Perception as an Inference Problem

Bruno A. Olshausen
Helen Wills Neuroscience Institute, School of Optometry and Redwood Center for Theoretical Neuroscience
UC Berkeley
What are the principles governing information processing in this system?
Chowdhury and DeAngelis (2008) replicate the finding that monkeys initially trained on coarse stereopsis show impaired coarse depth discrimination when muscimol is injected into MT. Remarkably, the same animals, after a second round of training on fine stereopsis, are unimpaired at either fine or coarse depth discrimination by similar injections. Moreover, recordings in MT show that neuronal responses are not altered by learning the fine stereopsis task. Given the differences between the tasks and the large number of visual areas containing disparity-sensitive neurons, one might not be surprised to find different areas involved in the two tasks. But it is quite unexpected that merely learning one task would change the contribution of areas previously involved in the other. Chowdhury and DeAngelis conclude that the change in outcome reflects a change in neural decoding—decision centers that decode signals to render judgments of depth, finding MT signals unreliable for the fine stereopsis task, switch their inputs to select some better source of disparity information. Candidates include ventral stream areas V4 or IT, where relative disparity signals have been reported (Orban, 2008) and which contain far more neurons (Figure 1). When challenged afresh with the coarse depth task, these same decision centers may now find that their new sources of information can solve the coarse task as well as the old ones. MT is no longer critical. Perhaps in other monkeys MT would never have a role in stereopsis at all. Chowdhury and DeAngelis' monkeys were trained simultaneously or previously to discriminate motion, which engages MT. Faced with a qualitatively similar random dot stimulus, it might make sense for the cortex to try to solve the new problem of stereopsis with existing decoding strategies. But if the animals were initially trained on a different task—say, a texture discrimination—MT might never be engaged at all. It would also be interesting to see the outcome if monkeys were trained on depth tasks that were less different and could be interleaved in the same sessions, for example noise-limited depth judgments using similar absolute or relative disparity

Figure 1. A Scaled Representation of the Cortical Visual Areas of the Macaque

Each colored rectangle represents a visual area, for the most part following the names and definitions used by Felleman and Van Essen (1991). The gray bands connecting the areas represent the connections between them. Areas above the equator of the figure (reds, browns) belong to the dorsal stream. Areas below the equator (blues, greens) belong to the ventral stream. Following Lennie (1998), each area is drawn with a size proportional to its cortical surface area, and the lines connecting the areas each have a thickness proportional to the estimated number of fibers in the connection. The estimate is derived by assuming that each area has a number of output fibers proportional to its surface area and that these fibers are divided among the target areas in proportion to their surface areas. The connection strengths represented are therefore not derived from quantitative anatomy and furthermore represent only feedforward pathways, though most or all of the pathways shown are bidirectional. The original version of this figure was prepared in 1998 by John Maunsell.
Two views of visual system function

• **Deduction**
  - feature extraction, classification
  - (Hubel & Wiesel; Fukushima; ‘deep learning’)

• **Inference**
  - generative models, recurrent computation
  - (Helmholtz; Nakayama; Kersten & Yuille; Geman; Lee & Mumford)
Hubel & Wiesel (1962, 1965)

Hypercomplex

Complex

Simple
input data ($x$)

output ($y$)

$y = f(x; w)$
Is this the goal of vision?
Visual Navigation in Box Jellyfish

To determine what medusae of T. cystophora, the approximately 5 m tall sand wasp, would see, we used the optical model to calculate the point-spread function of the optics at different retinal locations. Applying these point-spread functions to still images of Snell's window in the mangrove swamp, we made still pictures using a wide-angle camera to cover just the edge of Snell's window when the mangrove canopy was filled with water and medusae, which shifted from covering most of Snell's window to covering just the edge of Snell's window. Thus, the approximately 1 cm large lens eye is indeed specialized for looking up through the water to detect the mangrove canopy through surface ripple and the height of the mangrove trees.

The pictures were taken from just under the surface to make sure that the visual surroundings were detected even at a distance of 8 m (detection depends on the amount of surface ripple and the height of the mangrove trees). In the pictures, it was easy to follow the movement of medusae through the images. In the pictures, it was easy to follow the movement of medusae through the images.

To evaluate the possibility that the upper lens eye detects the position of the mangrove canopy through water currents, we used the optical model to determine what medusae of T. cystophora, which are not specialized for looking up through the water, would see when the tank was placed 2 or 4 m into the lagoon. At 8 m from the canopy edge, the medusae could still detect the mangrove canopy. This suggests that they responded to the displacement along the edges of the tank, constantly bumping into it, suggesting that they responded to the displacement along the edges of the tank, constantly bumping into it.

When the tank was placed 2 or 4 m into the lagoon, the medusae ceased feeding and swam toward the nearest mangrove trees. Most importantly, their mean swimming direction differed from the closest lagoon edge. At 8 m from the canopy edge, the medusae could still detect the mangrove canopy.

Experiments were conducted on wild populations of T. cystophora in the mangrove lagoons near La Parguera, Puerto Rico. Preliminary tests demonstrated that if a group of medusae was released in the tank, they would swim toward the nearest edge. When the tank was placed 2 or 4 m into the lagoon, the medusae ceased feeding and swam toward the nearest mangrove trees. This suggests that they responded to the displacement along the edges of the tank, constantly bumping into it.

The tank, with the trapped medusae in the mangrove lagoon, was filled with water, it was lightly buoyant such that the walls of the cylindrical tank were extended 1–2 cm above the external water surface, effectively sealing off the water around the animals but without affecting the visual surroundings. A group of medusae was released in the tank, and as long as the tank remained under the canopy, they were able to find their habitat between the mangrove prop roots. When the tank was placed 2 or 4 m into the lagoon, the medusae ceased feeding and swam toward the nearest mangrove trees.

