Computational Vision
U. Minn. Psy 5036
Daniel Kersten
Lecture 25: Perceptual integration, Cooperative Computation

Initialize

- **Spell check off**

```
Off[General::spell1];
```

Outline

Last time

Object recognition

Today

Putting ideas together

Integrating perceptual information

Modular vs. cooperative computation

For the most part, we've treated visual estimation as if it is done in distinct "modules", such as, surface-color-from-radiance (Land, 1959), shape-from-shading (Horn, 1975), optic flow (Hildreth, 1983) or structure-from-motion (Ullman, 1979).

In contrast to the modularity theories of vision, it is phenomenally apparent that visual information is integrated to provide a strikingly singular description of the visual environment. By looking at how human perception puts integrates scene attributes, we may get some idea of how vision modules in the brain interact, and what they represent.

Some basic graph types in vision (Review from Lecture 6)

See: Kersten, D., & Yuille, A. (2003) and Kersten, Mamassian & Yuille (2004)

Basic Bayes

$$p[S \mid I] = \frac{p[I \mid S]\, p[S]}{p[I]}$$

Usually, we will be thinking of the **Y** term as a random variable over the hypothesis space, and **X** as data. So for visual inference, **Y = S** (the scene), and **X = I** (the image data), and **I = f(S)**.

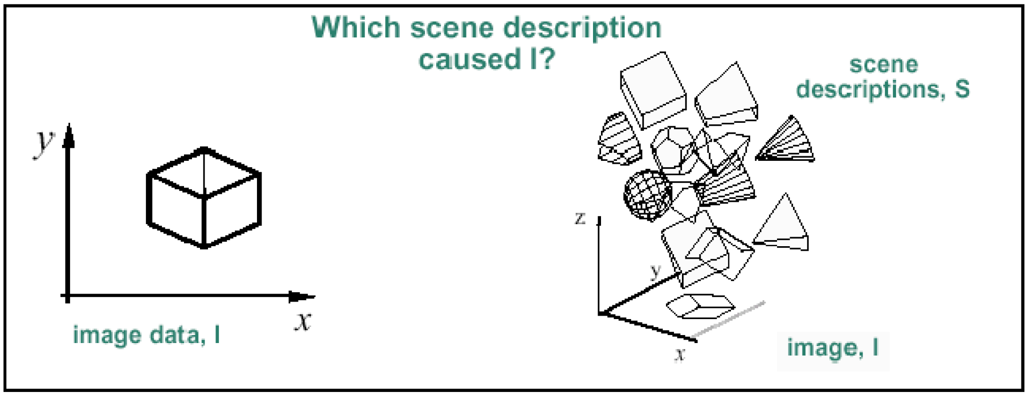
We'd like to have:

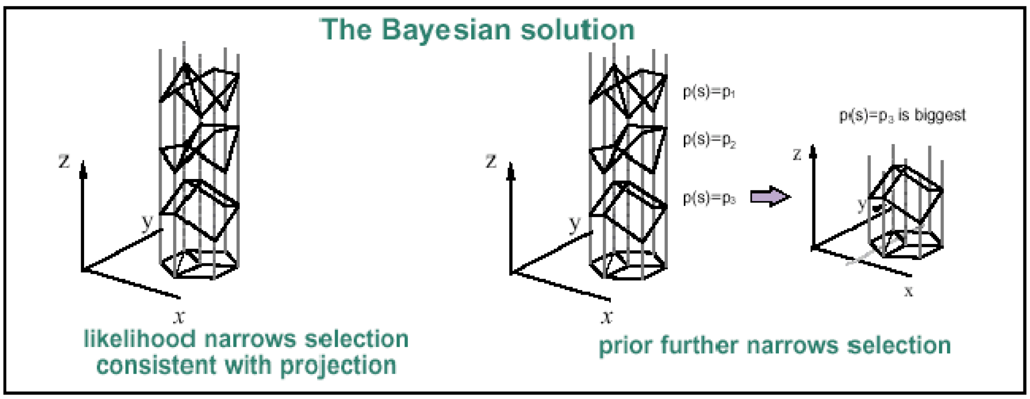
p(S|I) is the **posterior** probability of the scene given the image

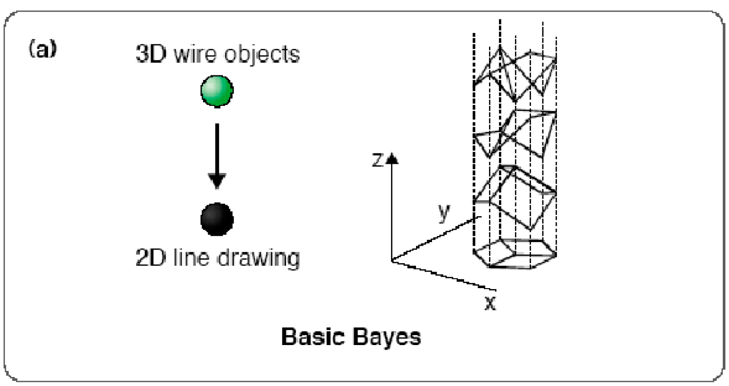
-- i.e. what you get when you condition the joint by the image data. The posterior is often what we'd like to base our decisions on, because as we discuss below, picking the hypothesis **S** which maximizes the posterior (i.e. maximum a posteriori or **MAP** estimation) minimizes the average probability of error.

p(S) is the **prior** probability of the scene.

p(I|S) is the **likelihood** of the scene. Note this is a probability of **I**, but not of **S**.







We've seen that the idea of prior assumptions that constrain otherwise underconstrained vision problems is a theme that pervades much of visual perception. Where do the priors come from? Some may be built in early on or hardwired from birth, and others learned in adulthood. See: Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. Nat Neurosci, 7(10), 1057-1058 for a recent example of learning the light from above prior for shape perception.

■ Low-level vision

We've seen a number of applications of Basic Bayes, including the algorithms for shape from shading and optic flow.

