & Hong, 2006; Roe et al., 2012).

Relations to electrophysiology (II)

- The findings of the electrophysiological experiments are summarized as follows:
  (1) There are two sets of neurons, with one set encoding regional properties (such as average brightness), and the other set coding boundary location (in agreement with J and I variable in the model, respectively).
  (2) The processes for computing the region and the boundary representations are tightly coupled, with both processes interacting with and constraining each other (as in the dynamical equations above).
  (3) During the iterative process, the regional properties diffuse within each region and tend to become constant, but these regional properties do not cross the region (in agreement with the model).
  (4) The interruption of the spreading of regional information by boundaries results in sharp discontinuities in the responses across two different regions (in agreement with the model). The development of abrupt changes in regional responses also results in a gradual sharpening of the boundary response, reflecting increased confidence in the precise location of the boundary.
- These findings are roughly consistent with neural network implementations of the weak membrane model. But other explanations are possible. For example, the weak membrane model requires lateral (sideways) interaction, and it is possible that the computations are done hierarchically using feedback from V2 to V1.

Relations to electrophysiology illustration

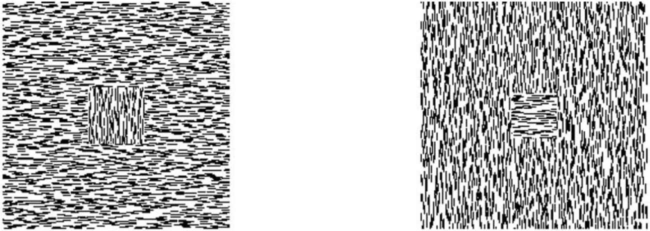

Figure 2: The stimuli for the experiments by TS Lee and his collaborators (Lee & Yuille, 2006).

Edge detection with spatial context (I)

- Our second example is to develop a model for detecting edges using spatial context. This relates to the phenomena known as association fields, see chapter figure 12.26 (left panel), where Gabor filters that are spatially aligned (in orientation and direction) get grouped into a coherent form.
- For this model, we have a set of neurons at every spatial position x, each tuned to a different angle $\theta_i : i = 1, ..., 8$, and a default cell at angle θ_0. The first cells are designed to detect edges at each orientation – i.e., they can be driven by the log-likelihood ratio of an edge detector at orientation θ_i at this position. The default cell is a dummy that is intended to fire if there is no edge present at this position. This organization forms a population of cells arrayed according to orientation (similar to a hypercolumn in V1).

Edge detection with spatial context (II)

We define a Gibbs distribution for the activity s_{x,θ_i} of the cells. The energy function $E(\vec{s})$ contains four types of terms: The first term, $\sum_x \sum_{i=0}^{8} s_{x,i}\phi(f_1, ..., f_M)$, represents the local evidence for an edge at each point and for its orientation. The second term $\sum_x (\sum_{i=0}^{8} s_{x,i} - 1)^2$., is intended to ensure that only one cell is active at any spatial position. This corresponds to an inhibitory interaction between cells in the same hypercolumn. The cells in the hypercolumn give alternative, and inconsistent, interpretations of the input – hence only one of them can be correct. The third term encourages edges to be continuous and change their directions smoothly. To define this term, we let $\vec{\theta_i} = (\cos\theta_i, \sin\theta_i)$ and $\vec{\theta_i}^T = (-\sin\theta_i, \cos\theta_i)$ denote the tangent to the edge and the normal. This term encourages there to be edges in the tangent direction, while the next term discourages them in the normal direction. This term is motivated by the intuition that curves are spatially smooth and can be justified by the statistics of natural images (Geisler & Perry, 2009; Elder & Goldberg, 2002).

