Binocular Stereo

A.L. Yuille and D. Kersten

Abstract

A. Binocular Stereo

We now consider a model for the task of binocular stereo. This requires determining the assignment between pixels i and a on the left and right images. Knowing this assignment enables us to estimate depth by trigonometry, see figure (1). The assignment problem can be simplified by imposing the epipolar line constraint which follows from the geometry of image formation. If the viewing directions of the two eyes are known, then points in one eye can only be matched to points on the other eye which lie on the corresponding epipolar line. If both eyes look directly forward then the epipolar lines are parallel.

Similarly, binocular stereo estimates depth of surfaces by solving the correspondence problem to match image features in the left and right images. Local matching, however, is often ambiguous and so additional requirements are needed such as assuming that the viewed surface is piecewise smooth. Finally, models of binocular stereo [1] (Arbib and Dev) emphasize that matching requires both excitation and inhibition, which is consistent with multi-electrode recordings [2],[3],[4].

Researchers in computer vision have discovered the advantages of having different stereo algorithms suitable for different types of images – for example, algorithms that can exploit the strong Manhattan-world structure that appears in city scenes (consisting of long edges aligned to a two-dimensional grid on the groundplane and a vertical direction).

This problem can be formulated in terms if correspondence variable $x_{ia} \in \{0,1\}$ defined for each pair of points i, a on the left and right image lattices. We require that a point in one eye can only be matched to one point in the other, which means that matches can *inhibit* each other. But we also assume that surfaces tends to be spatially smooth which means that neighboring points in one eye should be matched to neighboring points in the other, which gives *excitation.* The assumption of weak surface smoothness is required to ensure that the matching between the two eyes is unambiguous.

B. Local Stereo Cues: the Disparity Energy Model

The disparity energy model is formulated using Gabor filters and has some claim to biological plausibility. We follow the presentation in Qian (Neural Computation 94). It is based on models by Q... and Freeman. It is related to the receptive field properties of cells by ****. It assumes that we have a large set of cells, receiving input from both images, and which are tuned to different image frequencies and spatial phases. The disparity of the image can be computed from the response of these filters.

We give the presentation in one-dimension for simplicity. This is allowed because of the epipolar line constraint (illustrate this!!). It assumes that the cell receives input from both left and right eyes. It has receptive fields $f_l(x) = \exp\{-x^2/(2\sigma^2)\}\cos(\omega x + \phi_l)$ and $f_r(x) = \exp\{-x^2/(2\sigma^2)\}\cos(\omega x + \phi_r)$. In other words, they are Gabors where the Gaussian has variance σ^2, tuned to frequency ω and with phases ϕ_l, ϕ_r. The linear response is:

$$r = \int dx\{f_l(x)I_l(x) + f_r(x)I_r(x)\}. \tag{1}$$

This filter is be tuned to spatial frequency ω. Recall, from earlier lectures, that we can express the image by a Fourier expansion. The filter is most sensitive to the image component at this frequency. Hence we can represent the image as $I(\vec{x}) = \rho\cos(\omega x + \theta)$.

Now suppose that the right image is a displaced version of the left image $I_r(x) = I_l(x + D)$, where D is the disparity (define!!). Then we can ignore the Gaussian to calculate r. This give a good approximation if we allow the disparity to change smoothly with x – i.e. $D(x)$ – provided it changes slowly over the size of the Gaussian 2σ. This gives:

$$r_1 = \rho\{\cos(\theta - \phi_l) + \cos(\theta - \phi_r - \omega D)\}. \tag{2}$$

This can be re-expressed as:

$$r_1 = 2\rho\cos(\theta - \frac{\phi_l + \phi_r}{2} - \frac{\omega D}{2})\cos(\frac{\phi_l - \phi_r}{2} - \omega\frac{D}{2}). \tag{3}$$

Hence the response of the cell depends on the disparity. In particular, when $\phi_l - \phi_r = \omega D$ then the second cosine takes its biggest value of 1. But the cell also depends on image properties, i.e., the image phase θ which is an argument of the first cosine. This makes it unsuited for detecting disparity. But it can still be used if we include it with a cell which has a quadrature pair of filters (refer back to section!!).

