

Probabilities on Graphs: Undirected Graphs and Spatial Context

Alan Yuille and Dan Kersten

May 6, 2015

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. (Assuming the cameras are calibrated, 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 no scene structure. This shows that humans can solve stereo without needing to first recognize objects.

g

Stereo: The correspondence problem

- ▶ Stereo algorithms 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 weak assumptions about the spatial context such as being piecewise smooth (like the weak membrane model).
- ▶ A previous lecture developed a stereo algorithm using local image features alone (Gabor filters). But these algorithms are not very effective (unless hierarchical deep network filters are used).
- ▶ In this lecture we describe how spatial context, or prior knowledge, can be exploited and, in particular, piecewise smoothness. We describe an early stereo algorithm (Marr & Poggio 1976, Dev 1975) that captures the key concepts.

Technically this involves 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),

A cooperative stereo model (I)

- ▶ We specify the left and right images by \vec{l}_L, \vec{l}_R and denote features extracted from them by $\vec{f}(\vec{l}_L) = \{f(x_L) : x_L \in \mathcal{D}_L\}$, $\vec{f}(\vec{l}_R) = \{f(x_R) : x_R \in \mathcal{D}_R\}$.
- ▶ We define a discrete-valued correspondence variable $V(x_L, x_R)$ so that if $V(x_L, x_R) = 1$, the features at x_L, x_R in the two images correspond, and hence the disparity is $x_L - x_R$. If the features do not match, then we set $V(x_L, x_R) = 0$.
- ▶ 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).
- ▶ Similar algorithms can be applied to other matching problems in vision and elsewhere, This approach can even be applied to CS problems like the Travelling Salesman Problem.

A cooperative stereo model (II)

We specify a distribution $P(\vec{V}|\vec{f}(\vec{l}_L), \vec{f}(\vec{l}_2)) = \frac{1}{Z} \exp\{-E(\vec{V}; \vec{f}(\vec{l}_L), \vec{f}(\vec{l}_2))/T\}$, where the energy $E(\vec{V}; \vec{f}(\vec{l}_L), \vec{f}(\vec{l}_2))$ is given by:

$$\begin{aligned} E(\vec{V}; \vec{f}(\vec{l}_L), \vec{f}(\vec{l}_2)) &= \sum_{x_L, x_2} V(x_L, x_2) M(f(x_L), f(x_2)) \\ &+ A \sum_{x_L} \left(\sum_{x_2} V(x_L, x_2) - 1 \right)^2 + A \sum_{x_2} \left(\sum_{x_L} V(x_L, x_2) - 1 \right)^2 \\ &+ C \sum_{x_L, x_2} \sum_{y_L \in N(x_L)} \sum_{y_2 \in N(x_2)} V(x_L, x_2) V(y_L, y_2) \{(x_2 - x_L) - (y_2 - y_L)\}^2. \end{aligned} \quad (1)$$

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 in earlier lecture).
- ▶ 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 $P(\vec{V} | \vec{f}(\vec{l}_L), \vec{f}(\vec{l}_2)))$).
- ▶ 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

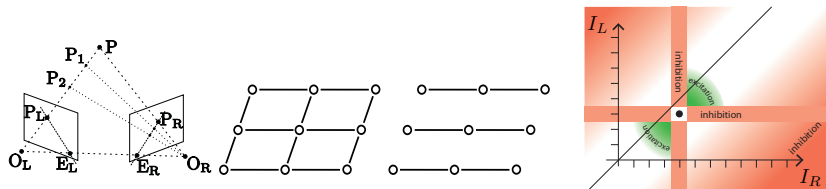


Figure 1: Far left and center: The geometry of stereo. A point P in 3-D space is projected onto points P_L and P_R . The projection is specified by the focal points O_L , O_R , and the directions of the cameras' gaze (the camera geometry). The geometry of stereo enforces that points in the plane specified by P , O_R , O_L 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 $P(\vec{V}|\vec{f}(\vec{l}_1), \vec{f}(\vec{l}_2))$. 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:

$$\begin{aligned} \frac{du(x_L, x_2)}{dt} = & -u(x_L, x_2) - M(f(x_L), f(x_2)) \\ & - 2A \left(\sum_{y_2 \neq x_2} q(x_L, y_2) - 1 \right) - 2A \left(\sum_{y_L \neq x_L} q(y_L, x_2) - 1 \right), \\ & - 2C \sum_{y_L \in N(x_L)} \sum_{y_2 \in N(x_2)} q(y_L, y_2) \{(x_2 - x_L) - (y_2 - y_L)\}^2. \end{aligned} \quad (2)$$

- ▶ 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 network which attempts to minimize the energy $E(\vec{V}; \vec{f}(\vec{l}_1), \vec{f}(\vec{l}_2))$ in equation (1).
- ▶ 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 $E(\vec{V}; \vec{f}(\vec{l}_1), \vec{f}(\vec{l}_2))$, 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 $\{D_i : i = 1, \dots, N\}$, summing the votes $v(D_i)$ for each disparity, 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

Model predicts tuning curve sharpening over time

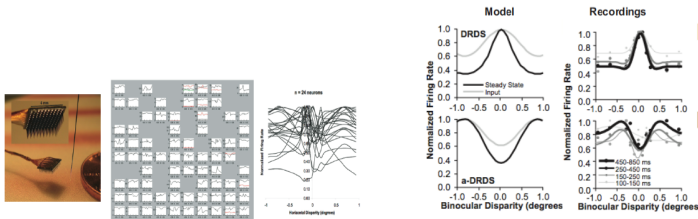


Figure 2: Experiments for testing stereo algorithms (Samonds et al., 2009, 2012). Left: The experimental setup. Right: The experiments give evidence for excitation between similar disparity and inhibition to prevent multiple matches.