Edge detection with spatial context (III)

We write it as $\sum_{x,y} \sum_{i,j=1}^{8} W^{T}_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j}$, where

$$W^{T}_{(x,\theta_i),(y,\theta_j)} = -\exp\{-|\vec{\theta}_i - \vec{\theta}_j|/K_1\} \exp\{-|x-y|/K_2\} \exp\{-|\hat{x}y - \vec{\theta}_i|/K_3\} \quad (5)$$

and $\hat{x}y$ is the unit vector in direction $x-y$. This term encourages edges that are in similar directions (first term) and nearby in position (second term), where the edge orientation is similar to the difference $x-y$ between the two points. This term is excitatory. The fourth and final term is inhibitory and discourages edges from being parallel to each other (if they are nearby). It is written as $\sum_{x,y} \sum_{i,j=1}^{8} W^{N}_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j}$. Here,

$$W^{N}_{(x,\theta_i),(y,\theta_j)} = \exp\{-|x-y|/K_4\} \exp\{-|\hat{x}y - \vec{\theta}_i^{T}|\} \quad (6)$$

Edge detection with spatial context (IV)

- The first term says this interaction decreases with distance. The second term discourages edges which are parallel to each other.
- This gives an overall energy:

$$E(\vec{s}) = \sum_{x} \sum_{i=0}^{8} s_{x,i} \phi(f_1, ..., f_M) + \hat{K}_0 \sum_{x} (\sum_{i=0}^{8} s_{x,i} - 1)^2$$
$$+ \hat{K}_1 \sum_{x,y} \sum_{i,j=1}^{8} W^{T}_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j} + \hat{K}_2 + \sum_{x,y} \sum_{i,j=1}^{8} W^{N}_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j}. \qquad (7)$$

- This yields a probability:

$$P(\vec{s}|\vec{f}) = \frac{1}{Z} \exp\{-E(\vec{s})\}.$$

- This model can be implemented in neural networks by defining either stochastic or deterministic neural dynamics (i.e., either Gibbs sampling or mean field theory). The resulting update equations are more complex than those defined for our earlier examples but have the same basic ingredients. Models of this type can qualitatively account for associative field phenomena.

Stereo models

This section introduces computational models for estimating depth by binocular stereo. The key problem to solve is the *correspondence problem* between the inputs in the two eyes to determine the *disparity*. Then the depth of the points in space can be estimated by trigonometry. (This presupposes that the eyes are *calibrated*, meaning that the distance between the eyes and the direction of gaze are known, which is beyond the scope of this chapter.) Julesz (1971) showed that humans could perceive depth from stereo if the images consisted of random dot stereograms, which minimize the effect of feature similarity cues, suggesting that human vision can solve this task by relying mainly on geometric regularities (assumed about the structure of the world). Other researchers (Bulthoff & Mallot, 1988) have studied human estimation of surface shape quantitatively and showed, among other things, bias toward fronto-parallel surfaces.

Stereo: The correspondence problem

Most stereo algorithm address the correspondence problem by assuming that (1) image features in the two eyes are more likely to correspond if they have similar appearance, and (2) the surface being viewed obeys prior knowledge, such as being piecewise smooth (e.g., like the weak membrane model). The first assumption depends on local properties of the images, while the second assumption uses nonlocal context. In an earlier lecture, we discussed how a population of Gabor filters could be used to match local image features. Here we describe how context can be used to impose prior knowledge about the geometry of the scene. We will study classic models, that assume that the surface is piecewise smooth. This leads to a Markov field model that includes excitatory connections, imposing the geometric constraints, with inhibitory connections that prevent points from one eye having more than one match in the second eye. This yields an algorithm that involves cooperation to implement the excitatory constraints, and competition to deal with the inhibitory constraints. This is consistent with findings from recent electrophysiological experiments (Samonds et al., 2009),(Samonds et al., 2012), which complement experiments (Ohzawa et al., 1990) that tested the local stereo models described earlier.

A cooperative stereo model (I)

- We now specify a computational model for stereo that for simplicity, we formulate in one dimension. There is a long history of this type of model, starting with the cooperative stereo algorithm (Dev, 1975; Marr & Poggio, 1976), and current computer vision stereo algorithms are mostly designed on similar principles.
- We specify the left and right images by $\vec{I}_L, \vec{I}_2$ and denote features extracted from them by $\vec{f}(\vec{I}_L) = \{f(x_L) : x_L \in \mathcal{D}_L\}$, $\vec{f}(\vec{I}_2) = \{f(x_2) : x_2 \in \mathcal{D}_2\}$. We define a discrete-valued correspondence variable $V(x_L, x_2)$ so that if $V(x_L, x_2) = 1$, the features at x_L, x_2 in the two images correspond, and hence the disparity is $x_L - x_2$. If the features do not match, then we set $V(x_L, x_2) = 0$. We encourage all data points to match one other data point, but allow some data points to be unmatched and others to match more than once (by paying a penalty).

