Probing sensory representations with metameric stimuli

Eero Simoncelli
HHMI / New York University
Where does all that visual information go?

[figure: Hubel ‘95]
Destiny of sensory information

Sensory input

Discard

Act

Remember (brain)
Metamers

- Two stimuli that are physically different, but appear the same to a human observer
- Classic example: trichromatic color perception
- Another example: texture perception
Spectral nature of light

[Newton, 1665]
Perceptual color matching experiment

Arbitrary test light

Mixture of 3 primary lights

[Young, Helmholtz, Grassman, etc., 1800’s; slide c/o D. Brainard]
Perceptual color matching experiment

Arbitrary test light

Mixture of 3 primary lights

[Young, Helmholtz, Grassman, etc, 1800’s; slide c/o D. Brainard]
Theory (Grassman, 1853): the visual system performs a **linear projection** of the wavelength spectrum onto a three-dimensional response space

- Predicts/explains perceptual “metamers” - lights that appear identical, but have physically distinct wavelength spectra (1800’s)
- Codified in CIE standards for color representation (1931)
- Underlying mechanism (cone photoreceptors) verified (1987)
Measured Primate L, M, S Spectral Sensitivities (Log Scale)

[Baylor, Nunn & Schnapf, 1987]
Visual texture

Homogeneous, with repeated structures

Let us say that to the extent that visible objects are different and far apart, they are forms. To the extent that they are similar and congregated they are a texture. A man has form; a crowd has man-texture. A leaf has form; an arbor has leaf texture, and so on.

[Lettvin, 1976]
Julesz (1962)

• Hypothesis: Two textures with identical Nth-order pixel statistics will appear the same (for some N).

• Hand-constructed counter-examples (N=3):

Julesz ‘78

Yellott ‘93
Physiologically-inspired Julesz-style texture model

[Portilla & Simoncelli, 2000]
Texture synthesis

[Portilla & Simoncelli, 2000]
Images

Manifold of images with identical model responses

Model responses
Experimental logic

If model captures the same properties as the visual system, images with identical model responses should appear identical to a human.
Pairs of images with identical model responses:

Top: original,  Bottom: synthesized

[Portilla & Simoncelli 2000]
Other explorations:

- inpainting
- interpolation
- seeds

As a demonstration of the flexibility of our approach, we can modify the algorithm to handle applications of constrained texture synthesis. In particular, consider the problem of extending a texture image beyond its spatial boundaries (spatial extrapolation). We want to synthesize an image in which the central pixels contain a copy of the original image, and the surrounding pixels are synthesized based on the statistical measurements of the original image. The set of all images with the same central subset of pixels is convex, and the projection onto such a convex set is easily inserted into the iterative loop of the synthesis algorithm. Specifically, we need only re-set the central pixels to the desired values on each iteration of the synthesis loop. In practice, this substitution is done by multiplying the desired pixels by a smooth mask (a raised cosine) and adding this to the current synthesized image multiplied by the complement of this mask. The smooth mask prevents artifacts at the boundary between original and synthesized pixels, whereas convergence to the desired pixels within the mask support region is achieved almost perfectly. This technique is applicable to the restoration of pictures which have been destroyed in some subregion ("filling holes") (e.g., Hirani and Totsuka, 1996), although the estimation of parameters from the defective image is not straightforward. Figure 19 shows a set of examples that have been spatially extrapolated using this method. Observe that the border between real and synthetic data is barely noticeable. An additional potential benefit is that the synthetic images are seamlessly periodic (due to circular boundary-handling within our algorithm), and thus may be used to tile a larger image.

Finally, we consider the problem of creating a texture that lies visually "in between" two other textures. The parameter space consisting of spatial averages of local functions has a type of convexity property in the limit as the image lattice grows in size. Figure 20 shows three images synthesized from parameters that are an average of the parameters for two example textures. In all three cases, the algorithm converges to an interesting-looking image that appears to be a patchwise mixtures of the two initial textures, rather than a new homogeneous texture that lies perceptually between them. Thus, in our model, the subset of parameters corresponding to textures (homogeneous RFs) is not convex!
Structural seeding [cf. “adversarial examples” - Szegedy et. al. 2014]
Can we generalize to inhomogeneous stimuli?