In 1985, Poggio, Torre and Koch showed that solutions to many of computational problems of low vision could be formulated in terms of maximum a posteriori estimates of scene attributes if the generative model could be described as a matrix multiplication, where the image I is matrix mapping of a scene vector S:

$$I = \mathbf{A}S$$

$$E = (I - \mathbf{A}S)^T (I - \mathbf{A}S) + \lambda S^T \mathbf{B} S$$

Then a solution corresponded to minimizing a cost function E, that simultaneously tries to minimize the cost due to reconstructing the image from the current hypothesis S, and a prior "smoothness" constraint on S. λ is a (often free) parameter that determines the balance between the two terms. If there is reason to trust the data, then λ is small; but if the data is unreliable, then more emphasis should be placed on the prior, thus λ should be bigger.

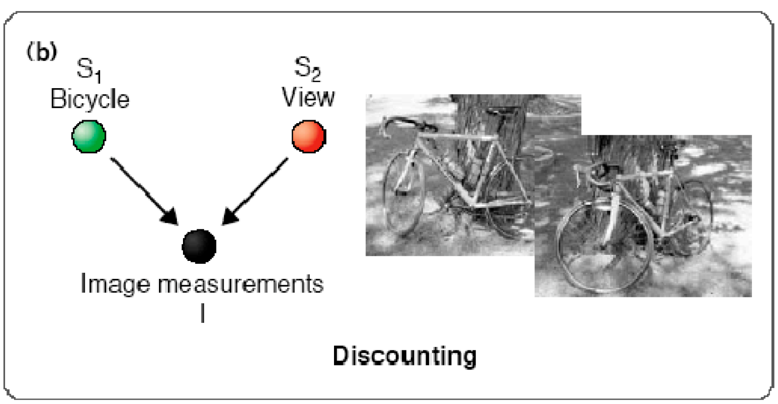
For example, S could correspond to representations of shape, stereo, edges, or motion field, and smoothness be modeled in terms of nth order derivatives, approximated by finite differences in matrix B.

The Bayesian interpretation comes from multivariate gaussian assumptions on the generative model:

$$p(I \mid S) = k \times \exp\left[-\frac{1}{2\sigma_n^2} (I - \mathbf{A}S)^T (I - \mathbf{A}S) \right]$$

$$p(S) = k' \times \exp\left[-\frac{1}{2\sigma_s^2} S^T \mathbf{B} S \right]$$

Discounting



This Bayes net describes the case where the joint distribution can be factored as:

$p(s_1, s_2, I) = p(I|s_1, s_2)p(s_1)p(s_2)$

Optimal inference for this task requires that we calculate the marginal posterior:

$p(s_1|I) \propto \int_{S_2} p(s_1, s_2 \mid I)\, ds_2$

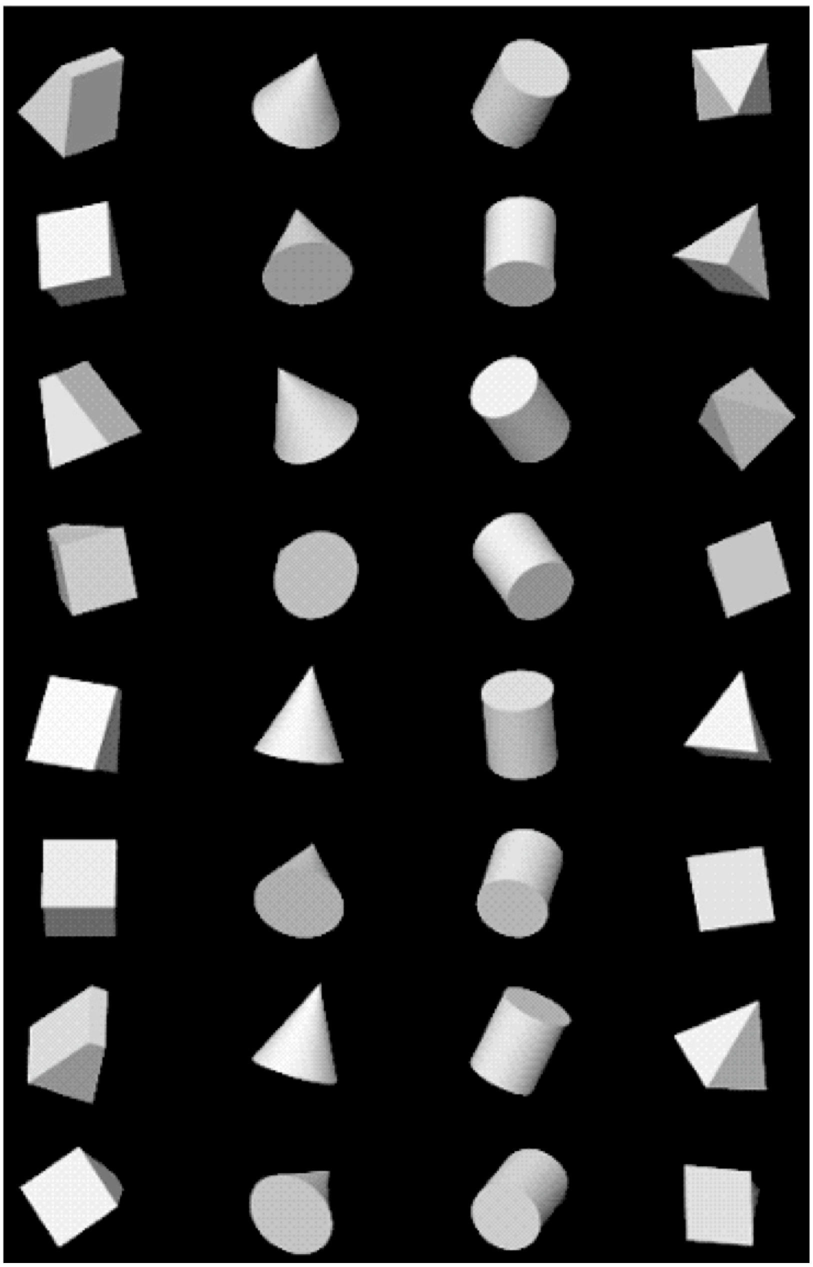
Liu, Knill & Kersten (1995) describe an example with:

I -> 2D x-y image measurements, s_1-> 3D object shape, and s_2-> view

Bloj et al. (1999) have an example estimating s_1-> surface chroma (saturation) with s_2-> illuminant direction.

■ Ideal observer for the "snap shot" model of visual recognition: Discounting views

Tjan et al. describe an application to object recognition in noise.