A cooperative stereo model (II)

We specify a distribution $P(\vec{V}|\vec{f}(\vec{I_L}),\vec{f}(\vec{I_2})) = \frac{1}{Z}\exp\{-E(\vec{V};\vec{f}(\vec{I_L}),\vec{f}(\vec{I_2}))/T\}$, where the energy $E(\vec{V};\vec{f}(\vec{I_L}),\vec{f}(\vec{I_2}))$ is given by:

$$\begin{aligned} E(\vec{V};\vec{f}(\vec{I_L}),\vec{f}(\vec{I_2})) &= \sum_{x_L,x_2} V(x_L,x_2)M(f(x_L),f(x_2)) \\ +A\sum_{x_L}(\sum_{x_2} V(x_L,x_2)-1)^2 &+ A\sum_{x_2}(\sum_{x_L} V(x_L,x_2)-1)^2 \\ +C\sum_{x_L,x_2}\sum_{y_L\in N(x_L)}\sum_{y_2\in N(x_2)} & V(x_L,x_2)V(y_L,y_2)\{(x_2-x_L)-(y_2-y_L)\}^2. \end{aligned} \tag{8}$$

A cooperative stereo model (III)

The first term imposes matches between image points with similar features; here $M(.,.)$ is a measure that takes small values if $f(x_L), f(x_2)$ are similar and large values if they are different. We will discuss at the end of this section how $M(f(x_L), f(x_2))$ relates to the model for local stereo discussed earlier. The second two terms penalize image points that are either unmatched or matched more than once. The third term encourages the disparities, $x_L - x_2$, to be similar for neighboring points (here $N(.)$ defines a spatial neighborhood as before). These models can be applied to two-dimensional images by solving the correspondence problem for each epipolar line separately (by maximizing $P(\vec{V}|\vec{f}(\vec{I_L}), \vec{f}(\vec{I_2})))$. This is shown in the figure that follows. The parameter T is the variance of the model, as for the line process model, and has default value $T = 1$.

A cooperative stereo model illustration

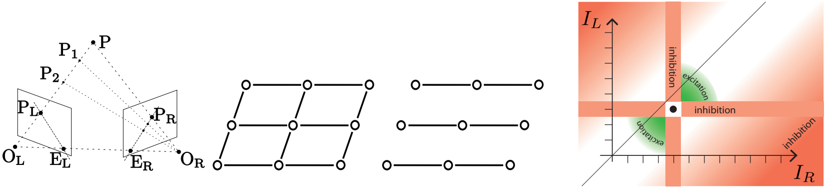

Figure 3: Far left and center: The geometry of stereo. A point P in 3-D space is projected onto points PL and PR. The projection is specified by the focal points OL, OR, and the directions of the cameras' gaze (the camera geometry). The geometry of stereo enforces that points in the plane specified by P, OR, OL must be projected onto corresponding lines EL, ER (the epipolar line constraint). If we can find the correspondence between the points on epipolar lines, then we can use trigonometry to estimate their depth, which is (roughly) inversely proportional to the disparity, which is the relative displacement of the two images. Far right: Binocular stereo requires solving the correspondence problem, which involves excitation (to encourage matches with similar depths/disparities) and inhibition (to prevent points from having multiple matches).