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-range 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:

$$P(D|U, V) = \frac{1}{Z} \exp\{-E[D; U, V]\},$$
$$P(U, V) = \frac{1}{Z} \exp\{-E(U, V)\}. \quad (3)$$



$$E[D; U, V] = \sum_{i \in \Lambda} \gamma_i (U_i \sin \theta_i + V_i \cos \theta_i - D_i)^2$$
$$E(U, V) = \alpha \sum_{i \in \Lambda} \{U_i^2 + V_i^2\} + \beta \sum_{i \in \Lambda} \sum_{j \in Nbh(i)} \{(U_i - U_j)^2 + (V_i - V_j)^2\}. \quad (4)$$

- ▶ 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:

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

$$0 = \alpha \hat{V}_i + \beta \sum_{j \in Nbh(i)} (\hat{V}_i - \hat{V}_j) - \gamma_i \{D_i - \sin \theta_i \hat{U}_i - \cos \theta_i (\hat{V}_i)\} \cos \theta_i, \quad \forall i \in \Lambda. \quad (5)$$

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:

$$\hat{U}_i = \frac{\beta \sum_{j \in Nbh(i)} \hat{U}_j}{\alpha + |Nbh|\beta}, \quad \hat{V}_i = \frac{\beta \sum_{j \in Nbh(i)} \hat{V}_j}{\alpha + |Nbh|\beta}. \quad (6)$$

- If there is no slowness (i.e. $\alpha = 0$) then the velocity estimate (\hat{U}_i, \hat{V}_i) 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

$$\hat{U}_i = \frac{\gamma_i D_i \sin \theta_i}{\alpha + \gamma_i \sin^2 \theta_i}, \quad \hat{V}_i = \frac{\gamma_i D_i \cos \theta_i}{\alpha + \gamma_i \cos^2 \theta_i}. \quad (7)$$

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 frames. 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 $\{x_i : i = 1, \dots, N\}$. 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 $\{V_i \in \{0, 1\} : i = 1, \dots, N\}$ 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 $U(\cdot)$ is the uniform distribution. This is a *mixture model*:

$$P(y_i|x_i, v, V_i) = P(y_i|x_i, v)^{V_i} U(y_i)^{1-V_i}. \quad (8)$$

We impose a prior on the $\{V_i : i = 1, \dots, N\}$ 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:

$$P(\{y_i\}|\{x_i\}, \{V_i\}, v) = \prod_{i=1}^N P(y_i|x_i, v)^{V_i} U(y_i)^{1-V_i},$$
$$P(\{V_i : i = 1, \dots, N\}) = \delta\left\{\sum_{i=1}^N V_i - CN\right\}, \quad P(v). \quad (9)$$

- ▶ 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:

$$P(v|\{y_i\}, \{x_i\}) = \frac{\sum_{\{V_i\}} P(\{y_i\}|\{x_i\}, \{V_i\}, v) P(\{V_i\}) P(v)}{\sum_v \sum_{\{V_i\}, t} P(\{y_i\}|\{x_i\}, \{V_i\}, v) P(\{V_i\}) P(v)}. \quad (10)$$

- ▶ 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 $V_{ia} \in \{0, 1\}$ 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 $P(\{y_i\}|\{x_i\}, \{V_i\}, \nu)$ by

$$P(\{y_a\}|\{x_i\}, \{V_{ia}\}, \nu) = \prod_{i=1, a=1}^{N, N} P(y_a|x_i, \nu)^{V_{ia}} U(y_i)^{1-V_{ia}}. \quad (11)$$

Then we modify our derivation of equation (10) to get:

$$P(\nu|\{y_a\}, \{x_i\}) = \frac{\sum_{\{V_{ia}\}} P(\{y_a\}|\{x_i\}, \{V_{ia}\}, \nu) P(\{V_{ia}\}) P(\nu)}{\sum_{\nu} \sum_{\{V_{ia}\}, t} P(\{y_a\}|\{x_i\}, \{V_{ia}\}, \nu) P(\{V_{ia}\}) P(\nu)}. \quad (12)$$

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:

$$P(\{y_i\}|\{x_i\}, \{V_{ia}\}, \{v(x_i)\}) = \prod_{i=1, a=1}^{N, N} P(y_a|x_i + v(x_i))^{V_{ia}}. \quad (13)$$

- ▶ 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:

$$\frac{\sum_{\{V_{ia}\}} P(\{y_a\}|\{x_i\}, \{V_{ia}\}, \{v(x)\}) P(\{V_{ia}\}) P(\{v(x)\})}{\sum_{\{v(x)\}} \sum_{\{V_{ia}\}, t} P(\{y_a\}|\{x_i\}, \{V_{ia}\}, \{v(x)\}) P(\{V_{ia}\}) P(\{v(x)\})}. \quad (14)$$

- ▶ 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:

$$P(y_i|x_i, t, V_i) = P(y_i|x_i, t)^{V_i} P(y_i|x_i, -t)^{1-V_i}. \quad (15)$$

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 $V_i = 1$. 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 lecture described models for binocular stereo and motion. We stressed how context can include excitatory and inhibitory interactions.
- ▶ 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.
- ▶ Most recent work on these topics use neural networks. There were delays because of the difficulties of annotating stereo and motion. Synthetic data was used, but this has only recently become sufficiently realistic to enable transfer to real world images. CNNs are good for extracting features which could be used for matching. But the self-attention mechanisms in transformers make them more suitable for these correspondence tasks.