Eight views of four objects. (See Tjan B., Braje, W., Legge, G.E. & Kersten, D. (1995) Human efficiency for recognizing

3-D objects in luminance noise.Vision Research, 35, 3053-3069.)

Let **X** = the vector describing the image data. Let $\boldsymbol{O_i}$ represent object i, where i = 1 to N. Suppose that $\boldsymbol{O_i}$ is represented in memory by M "snap shots" of each object, call them views (or templates) $\boldsymbol{V}_{\mathbf{ij}}$, where j = 1, M.

$$p\,(O_i \mid X) = \sum_{j=1}^{M} p\,(V_{ij} \mid X)$$

$$= \sum_{j=1}^{M} \frac{p\,(X \mid V_{ij})\,p\,(V_{ij})}{p\,(X)}$$

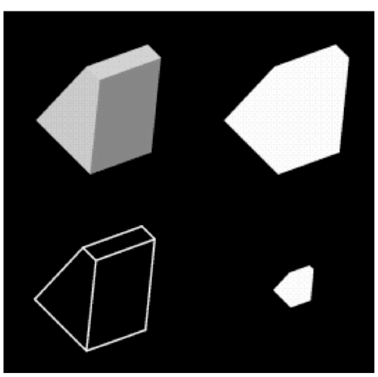
Given image data, Ideal observer chooses i that maximizes the posterior $p\,(O_i \mid X)$. If we assume that the p(X) is uniform, the optimal strategy is equivalent to choosing i that maximizes:

$$L\,(i) = \sum_{j=1}^{M} p\,(X \mid V_{ij})\,p\,(V_{ij})$$

If we assume i.i.d additive gaussian noise (as we did for the signal-known-exactly detection ideal), then

$$p\,(X \mid V_{ij}) = \frac{1}{\left(\sigma\sqrt{2\,\pi}\right)^{p}}\exp\left(-\frac{1}{2\,\sigma^2}\,\|X - V_{ij}\|^2\right)$$

where p is the number of pixels in the image.



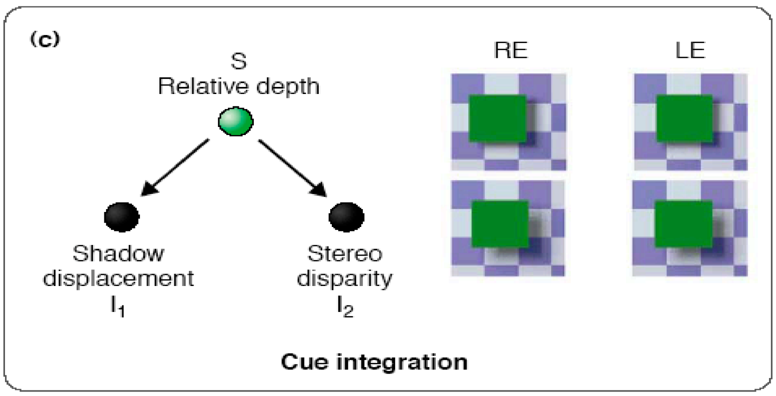
Tjan et al. showed that size, spatial uncertainty and detection efficiency played large roles in accounting for human object recognition efficiency. Interestingly, highest recognition efficiencies (~7.8%) were found for small silhouettes of the objects. (The small silhouettes were 0.7 deg, vs. 2.4 deg for the large silhouettes).

- **now..more on Cue integration and "Explaining away"**

Cue integration

- **Weak fusion**

Clark & Yuille, Landy & Maloney, Knill & Kersten, Schrater & Kersten.



This Bayes net describes the factorization:

p(S,I1,I2) = p(I1|S)p(I2|S)p(S)

- **Maximum a posteriori observer for cue integration: conditionally independent cues**

We'll change notation, and let x_1 and x_2 be image measurements or cues. The simple Bayes net shown above describes the case where the two cues are conditionally independent. In other words, p(x1,x2|s) = p(x1|s)p(x2|s).

Let's consider the simple Gaussian case where $x_i = \mu_{\text{cue}\,i} + n_i$. We'll show that optimal combined cue estimate is a weighted average of the cues.

p(s|x1,x2)=p(x1,x2|s)p(s)/p(x1,x2) $\propto$ p(x1|s)p(x2|s) =$e^{-(x_1-s)^2/2\sigma_1{}^2} e^{-(x_2-s)^2/2\sigma_2{}^2}$

```
PowerExpand[Log[E^(-(x1-μ)^2/(2 σ1^2)) E^(-(x2-μ)^2/(2 σ2^2))]]
```

$$-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}$$

$$D\left[-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}-\frac{(x_2-\mu)^2}{2\,\sigma_2^2},\ \mu\right]$$

$$\frac{x_1-\mu}{\sigma_1^2}+\frac{x_2-\mu}{\sigma_2^2}$$

$$\text{Solve}\left[\frac{x_1-\mu}{\sigma_1^2}+\frac{x_2-\mu}{\sigma_2^2}==0,\ \mu\right]$$

$$\left\{\left\{\mu\to\frac{x_2\,\sigma_1^2+x_1\,\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right\}\right\}$$

$$\left\{\left\{\mu\to\frac{x_2\,\sigma_1^2+x_1\,\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right\}\right\}\ /.\ \left\{\sigma_1^2\to 1/r_1,\ \sigma_2^2\to 1/r_2\right\}$$

$$\left\{\left\{\mu\to\frac{\frac{x_1}{r_2}+\frac{x_2}{r_1}}{\frac{1}{r_2}+\frac{1}{r_1}}\right\}\right\}$$

where $r_i\left(=\frac{1}{\sigma_i{}^2}\right)$, is called the reliability.