A cooperative stereo model (IV)

- We obtain a neural circuit model by performing mean field theory on $P(\vec{V}|\vec{f}(\vec{I_L}), \vec{f}(\vec{I_2}))$. This replaces $V(x_L, x_2) \in \{0,1\}$ by continuous-valued $q(x_L, x_2) \in [0,1]$ and an associated variable $u(x_L.x_2) = T \log \frac{q(x_L,x_2)}{1-q(x_L,x_2)}$ with $q(x_l, x_2) = \frac{1}{1+\exp\{-u(x_L,x_2)\}}$.
- The update equation is:

$$\frac{du(x_L, x_2)}{dt} = -u(x_L, x_2) - M(f(x_L), f(x_2))$$
$$-2A(\sum_{y_2 \neq x_2} q(x_L, y_2) - 1) - 2A(\sum_{y_L \neq x_L} q(y_L, x_2) - 1),$$
$$-2C \sum_{y_L \in N(x_L)} \sum_{y_2 \in N(x_2)} q(y_L, y_2)\{(x_2 - x_L) - (y_2 - y_L)\}^2. \tag{9}$$

- This update includes the standard integration term (first term), and the second term encourages matches where the features agree. There is also inhibition between competing matches (the third and fourth term), and excitation for matches that are consistent with a smooth surface (last term).

A cooperative stereo model: Interactive demo

There is a variant of this algorithm that is a discrete Hopfield network which attempts to minimize the energy $E(\vec{V}; \vec{f}(\vec{I_L}), \vec{f}(\vec{I_2}))$ in equation (8). The algorithm starts by assigning initial values, 0 or 1, to each state variable $V(x_L, x_2)$. The algorithm proceeds by selecting a state variable, changing its value (e.g., changing $V(x_L, x_2) = 1$ to $V(x_L, x_2) = 0$), calculating if this change reduces the energy $E(\vec{V}; \vec{f}(\vec{I_L}), \vec{f}(\vec{I_2}))$, and keeping the change if it does. This process repeats until the algorithm converges (i.e., all possible changes raise the value of the energy).

A cooperative stereo model and the local model

How does the cooperative stereo algorithm relate to our earlier algorithm for computing stereo disparity locally? Recall that the algorithm estimated the disparity at a single point by having a set of neurons tuned to different disparities $\{D_i : i = 1, ..., N\}$, summing the votes $v(D_i)$ for each disparity by equation (**??**), and selecting the disparity with the most votes. Using the cyclopean coordinate system (Jules, 1971), we express the disparity by $D(x) = \frac{1}{2}(x_2 - x_L)$, where $x = \frac{1}{2}(x_2 + x_L)$. At each point x we specify a population of neurons that encodes the votes $v(D(x))$ for the different disparities. Then, instead of using winner-take-all to make a local decision, we feed the responses $v(D(x))$ back into cooperative stereo algorithm by defining $M(f(x_L), f(x_2)) = \exp\{-v(\frac{1}{2}(x_2 - x_L))\}$ (the negative exponential $\exp\{-\}$ is required so the $M(f(x_L), f(x_2))$ is small if the vote for disparity $D(x) = \frac{1}{2}(x_2 - x_L)$ is large).

A cooperative stereo model and electrophysiology

Analyses of electrophysiological studies (Samonds et al., 2009),(Samonds et al., 2012) were in general agreement with the predictions of this type of stereo algorithm. In particular, studies showed that neural population responses included excitation between cells tuned to similar disparities at neighboring spatial positions as well as inhibition between cells tuned to different disparities at the same position. In addition, Samonds et al. (2013) implemented a variant of the stereo algorithm described above and showed that it could account for additional phenomena, such as sharper tuning to the disparity for larger stimuli and performance on anticorrelated stimuli (where the left and right images have opposite polarity).

A cooperative stereo model and electrophysiology illustration

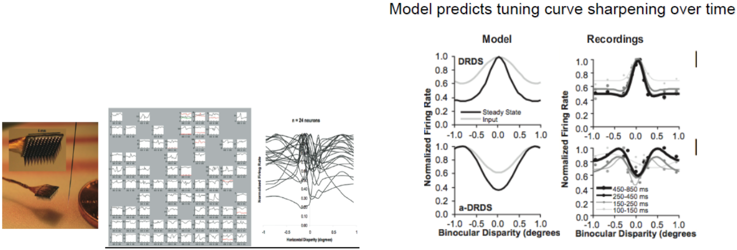

Figure 4: Experiments for testing stereo algorithms (Samonds et al., 2009, 2012). Left: The experimental setup. Right: The experiments give evidence for excitation between similar disparity and inhibition to prevent multiple matches.