Can we make the model more physiological?
Retina
Optic Nerve
LGN
Optic Tract
Visual Cortex

(V1)

[figure: Hubel ‘95]
Retina

Dorsal pathway: V1->V3->V5
position, motion, action

Ventral pathway: V1->V2->V4-> IT
spatial form, recognition, memory

[Ungerleider & Mishkin, 1982]
• Visual neurons respond to content within a small region of the visual input known as the **Receptive Field (RF)**

• In each visual area, we assume RFs cover the entire visual field
Inhomogeneity - RF sizes grow with eccentricity

Retinal ganglion (midget) cell receptive fields (macaque, magnified x10) [Perry et.al., 1984; Watanabe & Rodiek, 1989]

Modified Snellen acuity chart (threshold, x10) [after Anstis, 1973]
[after Geisler et al., 1999]
RF sizes grow with eccentricity

[Freeman & Simoncelli 2011, data from Gattass et. al., 1981; Gattass et. al., 1988; Perry et. al., 1984]
V1 simple cell

[Hubel & Wiesel, 1962]

V1 complex cell

linear weights  rectifying nonlinearity
Local texture representation in the ventral stream

V2 receptive fields  V1 cells  Joint statistics

3.1
1.4
12.5
Local texture representation in the ventral stream

Local correlational statistics can be re-expressed as a “subunit” model...
Canonical computation in the ventral stream

Substantial information loss $\Rightarrow$ model predicts metamers
Canonical sensory computation

• Linear filter (determines pattern selectivity)

• Rectifying nonlinearity

• Local pooling (e.g., average, max)

• Local gain control

• Noise

Cascaded ...  

[eg. Douglas, 1989;  
Heeger, Simoncelli & Movshon 1996;  
Heeger & Carandini 2014]
[Koch & Poggio, 1999; cf. Fukushima, 1980; Serre, Oliva, Poggio 2007; etc]
Synthesizing Ventral Stream Metamers

Original image

Model responses

3.1
1.4
12.5
.
.
.

[Freeman & Simoncelli, 2011]
Synthesizing Ventral Stream Metamers

Model responses

3.1
1.4
12.5
.
.
.

[Freeman & Simoncelli, 2011]
Model RF size (diam / eccentricity)

[Freeman & Simoncelli, 2011]
Scaling (radius / eccentricity)

Proportion correct

Model RF size (diam / eccentricity)

chance

Freeman & Simoncelli, 2011
\[ D = \Phi \left( \left[ 1 - \frac{s^2_{\text{human}}}{s^2_{\text{model}}} \right] \right) \]

[Freeman & Simoncelli, 2011]
Scaling (diameter / eccentricity)

S1  S2  S3  S4  Average

[Freeman & Simoncelli, 2011]
RF sizes grow with eccentricity

![Graph showing the relationship between receptive field size and eccentricity, with lines indicating growth in receptive field size across different visual areas (V1, V2, V4).]

[Freeman & Simoncelli 2011, from Gattass et. al., 1981; Gattass et. al., 1988; Perry et. al., 1984]
[Allman & Kaas, 1971; Allman & Kaas, 1974; Gattass et al., 1981; van Essen et al., 1984; Maguire & Baizer, 1984; Burkhalter & van Essen, 1986; Gattass et al., 1987; Desimone & Schein, 1987; Gattass et al., 1988; Cavanaugh et al., 2002]

[Freeman & Simoncelli, 2011]
Extended presentation

Directed attention

Scaling (diameter / eccentricity) of receptive fields in synthesis model

Proportion correct

$r^2 = 0.91, 0.89$

$r^2 = 0.94, 0.91$

$r^2 = 0.95, 0.85$

$r^2 = 0.97, 0.95$

Figure 4. Metamer control experiments. Each column shows data and fitted psychometric functions for an individual observer. Both experiments use stimuli generated by the mid-ventral model. (a) Metamer experiment with extended presentation time. Dark gray points: 200 ms presentation time (replotted from Fig. 3). Light gray points: 400 ms presentation time. Shaded region: 68% confidence interval obtained using bootstrapping. Gray horizontal line: chance performance. (b) Metamer experiment with directed attention. Dark gray points: undirected attention (replotted from Fig. 3). Light gray points: subjects were directed with an attentional cue indicating the region with the largest change (see Methods). Shaded region: same as panel a. [Freeman & Simoncelli, 2011]
Scaling (diameter / eccentricity) of receptive fields in synthesis model

Figure 3. Metamer experiment. Each panel shows, for an individual observer, the proportion of correct responses in the ABX task, as a function of the scaling parameter (ratio of receptive field diameter to eccentricity) of the model used to generate the stimuli. Dark gray points: mid-ventral model (see Fig. 2). Light gray points: V1 model (see Supplementary Fig. 2). Shaded region, 68% confidence interval obtained using bootstrapping. Gray horizontal line: Chance performance. Black lines: Performance of observer model with best fitting critical scaling and gain parameters (see Methods and Results). $r^2$ values for the fits indicated at the bottom of each plot.