$$\mu\to\frac{r_1\,x_1}{r_1+r_2}+\frac{r_2\,x_2}{r_1+r_2}$$

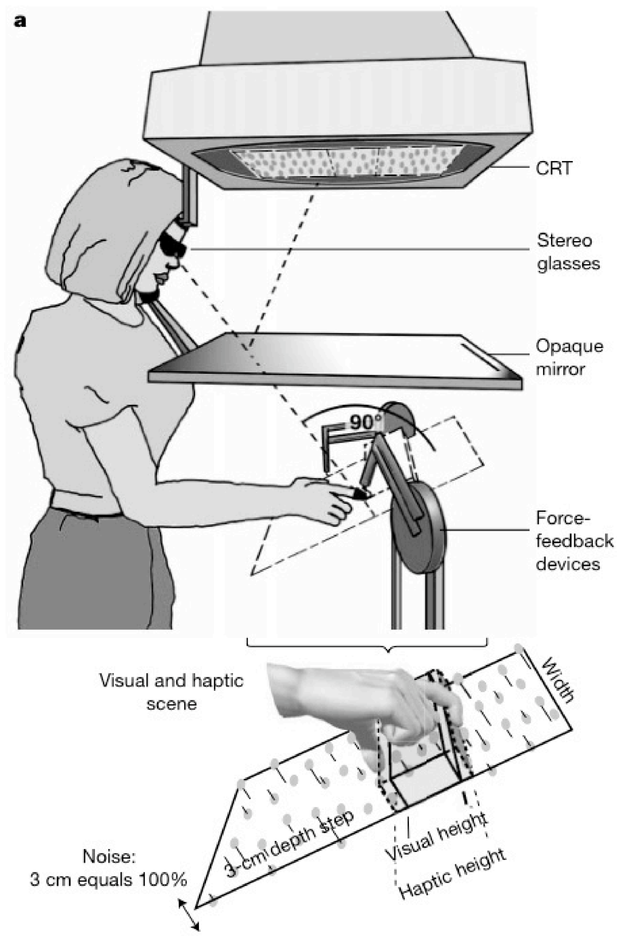
$$\mu\to\frac{r_1\,x_1}{r_1+r_2}+\frac{r_2\,x_2}{r_1+r_2}$$

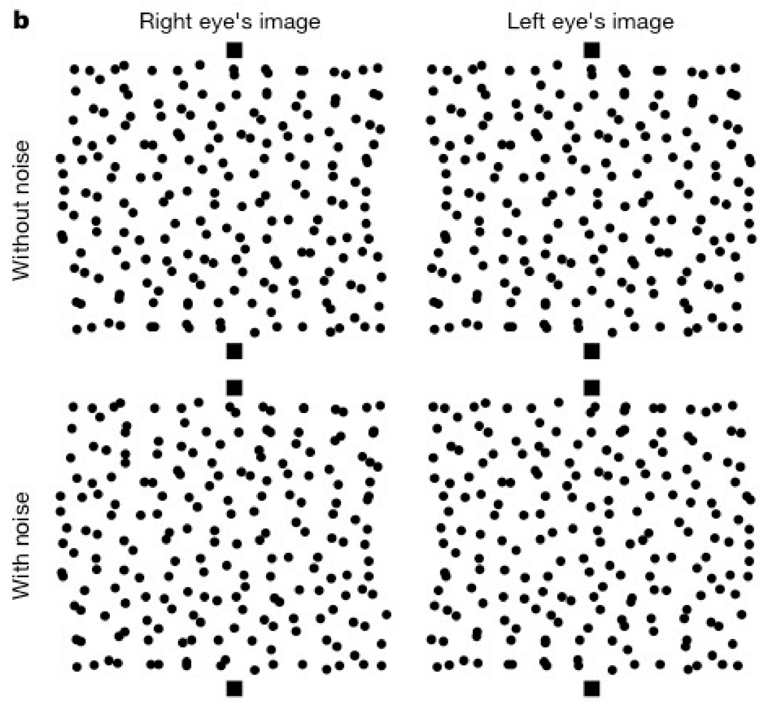
In general, one can show that the combined estimate is the weighted sum of the separate estimates, where the weights w_i are determined by the relative reliabilities :

$$\hat{\mu}_{\text{combined}}=\hat{\mu}_{\text{cue1}}\,w_1+\hat{\mu}_{\text{cue2}}\,w_2=\hat{\mu}_{\text{cue1}}\,\frac{r_1}{r_1+r_2}+\hat{\mu}_{\text{cue2}}\,\frac{r_2}{r_1+r_2}.$$

■ An application to integrating cues from vision and haptics (touch)

When a person looks and feels an object, vision often dominates the integrated percept--e.g. the perceived size of an object is typically driven more strongly by vision than by touch. Why is this? Ernst and Banks showed that the reliability of the visual and haptic information determines which cue dominates. They first measured the variances associated with visual and haptic estimation of object size. They used these measurements to construct a maximum-likelihood estimator that integrates both cues. They concluded that the nervous system combines visual and haptic information in a fashion that is similar to a maximum-likelihood ideal observer. Specifically, visual dominance occurs when the variance associated with visual estimation is lower than that associated with haptic estimate.





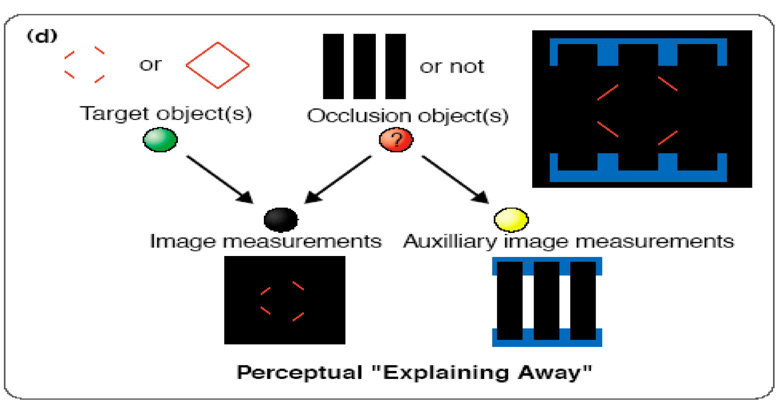
See Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Perceptual explaining away, Cooperative computation

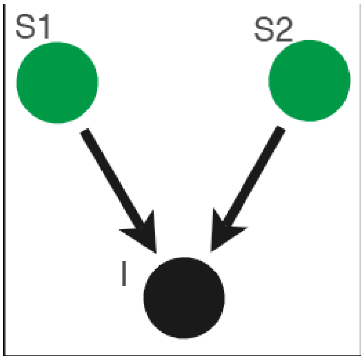
■ Perception as puzzle solving

Rock, I. (1983). *The Logic of Perception*. Cambridge, Massachusetts: M.I.T. Press.