Motion

- Similar models have been applied to a range of motion phenomena. The input is a sequence of images taken at subsequent times. The task is to estimate the correspondence between pixels.
- In short-rang motion, the images are taken at thirty frames per second (or faster). In long-range motion, there are larger time gaps between the images. Short-range motion is differentiable (after some smoothing) but long-range motion is not. Short-range motion suffers from the *aperture problem* where only one component of the motion/velocity can be directly estimated. Long-range motion has a correspondence problem (without the epipolar line constraint unless the scene is rigid).
- Early computational studies (Ullman, 1979) showed that several perceptual phenomena of long-range motion could be described by a "minimal mapping" theory that uses a slowness prior. Smoothness priors accounted for findings on short-range motion (Hildreth, 1984).
- Yuille and Grzywacz (1988) qualitatively showed that a slow-and-smooth prior could account for a large range of motion perceptual phenomena – including motion capture and motion cooperation – for short- and long-range motion. Weiss and his collaborators showed that slow (Weiss & Adelson, 1998) and slow-and-smooth priors (Weiss et al., 2002) could explain other short-range motion phenomena, such as how percepts can change dramatically as we alter the balance between the likelihood and prior terms (i.e., for some stimuli the prior dominates the likelihood and vice versa).

Motion Phenomena

- All these models combine local estimates of the motion, such as those described in the previous section, with contextual cues implementing slow-and-smooth priors. They can be formulated using the same mathematical techniques.
- There are a range of other phenomena – motion transparency and depth estimation – which require other types of models.
- See `http://www.michaelbach.de/ot/mot-motionBinding/` to see how spatial context can be affected by other cues such as occlusion. It is also possible to perceive three-dimensional structure by observing a motion sequence (somewhat similar to binocular stereo) as can be seen in `http://michaelbach.de/ot/mot-ske/`.

Motion: Short Range Slow-and-Smooth

- We present a simple slow-and-smooth model.
- The model is formulated as estimating the two dimensional velocities $(U, V) = \{(U_i, V_i) : \; i \in \Lambda\}$ defined over an image lattice Λ. Our goal is to estimate the motion, or velocity, (U, V). Smoothness is defined over a local neighborhood $Nbh(i)$ defined on the lattice,
- The likelihood functions and the slow-and-smoothness prior are defined by Gibbs distributions:

$$P(D|U, V) = \frac{1}{Z} \exp\{-E[D; U, V]\},$$
$$P(U, V) = \frac{1}{Z} \exp\{-E(U, V)\}. \quad (10)$$

-

$$E[D; U, V] = \sum_{i \in \Lambda} \gamma_i (U_i sin\theta_i + V_i \cos\theta_i - D_i)^2$$
$$E(U, V) = \alpha \sum_{i \in \Lambda} \{U_i^2 + V_i^2\} + \beta \sum_{i \in \Lambda} \sum_{j \in Nbh(i)} \{(U_i - U_j)^2 + (V_i - V_j)^2\}. \quad (11)$$

- The *data term* assumes that we can only observe one component of the velocity specified by a known angle θ_i. The parameter $\gamma_i = 0$ if there are no observations at lattice site i, and otherwise $\gamma_i = 1/(2\sigma_i^2)$ where σ_i^2 is the variance of the data at i. *The prior terms* imposes both slowness and smoothness terms – weighted by α and β respectively.

Motion: Slow-and-Smooth

- The posterior distribution $P(U, V|D) \propto P(D|U, V)P(U, V)$ is a Gaussian. This is because both $P(D|U, V)$ and $P(U, V)$ are Gaussians (and the conjugate of a Gaussian is also a Gaussian).
- We estimate the most probable motion $(\hat{U}, \hat{V})$ from $P(U, V|D)$. For Gaussian distributions, the MAP estimate and the mean estimate are identical. Both reduce to minimizing the energy function $E(U, V) + E(D; U, V)$ which is quadratic in (U, V). This is performed by solving the linear equations:
-

$$0 = \alpha \hat{U}_i + \beta \sum_{j \in Nbh(i)} (\hat{U}_i - \hat{U}_j) - \gamma_i \{D_i - \sin\theta_i \hat{U}_i - \cos\theta_i (\hat{V}_i)\} \sin\theta_i, \ \forall i \in \Lambda$$

$$0 = \alpha \hat{V}_i + \beta \sum_{j \in Nbh(i)} (\hat{V}_i - \hat{V}_j) - \gamma_i \{D_i - \sin\theta_i \hat{U}_i - \cos\theta_i (\hat{V}_i)\} \cos\theta_i, \ \forall i \in \Lambda. \quad (12)$$