Now suppose that we consider quadrature pairs of the two cells tuned to the same ω. Where one cell has phases ϕ_l, ϕ_r and the other has phases ϕ'_l, ϕ'_r, where $(\phi_l - \phi_r) = (\phi'_l - \phi'_r)$ and $\phi'_l + \phi'_r = \phi_l + \phi_r + \frac{\pi}{2}$. Then the second cell has response $r_2 = 2\rho\cos(\theta - \frac{\phi_l+\phi_r}{2} - \frac{\omega D}{2})\cos(\frac{\phi_l-\phi_r}{2} - \omega\frac{D}{2}) = 2\rho\sin(\theta - \frac{\phi_l+\phi_r}{2} - \frac{\omega D}{2})\cos(\frac{\phi_l-\phi_r}{2}$. Hence if we squares and add the responses of the two cells we obtain:

$$r_1^2 + r_2^2 = \cos^2(\frac{\phi_l - \phi_r}{2} - \omega\frac{D}{2}). \tag{4}$$

This response depends only on the disparity D and the image phase ω. It takes largest values when $\phi_l - \phi_r = \omega D$. Hence a population of quadrature cells tuned to different phases ϕ_l, ϕ_r and frequencies ω. From this population of cells we can estimate D from the cells with biggest response.

Note that this method effectively consists of matching image windows in the left and right image to each other (as done in computer vision).

Technical notes. These results are derived using the two following formulae:

$$\int dx \cos(\omega x + \phi_l)\cos(\omega x + \theta) = \frac{1}{2}\cos(\theta - \phi_l). \tag{5}$$

This can be obtained by expressing $\cos\theta = \frac{1}{2}\{e^{i\theta} + e^{-i\omega}\}$. Then the integrand can be expressed in terms of exponentials which are easy to integrate. The second formula is

$$\cos 2\theta + \cos 2\alpha = 2\cos(\theta + \alpha)\cos(\theta - \alpha). \tag{6}$$

This also can be derived by expressing the cosines in term of exponentials.

C. *Models with Piecewise smooth surfaces*

We now specify one of the first class of stereo algorithm. The so-called cooperative stereo algorithm [1] (Arbib and Dev). Reformulating this in terms of probabilities leads to an distribution of form:

$$\begin{aligned} P(\mathbf{x}|\mathbf{z}_L, \mathbf{z}_R) &= \frac{1}{Z}\exp\{-E(\mathbf{x}; \mathbf{z}_L, \mathbf{z}_R)\}, \\ \text{where } E(\mathbf{x}) &= \sum_{ia} x_{ia} M(z_L^i, z_R^a) + A\sum_i(\sum_a x_{ia} - 1)^2 + B\sum_a(\sum_i x_{ia} - 1)^2 \\ &+ C\sum_{ia}\sum_{j\in Nbd(i), b\in Nbh(a)} x_{ia}x_{jb}F(\{p_1^i - p_1^a\} - \{p_1^j - p_1^b\})G(p_2^i - p_2^a)G(p_2^j - p_2^b). \end{aligned} \tag{7}$$

Here $M(z_L^i, z_R^a)$ is a measure of similarity between images features in the left and right eyes and hence encourages matches between lattice sites which have similar image features. $\vec{p}^1 = (p_1^i, p_2^i)$ denotes the position of lattice point i. The function $G(.)$ is strongly peaked at zero, imposing the condition that matches only happen between points with the same values $p_2^1 = p_2^a$ (i.e. the epipolar line constraint). The disparity $p_1^i - p_1^a$ between i and a is a measure of the depth (if the two points are matched). The function $F(.)$ is weakly peaked about 0 and encourages neighboring points to have similar disparities, hence corresponding to a smooth underlying surface. The size of the neighborhood is specified by $Nbh()$ and typically only includes nearest neighbors in the lattice.

The cooperative stereo algorithm performed (discrete) steepest descent on the energy $E(\mathbf{z}, \mathbf{z}_L, \mathbf{z}_R)$. More recent algorithms have more complex energy functions and use algorithms such as dynamic programming or belief propagation which are better at estimating the most probable match $\mathbf{z}^* = \arg\max P(\mathbf{x}|\mathbf{z}_L, \mathbf{z}_R)$.