■ Perceptual explaining away



Both causes S1 and S2 can be primary variables.



The above Bayes net describes the factorization:

p(S1,S2,I) = p(I|S2,S2) p(S1)p(S2)

If we average over I, S1 and S2 are independent. However, knowledge of I makes S1 and S2 conditionally dependent. The two causes S1 and S2 can behave like competing hypotheses to explain the data I.

In general, “explaining away” is a phenomenon that occurs in probabilistic belief networks in which two (or more) variables influence a third variable whose value can be measured (Pearl, 1988). Once measured, it provides evidence to infer the values of the influencing variables.

Imagine two coins that can be flipped independently, and the results (heads or tails) have an influence on a third variable. For concreteness, assume the third variable’s value is 1 if both coins agree, and 0 if not (NOT-XOR). If we are ignorant of the value of the third variable, knowledge of one influencing variable doesn't help to guess the value of the other—the two coin variables are independent. (This is called marginal independence, “marginal” with respect to the third variable, I)

But if the value of the third variable is measured (suppose it is 1), the two coin variables become coupled, and they are said to be *conditionally dependent*. Now knowing that one coin is heads guarantees that the other one is too.

The phrase “explaining away” arises because coupling of variables through shared evidence often arises in human reasoning, when the influences can be viewed as competing causes.

Suppose we have a prior reason to believe that both coins are heads, but we believe the second coin is even more likely than the first to be heads. Now we make a measurement, and discover the evidence is 0. The evidence explains away the "competing hypothesis" that both coins are heads, and in particular that the first coin is heads, because the first coin must be tails if the second coin is heads.

Human reasoning is particularly good at these kinds of inferences.

“Explaining away” is also a characteristic of perceptual inferences, for example when there are alternative perceptual groupings consistent with a set of identical or similar sets of local image features.

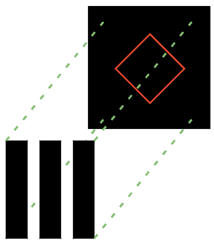
Demonstrations of cooperative computation and explaining away in perception

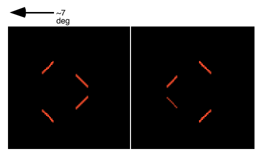
Several perceptual phenomena that we've seen before can be interpreted as "explaining away".

Occlusion & motion: Lorenceau & Shiffrar, Sinha

Recall translating diamond used to illustrate the aperture problem.

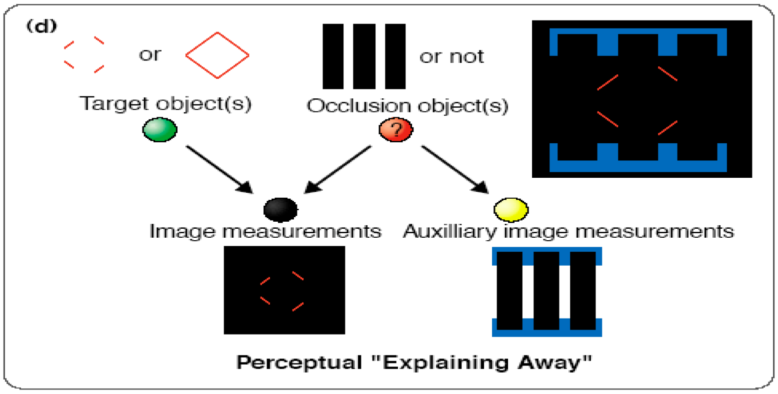
When the diamond is seen as coherently translating, one often also interprets the vertices as being covered by rectangular occluders.





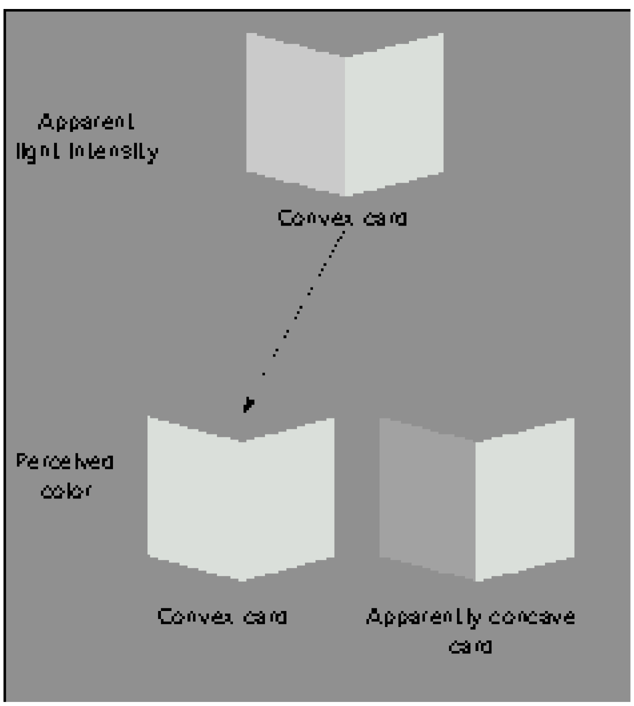
■ Translating diamond with "occluding occluders"

Occlusion as explaining away:



Lightness & surface geometry

■ Mach card

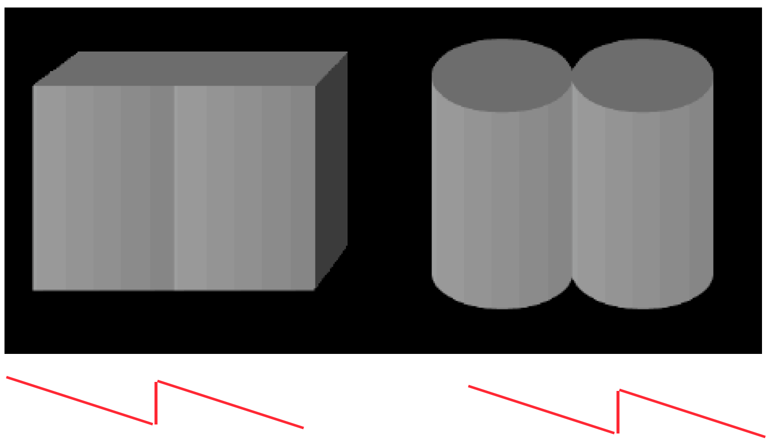


■ Lightness and shape

Recall the lightness demonstration that is similar to the Craik-O'Brien-Cornsweet effect, but difficult to explain with a simple filter mechanism (Knill, D. C., & Kersten, D. J., 1991). The idea is that the lightness of a pair of luminance gradients on the left of the figure below look different, whereas they look similar for the pair luminance gradients on the right. The reason seems to be due to the fact that the luminance gradients on the right are attributed to smooth changes in shape, rather than smooth changes in illumination.

http://vision.psych.umn.edu/www/kersten-lab/demos/lightness.html

These demonstrations suggest the existence of scene representations in our brains for shape, reflectance and light source direction.