Motion: Slow-and-Smooth Examples

- First, at a position where there is no observation and so $\gamma_i = 0$. The estimated velocity at i is a sub-average of the velocities of its neighbors:

$$\hat{U}_i = \frac{\beta \sum_{j \in Nbh(i)} \hat{U}_j}{\alpha + |Nbh|\beta}, \;\; \hat{V}_i = \frac{\beta \sum_{j \in Nbh(i)} \hat{V}_j}{\alpha + |Nbh|\beta}. \tag{13}$$

- If there is no slowness (i.e. $\alpha = 0$) then the velocity estimate $(\hat{U}_i, \hat{U}_j)$ is an average of the velocity of its neighbors. But if $\alpha > 0$ then the estimates are lower, meaning that the estimate of motion speed decreases in regions where there are no observations (agrees with experiments). If there is no smoothness (i.e. $\beta = 0$) then the estimate of velocity is zero at i.
- Second, at a lattice node with an observation the model encourages similarity to the motion of the neighbors and agreement with the observations.
-

$$\hat{U}_i = \frac{\beta \sum_{j \in Nbh(i)} \hat{U}_j + \gamma_i D_i \sin\theta_i}{\alpha + \beta|Nbh| + \gamma_i \sin^2\theta_i}, \quad \hat{V}_i = \frac{\beta \sum_{j \in Nbh(i)} \hat{V}_j + \gamma_i D_i \cos\theta_i}{\alpha + \beta|Nbh| + \gamma_i \cos^2\theta_i}.$$

- A special case occurs when we set $\beta = 0$ which removes the smoothness constraint yielding

$$\hat{U}_i = \frac{\gamma_i D_i \sin\theta_i}{\alpha + \gamma_i \sin^2\theta_i}, \;\; \hat{V}_i = \frac{\gamma_i D_i \cos\theta_i}{\alpha + \gamma_i \cos^2\theta_i}. \tag{14}$$

This encourages the estimated motion to be in direction $(\sin\theta_i, \cos\theta_i)$.

Motion: Slow-and-Smooth Gaussians

- A more advanced model (Yuille and Grzywacz 1988) imposes a slow-and-smooth prior which includes higher-order derivatives on the velocity field.
- In this theory, the velocity estimates can be expressed as linear weighted sums of Gaussian distributions centered on the observations. This predicts how the velocity falls off with spatial distances.
- This theory helped inspire Poggio's theory of learning by radial basis functions.
- The theory is also used for the related problem of shape matching.

Motion: Long-Range Motion

- In long-range motion there is a large time difference between time franes. This means that we have a correspondence problem and not an aperture problem. Ullman formulated this a minimal mapping problem. (1979). His theory essentially assumed that the velocity was as slow as possible. Experiments showed that human perception was more consistent with slow-and-smooth. This type of theory will be discussed in a few slides.
- First, we discuss an *ideal observer* study of long-range motion perception (Barlow and Tripathy 1997). This addressed the ability of humans to perceive coherent long-range motion in the presence of background clutter.
- This model is interesting because it compares human ability to perform this visual task with an ideal observer model which knows the statistical properties of the stimuli. Not surprising the ideal observer model does better (by many orders of magnitude). Human perception is much more consistent with a slow-and-smooth model (Lu and Yuille 2006).