Note that this probabilistic formulation will inhibit matches in one direction, by requiring that each point has at most one match, but will excite matches which have similar disparities and hence correspond to smooth surfaces.

Experiments by Lee *et al.* using multi-electrode recordings showed evidence for excitation and inhibition as predicted by this model, see [2],[4], and previous lectures.

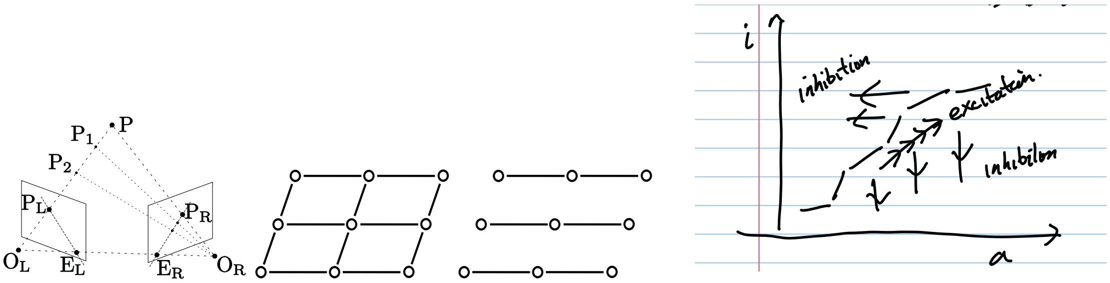

Fig. 1. Stereo. The geometry of stereo (left). A point P in 3-D space is projected onto points PL; PR in the left and right images. The projection is specified by the focal points OL,OR and the directions of gaze of the cameras (the camera geometry). The geometry of stereo enforces that points in the plane specified by P,OR OL, must be projected onto corresponding lines EL;ER in the two images (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. Right Panel: 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).

We can obtain a "neural model" by applying mean field theory to the model above. This requires finding an approximate distribution $Q(x)$ to $P(x)$ by defining a free energy function $F(q)$. We set $Q(x) = \prod_{ia} q_{ia}(x_{ia})$ and let $q_{ia}(1) = q_{ia}$. Then the free energy can be expressed as $F(q) = E(q) + \sum_{ia}\{q_{ia}\log q_{ia} - (1 + q_{ia})\log(1 - q_{ia})\}$. (See earlier lectures).

Then we define an update equation:

$$\frac{dq_{ia}}{dt} = -q_{ia}(1 - q_{ia})\frac{\partial F(q)}{\partial q_{ia}}. \tag{8}$$

This algorithm is guaranteed to decrease F monotonically and hence converge to a local minimum of $F(q)$. (Recall earlier lectures). Then we can re-express this update rule as:

$$\frac{du_{ia}}{dt} = -M(z_L^i, z_R^a) - 2\sum_{j,b} T_{ijab} q_{jb} - \theta_{ia} - u_{ia}. \tag{9}$$

Here $u_{ia} = \log(q_{ia}/(1 - q_{ia}))$ so $q_{ia} = \sigma(u_{ia})$ (sigmoid function). Also $T_{ijab} = A\delta_{ab} + B\delta_{ij} + F((i - a) - (j - b))G(i - a)G(j - b)$.

As in earlier lecture, the mean field theory approximation to the probabilistic model yields update rules that are similar to the "neural network" models proposed by Lee and his collaborators to account for their experiments.

References

[1] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–287, 1976.

[2] J. Samonds and B. Potetz. Cooperative and Competitive Interactions Facilitate Stereo Computations in Macaque Primary Visual Cortex. *The Journal of Neuroscience*, 29(50):15780–15795, 2009.

[3] J. Samonds and B. Potetz. Relative luminance and binocular disparity preferences are correlated in macaque primary visual cortex, matching natural scene statistics. In *Proceedings of the* . . . , 2012.

[4] J. M. Samonds, B. Potetz, and T. S. Lee. Relative luminance and binocular disparity preferences are correlated in macaque v1, matching natural scene statistics. *Proc Nat Acad Sci USA (PNAS)*, 109(16):6313–6318, April 2012.