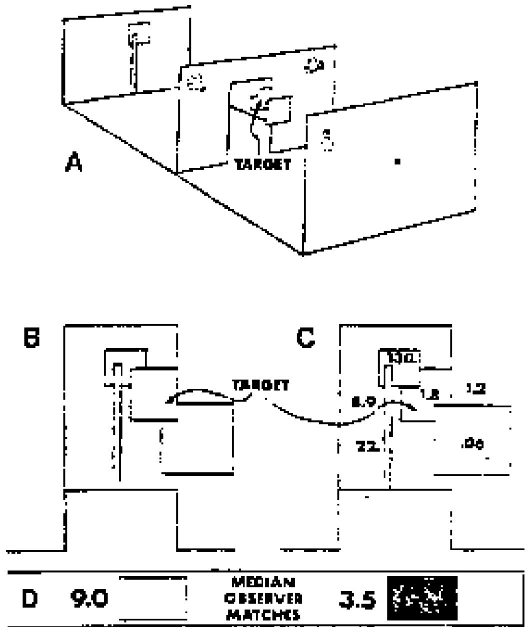


Draw a diagram to illustrate the above illusion in terms of "explaining away"

■ Dependence of lightness on spatial layout

Gilchrist:

In the 1970's, Alan Gilchrist was able to show that the lightness of a surface patch may be judged either dark-gray, or near-white with only changes in perceived spatial layout! (Gilchrist, A. L. (1977). How did he do this? What is going on? Interpret lightness as reflectance estimation.



igure 5. (A) Perspective view of the parallel planes display, owing hidden light bulbs. The display was seen through the inholes in which the target appeared to be located either (B) in e near plane or (C) in the far plane, with luminances shown foot-Lamberts. (D) The average match from a Munsell chart the two displays.

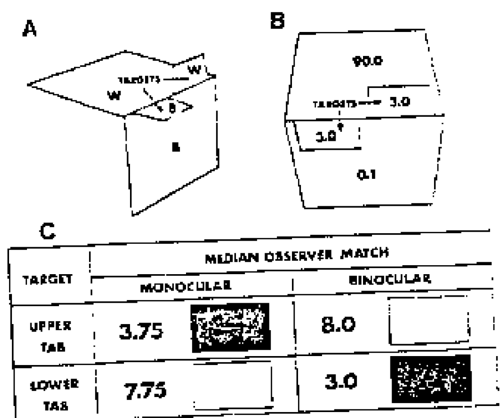


Figure 4. (A) Perspective view of the stimulus display used in the critical test, showing color (B, black; W, white) of each part. (B) Monocular retinal pattern showing luminances in foot-Lamberts. (C) Average Munsell matches for monocular and ... conditions.

o The Room-in-a-Shoe-Box experiment

o Coplanar card experiment

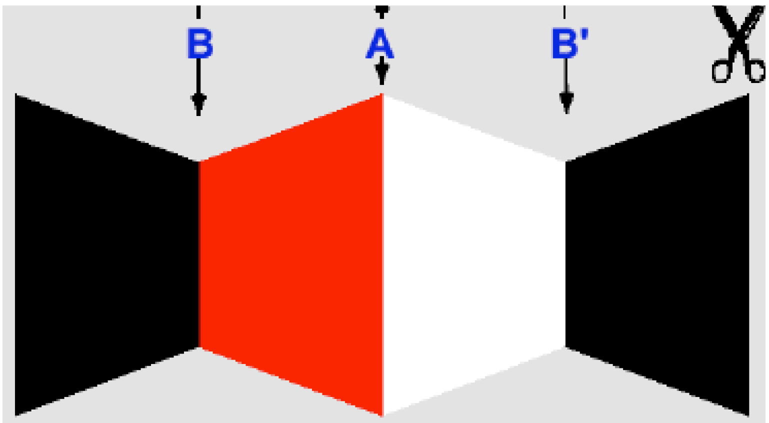
The left and right inner gray disks in the above figure are the same intensity. In classic simultaneous contrast, the brighter annulus on the right makes the inner disk appear darker.