Motion: Long Range Motion: Ideal Observer

- There are N points in the first time frame at positions $\{x_i : i = 1, ..., N\}$. A proportion of these CN move coherently by an amount $v + \delta$ between each time frame where v is a constant (fixed translation) and $\delta \sim \mathcal{N}(0, \sigma)$ is zero mean additive Gaussian noise. The remaining $(1 - C)N$ points move at random.
- To model this we introduce a set of binary-values variables $\{V_i \in \{0, 1\} : i = 1, ..., N\}$ so that if $V_i = 1$ then dot x_i moves coherently – i.e. $P(y_i|x_i, v, V_i = 1) = P(y_i|x_i, v) = \mathcal{N}(x_i + v, \sigma)$ – while if $V_i = 0$ then $P(y_i|x_i, V_i = 0) = U(y_i)$, where $U(.)$ is the uniform distribution. This is a *mixture model*:

$$P(y_i|x_i, v, V_i) = P(y_i|x_i, v)^{V_i} U(y_i)^{1-V_i}. \tag{15}$$

We impose a prior on the $\{V_i : i = 1, .., N\}$ which ensures that CN dots move coherently – so $\sum_{i=1}^{N} V_i = CN$ – and a prior $P(v)$ on the velocity.

Motion: Long Range Motion Ideal

- This gives a model:

$$P(\{y_i\}|\{x_i\},\{V_i\},v) = \prod_{i=1}^{N} P(y_i|x_i,v)^{V_i} U(y_i)^{1-V_i},$$

$$P(\{V_i : i = 1,..,N\}) = \delta\{\sum_{i=1}^{N} V_i - CN\}, \quad P(v). \tag{16}$$

- The experiments by Barlow and Tripathy (1997) require human subjects to estimate the velocity v for the stimuli. This is sometimes constrained so that v can either move to the left or the right by a fixed amount t – e.g., $v \in \{\pm t\}$ for fixed t. We can model this by requiring that $P(v) = (1/2)\delta(v-t) + (1/2)\delta(v+t)$.
- We can compare human performance on estimating velocity – e.g., false positives and false negatives – to the model prediction obtained from:

$$P(v|\{y_i\},\{x_i\}) = \frac{\sum_{\{V_i\}} P(\{y_i\}|\{x_i\},\{V_i\},v)P(\{V_i\})P(v)}{\sum_v \sum_{\{V_i\},t} P(\{y_i\}|\{x_i\},\{V_i\},v)P(\{V_i\})P(v)}. \tag{17}$$

- This computation is demanding since it requires summing over all possible $\{V_i\}$. There are $N!/(NC)!(N(1-C))!$ possible values.

Motion: Long Range Motion Ideal

- In fact, the computation is even worse because our formulation has assumed that we know the correspondence between dots in the first and second frame. To model this ambiguity, we need to replace the $\{V_i\}$ by correspondence variables $\{V_{ia}\}$ where each $V_{ia}\{0,1\}$ take only binary-values. This correspondence variable must obey the following constraints which we impose in the prior $P(\{V_{ia}\})$.
- Firstly, we set $V_{ia} = 1$ if x_i in the first frame corresponds to y_a in the second frame.
- Secondly, to avoid matching ambiguity we require that if $V_{ia} = 1$ then $V_{ib} = 0$ for all $b \neq a$ – i.e. a dot x_i can have at most one match y_a in the second frame. Thirdly, we impose the constraint $\sum_{i=1,a=1}^{N,N} V_{ia} = CN$ to ensure that a fraction CN of dots are matched.
- Finally, we replace the term $P(\{y_i\}|\{x_i\},\{V_i\},v)$ by

$$P(\{y_a\}|\{x_i\},\{V_{ia}\},v) = \prod_{i=1,a=1}^{N,N} P(y_a|x_i,v)^{V_{ia}} U(y_i)^{1-V_{ia}}. \qquad (18)$$

Then we modify our derivation of equation (17) to get:

$$P(v|\{y_a\},\{x_i\}) = \frac{\sum_{\{V_{ia}\}} P(\{y_a\}|\{x_i\},\{V_{ia}\},v)P(\{V_{ia}\})P(v)}{\sum_v \sum_{\{V_{ia}\},t} P(\{y_a\}|\{x_i\},\{V_{ia}\},v)P(\{V_{ia}\})P(v)}. \qquad (19)$$