The results thus predict that if the amount of surface ripple and the height of the mangrove trees is detected even at a distance of 8 m, the lens eye can be used to detect the mangrove canopy. This suggests that the lens eye is indeed specialized for looking up through the water to detect the mangrove canopy through surface ripple and the height of the mangrove trees.

The lens eye is used to detect the mangrove canopy through surface ripple and the height of the mangrove trees. This suggests that the lens eye is indeed specialized for looking up through the water to detect the mangrove canopy through surface ripple and the height of the mangrove trees.

Figure 1. Rhopalial Orientation and Visual Field

Figure 2

Figure 3

Table S1

*: *C14
\[ \tau \dot{s} + s = g(s, x, a; w) \]
\[ a = f(s) \]
Vision as inference

World --------> Image <-------- Model
Separation of shape and reflectance

From a physical point of view, the problem of lightness constancy is as follows. An illuminance image, \( E(x,y) \), is defined as the perceived intensity of light at a point \((x,y)\) in the scene, and a luminance image, \( L(x,y) \), is defined as the perceived reflectance of a surface at that point.

Luminance, illuminance, and reflectance, are physical quantities that can be measured by physical devices. There are also two subjective variables that must be discussed. Patches differ in brightness. Patches differ in lightness. Brightness is sometimes defined as perceived luminance.

Intrinsic image decomposition (Arend, 1994; Adelson and Pentland, 1996) has been proposed for understanding lightness perception. Unfortunately, Land and McCann's constraints fail when applied to the checker-block image. Figure 24.8(a) shows the checker-block and its analysis into two intrinsic images. Figure 24.6 demonstrates that the block is made of a 2x2 set of cubes, each colored either gray or black.

Note that Land and McCann's constraints fail when applied to the checker-block image. Figure 24.8(a) shows the checker-block and its analysis into two intrinsic images. Unfortunately, they happen to have the same luminance, because they have the same reflectance, but different illuminance. If they share the same reflectance, but different illuminance, it is counterbalanced by its higher luminance.
Possible neural circuits for inferential computation in V1

1. Sparse coding

2. Separating form and motion from time-varying images
Sparse coding image model

(Olshausen & Field, 1996; Chen, Donoho & Saunders 1995)

\[ I(\vec{x}) = \sum_{i=1}^{M} a_i \phi_i(\vec{x}) + \epsilon(\vec{x}) \]
Energy function

\[ E = \frac{1}{2} |I - \Phi a|^2 + \lambda \sum_i C(a_i) \]
Energy function

\[ E = \frac{1}{2} |I - \Phi a|^2 + \lambda \sum_i C(a_i) \]

\[-\log P(I | a) \quad P(a)\]
Coefficients $a_i$ may be computed via thresholding and lateral inhibition (‘LCA’ - Rozell, Johnson, Baraniuk & Olshausen, 2008)

$$G_{ij} = \langle x_i \rangle \langle x_j \rangle$$

$$b_i = \sum_x \phi_i(x) I(x)$$

$$G_{ij} = \sum_x \phi_i(x) \phi_j(x)$$
Two examples

1. Sparse coding

2. Separating form and motion from time-varying images
Visual perception requires separation of form and motion from time-varying retinal images

(eye movement data from Austin Roorda, UC Berkeley)
Simple averaging is not sufficient
The problem

\[ I(\vec{x}, t) = S(\vec{x} - \Delta \vec{x}(t)) + \epsilon(\vec{x}, t) \]

\[ \Delta \vec{x}(t) = \arg \min_{\Delta \vec{x}(t)} |I(\vec{x}, t) - S(\vec{x} - \Delta \vec{x}(t))|^2 \]

\[ \hat{S}(\vec{x}) = \int I(\vec{x} + \Delta \vec{x}(t)) \, dt \]
Traditional models compute motion and form independently

- Time-varying image
  - Motion energy and pooling
    - Optic flow
  - Feature extraction and pooling
    - Invariant pattern recognition
Traditional models compute motion and form independently.
Motion \textit{and} form must be estimated simultaneously.
Graphical model for separating form and motion
(Alex Anderson, Ph.D. thesis)

\[ \hat{S} = \arg \max_S \log P(R|S) \]
Given current estimate of position (X), update S.
Given current estimate of form (S), update X

\[ P(X_t \mid R_{0:t}) \]

\[ P(X_{t+1} \mid R_{0:t}) \]

\[ R_{t+1} \]

\[ S = S^t \]

\[ P(R_{t+1} \mid X_{t+1}, S = S^t) \]

\[ P(X_{t+1} \mid R_{0:t+1}) \]
Joint estimation of form (S) and position (X)
Including a prior over form ($S$)

Eye position

Spikes
(from LGN afferents)

Pattern

Dictionary

Sparse representation

$$\hat{A} = \arg \max_A \log P(R|A) + \log P(A)$$

sparse
Learned dictionary $D$
Prior over form (S) improves inference
Form prior improves inference

SNR as a function of time, DC = 100.0

- **Sparse Prior**
- **Independent Pixel Prior**

![SNR graph](image)
Main points

• Perception seems better described as an inference problem that attempts to disentangle underlying causes from image data.

• Inference involves bidirectional information flow both within and between levels of representation.

• This moves us away from thinking of ‘receptive fields’ and instead toward how populations of neurons interact to perform collective computations.
Papers


http://redwood.berkeley.edu/bruno/papers/rozell-sparse-coding-nc08.pdf

http://redwood.berkeley.edu/bruno/papers/highly-overcomplete-SPIE.pdf