Color & shape

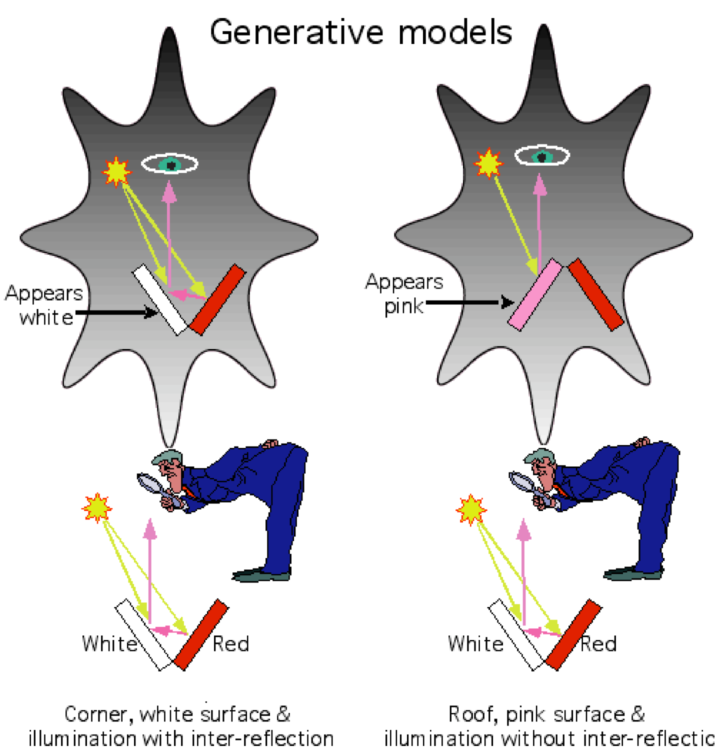
■ Bloj, Kersten & Hurlbert

Demo

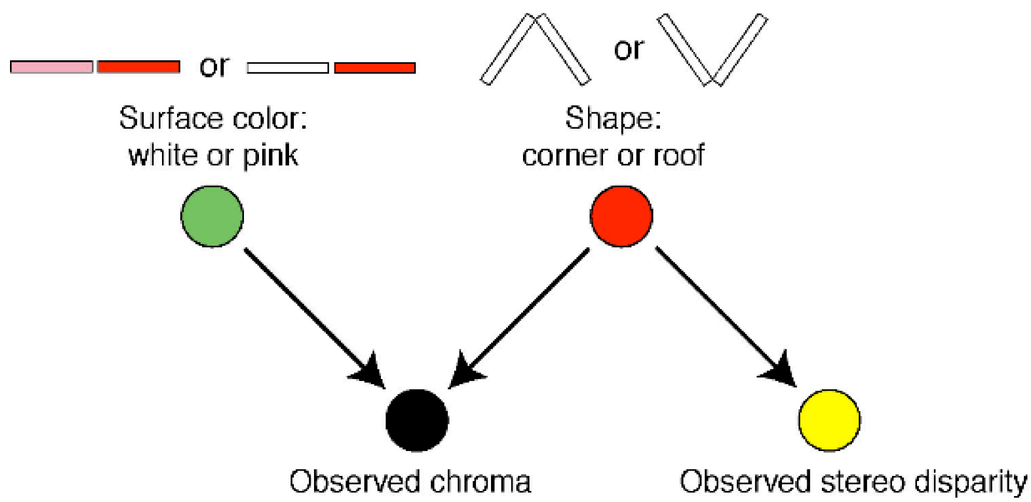
http://gandalf.psych.umn.edu/users/kersten/kersten-lab/Mutual_illumination/BlojKerstenHurlbertDemo99.pdf



Interpretation



Interreflection as explaining away. Stereo can be used as an auxiliary cue to change the perceived shape from concave to convex.



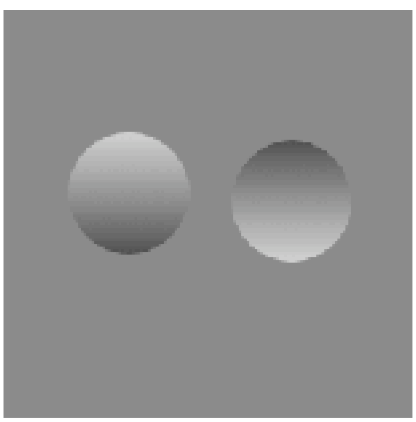
Dependence of shape on perceived light source direction

Dependence of shape on perceived light source direction

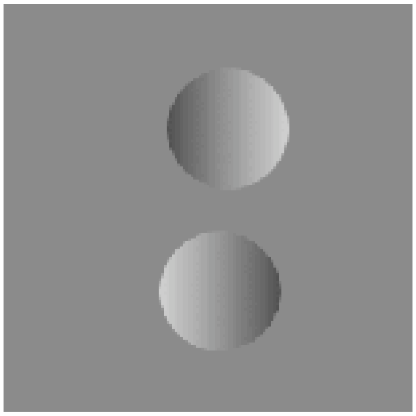
Brewster (1926), Gibson, Ramachandran, V. S. (1990), crater illusion and the single light source assumption

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. Nat Neurosci, 7(10), 1057-1058.

■ Vertical light direction



■ Horizontal light direction



Transparency

■ Motion and transparency (Kersten et al., 1992)

Dependence of transparency on perceived depth

Kersten and Bülthoff

o orientation and transparency

o transparency and depth from motion--computer demo

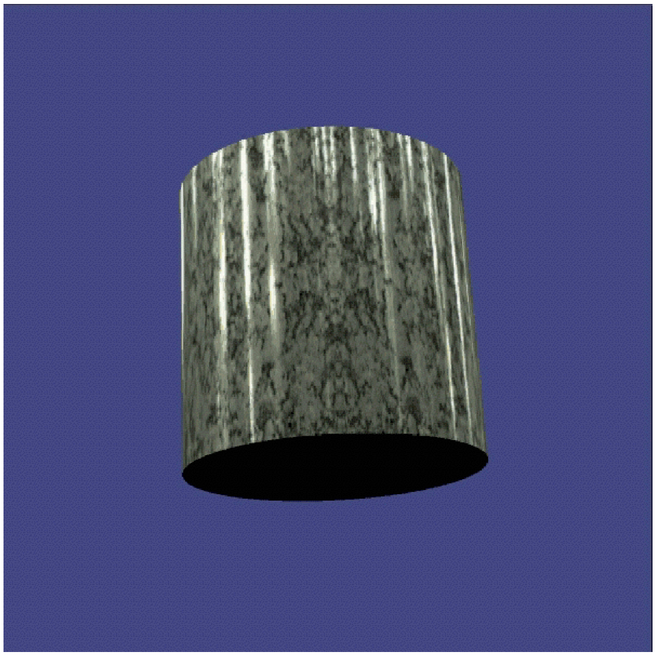
http://gandalf.psych.umn.edu/users/kersten/kersten-lab/demos/transparency.html