Motion: Long Range Motion Ideal

- The EM algorithm enables us to estimate $v^* = \arg\max P(v|\{y_a\}, \{x_i\})$ well in practice. This algorithm iterates between estimating the velocity v (or t if we allow only two velocities) then estimating a distribution $Q(\{V_{ia}\})$ for the correspondence variables.
- Lu and Yuille (2005) computed the Bayes risk for this model precisely (Barlow and Tripathy had made approximate estimates of it).
- Their analysis showed that human observers were many orders of magnitude worse than the performance predicted by the model. Even assuming that human observers had degraded models – e.g., wrong priors for $P(v)$, noise in their measurements of $\{x_i\}$ and $\{y_i\}$ – were enable to account for the difference. Nevertheless this model did predict the trends of the data, for example how performance changed as number N of dots varied, as C varied, and as t varied.
- Lu and Yuille suggested that the enormous difference between human and model performance arose because humans used a general purpose model of motion perception suited to the statistics of the visual stimuli that occur in the real world and not those that appear in laboratory experiments.

Motion: Long Range Motion Ideal

- An alternative model for motion estimation which assumed that the motion $\{v(x)\}$ can vary spatially but obeying a slow-and-smooth prior $P(\{v(x)\})$ (see earlier chapter). The correspondence prior $P(\{V_{ia}\})$ is modified to require that all dots are matched $\sum_{ia} V_{ia} = N$.
- The prediction equation is modified to be:

$$P(\{y_i\}|\{x_i\},\{V_{ia}\},\{v(x_i)\}) = \prod_{i=1,a=1}^{N,N} P(y_a|x_i + v(x_i))^{V_{ia}}. \qquad (20)$$

- The velocity can then be estimated by solving $v(x)^* = \arg\max P(\{v(x)\}|\{x_i\},\{y_a\})$ where $P(\{v(x)\}|\{x_i\},\{y_a\})$ is given by:

$$\frac{\sum_{\{V_{ia}\}} P(\{y_a\}|\{x_i\},\{V_{ia}\},\{v(x)\})P(\{V_{ia}\})P(\{v(x)\})}{\sum_{\{v(x)\}}\sum_{\{V_{ia}\},t} P(\{y_a\}|\{x_i\},\{V_{ia}\},\{v(x)\})P(\{V_{ia}\})P(\{v(x)\})}. \qquad (21)$$

- The solution for $v(x)^*$ can also be found by applying the EM algorithm (Lu and Yuille 2005). It can be shown that this model gave very good fits to human performance on the data described by Barlow and Tripathy and also on novel experiments.
- This suggests that human performance, at least for visual perception, may be based on models and prior assumptions which are valid in the natural environment. Humans may not be unable to adapt to the statistics chosen, somewhat arbitrarily, by the experimenter in a laboratory setting.

Motion: Long Range Motion Transparency

- We can also modify the model above to deal with transparent motion where there are two types of motion occurring simultaneously. The simplest case involves motion moving either to the left with average velocity t or to the right with average velocity $-t$.
- We modify the to be:

  $$P(y_i|x_i, t, V_i) = P(y_i|x_i, t)^{V_i} P(y_i|x_i, -t)^{1-V_i}. \tag{22}$$

  From this we can estimate the probability of t and of the $\{V_i\}$ enabling us to deal with transparent motion and estimate the velocities $\pm t$ and which dots move to the left $V_i = 0$ and which to the right $V_i = 1$.
- This transparency motion model is called a layered model since it divides the data into two-layers, with $V_i = 0$ or $V_i = 0$. The model can be extended to allowing that the velocities are allowed to vary within each layers – i.e. replace v by $\{v(x)\}$ – and by using correspondence variables.
- These transparency motion models are shown to perform well on real world motion stimuli and also to qualitatively account for human performance on such stimuli (Weiss 1997).

Summary of models with context

This section illustrated how neural networks and Markov models could be used to apply context to visual tasks. We concentrated on edge detection, segmentation, and binocular stereo. We stressed how context can include excitatory and inhibitory interactions. And how inference can be performed using stochastic neurons (e.g., Gibbs sampling) or dynamic neural networks (e.g., mean field approximations). These models have some relations to psychophysics and electrophysiology. But we stress that detailed biological evidence in favor of these models remains preliminary due to the current limitations of experimental techniques. We note that current computer vision algorithms that address similar visual tasks are more complex although based on similar principles (Blake et al., 2011).