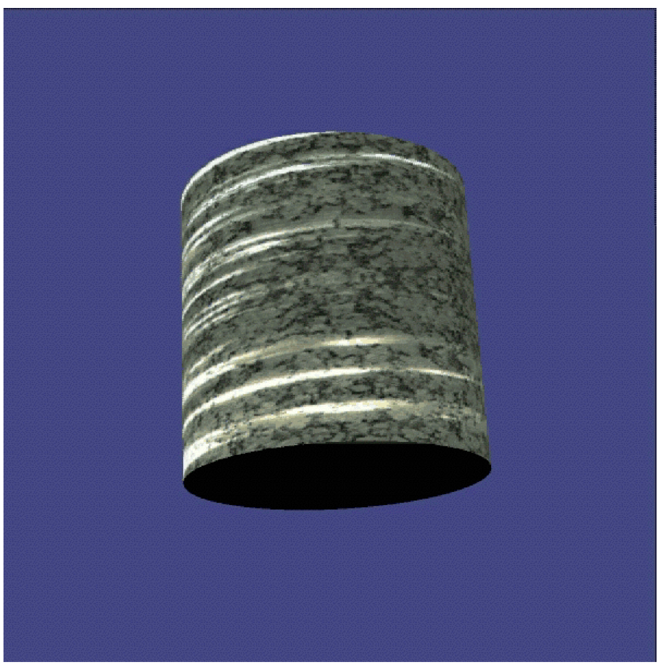
Nakayama, Shimojo (1992)

o transparency and depth from stereo demos, neon color spreading

Dependence on curvature

(Dr. Bruce Hartung)





Application to image parsing, object recognition

Incorporating higher-level knowledge--Image parsing and recognition using cooperative computation

For current work in this area applied to segmentation and recognition, see: Tu Z, Zhu S-C (2002), Zhu and Tu (2000). For a review, see: Yuille and Kersten (2006).

http://gandalf.psych.umn.edu/users/kersten/kersten-lab/courses/Psy5036W2008/Lectures/24.%20ObjectRecognition-Clutter/Lecture24TopDown.pdf

References

■ Cue integration, cooperative computation

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. Nat Neurosci, 7(10), 1057-1058.

Barrow, H. G., & Tenenbaum, J. M. (1978). Recovering Intrinsic Scene Characteristics from Images. In A. R. Hanson, & E. M. Riseman (Ed.), Computer Vision Systems (pp. 3-26). New York: Academic Press.

Bergstrom, S. S. (1977). Common and Relative Components of Reflected Light as Information About the Illumination, Colour and Three-Dimensional Form of Objects. 18, 180-186).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Brewster, D. (1826). On the optical illusion of the conversion of cameos into intaglios and of intaglios into cameos, with an account of other analogous phenomena. *Edinburgh Journal of Science, 4*, 99-108.

Clark, J. J., & Yuille, A. L. (1990). Data Fusion for Sensory Information Processing . Boston: Kluwer Academic Publishers.

Ernst MO, Banks MS, Bulthoff HH (2000) Touch can change visual slant perception. Nat Neurosci 3:69-73.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Gilchrist, A. L. (1977). Perceived Lightness Depends on Perceived Spatial Arrangement. Science, 195, 185-187.

Gibson, J. J. (1950). *The Perception of the Visual World*. Boston, MA: Houghton Mifflin.

Hillis JM, Ernst MO, Banks MS, Landy MS (2002) Combining sensory information: mandatory fusion within, but not between, senses. Science 298:1627-1630.

Humphrey, K. G., Goodale, M. A., Bowen, C. V., Gati, J. S., Vilis, T., Rutt, B. K., & Menon, R. S. (1996). Differences in Perceived Shape from Shading Correlate with Activity in Early Visual Areas.,, 1-16.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3, 79-87.

Jacobs RA (2002) What determines visual cue reliability? Trends Cogn Sci 6:345-350.

Kersten, D. J. (1991). Transparency and the Cooperative Computation of Scene Attributes. In M. Landy, & A. Movshon (Ed.), Computational Models of Visual Processing. Cambridge, Massachusetts: M.I.T. Press.

Kersten, D., Bülthoff, H. H., Schwartz, B., & Kurtz, K. (1992). Interaction between transparency and structure from motion. Neural Computation, 4(4), 573-589.

Kersten D, Yuille A (2003) Bayesian models of object perception. Current Opinion in Neurobiology 13:1-9.

Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian Inference. Annual Review of Psychology 55:271-304.

Knill, D. C., & Kersten, D. (1991). Apparent surface curvature affects lightness perception. Nature, 351, 228-230.

Mach, E. (*1886*, 1959). The Analysis of Sensations . New York: Dover.

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. Science, 257, 1357-1363.

Pearl J (1988) Probabilistic reasoning in intelligent systems : networks of plausible inference, Rev. 2nd printing. Edition. San Mateo, Calif.: Morgan Kaufmann Publishers.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314-319.

Poggio, T., Gamble, E. B., & Little, J. J. (1988). Parallel integration of vision modules. Science, 242, 436-440.

Ramachandran, V. S. (1990). Visual perception in people and machines. In A. Blake, & T. Troscianko (Ed.), A.I. and the Eye John Wiley & Sons Ltd.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks . Cambridge, UK: Cambridge University Press.

Schrater PR, Kersten D (2000) How optimal depth cue integration depends on the task. *International Journal of Computer Vision* 40:73-91.

Tjan B., Braje, W., Legge, G.E. & Kersten, D. (1995) Human efficiency for recognizing 3-D objects in luminance noise.Vision Research, 35, 3053-3069.

Todd, J. T., & Mingolla, E. (1983). Perception of Surface Curvature and Direction of Illumination from Patterns of Shading. Journal *of Experimental Psychology: Human Perception & Performance, 9*(4), 583-595.

Tu Z, Zhu S-C (2002) Parsing Images into Region and Curve Processes. In: Proc. of the 7th European Conference on Computer Vision, p 393 ff. Copenhagen, Denmark: Springer-Verlag, Berlin Heidelberg.

von der Heydt R, Friedman H, Zhou HS (2003) Searching for the neural mechanisms of color filling-in. In: Filling-in: From Perceptual Completion to Cortical Reorganization (Pessoa L, P DW, eds), pp 106-127. Oxford: Oxford University Press.

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In K. D.C., & R. W. (Ed.), Perception as Bayesian Inference Cambridge, U.K.: Cambridge University Press.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci, 10(7), 301-308. http://gandalf.psych.umn.edu/~kersten/kersten-lab/papers/yuillekerstenTICs2006.pdf

Zhu S-C, Zhang R, Tu Z (2000) Integrating Bottom-up/Top-Down for Object Recognition by Data Driven Markov Chain Monte Carlo. In: Proc. of Int'l Conf. on Computer Vision and Pattern Recognition. SC.