[Freeman & Simoncelli, 2011]
“V2 model”  ●  Main experiment  ●  Extended presentation  ●  Directed attention

“V1 model”  ●

Scaling (diameter / eccentricity)

V4

V2

V1

S1  S2  S3  S4  Average

[Allman & Kaas, 1971; Allman & Kaas, 1974; Gattass et.al., 1981; van Essen et.al., 1984; Maguire & Baizer, 1984; Burkhalter & van Essen, 1986; Gattass et.al., 1987; Desimone & Schein, 1987; Gattass et.al., 1988; Cavanaugh et. al., 2002]

[Freeman & Simoncelli, 2011]
Reading

Figure 7. Effects of crowding on reading and searching. (a) Two metamers, matched to the model responses of a page of text from the first paragraph of Herman Melville's "Moby Dick". Each metamer was synthesized using a different foveal location (the letter above each red dot). These locations are separated by the distance readers typically traverse between fixations. In each metamer, the central word is largely preserved; farther in the periphery the text is letter-like but scrambled, as if printed with non-latin characters. Note that the boundary of readability in the first image roughly coincides with the location of the fixation in the second image. We emphasize that these are samples drawn from the set of images that are perceptually metameric; although they illustrate the kinds of distortions that result from the model, no single example represents "what an observer sees" in the periphery. (b) The notoriously hard-to-find "Waldo" (character with the red and white striped shirt) blends into the distracting background, and is only recognizable when we (or the model) look right at him. Cross-hairs surrounding each image indicate the location of the model fovea. (c) A soldier in Afghanistan wears sandy-stone patterned clothing to match the stoney texture of the street, and similarly blends into the background.

[Freeman & Simoncelli, 2011]
Camouflage

Figure 7. Effects of crowding on reading and searching.

(a) Two metamers, matched to the model responses of a page of text from the first paragraph of Herman Melville’s “Moby Dick”. Each metamer was synthesized using a different foveal location (the letter above each red dot). These locations are separated by the distance readers typically traverse between fixations. In each metamer, the central word is largely preserved; farther in the periphery the text is letter-like but scrambled, as if printed with non-latin characters. Note that the boundary of readability in the first image roughly coincides with the location of the fixation in the second image. We emphasize that these are samples drawn from the set of images that are perceptually metameric; although they illustrate the kinds of distortions that result from the model, no single example represents “what an observer sees” in the periphery.

(b) The notoriously hard-to-find “Waldo” (character with the red and white striped shirt) blends into the distracting background, and is only recognizable when we (or the model) look right at him. Cross-hairs surrounding each image indicate the location of the model fovea.

(c) A soldier in Afghanistan wears sandy-stone patterned clothing to match the stoney texture of the street, and similarly blends into the background.

[Freeman & Simoncelli, 2011]
Can we drive individual V2 neurons using local texture stimuli?

Top: synthetic textures, full model
Bottom: “spectral noise” (matched only for “V1” statistics)

[Freeman, et. al. 2013]
[Freeman, et. al. 2013]
15% of V1 neurons significantly positively modulated

63% of V2 neurons significantly positively modulated

[Freeman, et. al. 2013]
Texture category

Firing Rate (ips)

0.23

-0.03

-0.05

0.05

0.23

[Freeman, et. al. 2013]
We use the term “modulation” to capture the differential responses to textures and noise, and index its magnitude by taking the difference of responses divided by the sum (Fig. 2.4c). The average modulation index of neurons in V1 was near zero for most of the response time course, except for a modest late positive modulation (Fig. 2.4c). Neurons in V2 showed a substantial diversity of modulation across families in V2. (a) Firing rates for three single units in V1 (green) and V2 (blue) to naturalistic (dark dots) and noise (light dots), separately for the 15 texture families. Families are sorted according to the ranking in panel b. Gray bars connecting points are only for visualization of the differential response. Modulation indices (averaged across texture families) are reported in the upper right of each panel. Error bars indicate s.e.m. across the 15 samples of each texture family. (b) Diversity in modulation across texture families, averaged across all neurons. Error bars indicate s.e.m. across neurons. Gray bar indicates 2.5th and 97.5th percentiles of the null distribution of modulation expected due to chance.
Subject 1, Right hemisphere

Texture (9 sec)

Subject 2, Right hemisphere

Spectral Noise (9 sec)

[Freeman, et. al. 2013]
Predicting discriminability

Different families

[Ziemia, Freeman, Movshon, Simoncelli - unpublished]
Anesthetized macaque
  • V1: 102 neurons
  • V2: 103 neurons

Stimuli presented for 100ms
within a 4° aperture

20 repetitions each

[Ziemba, Freeman, Movshon, Simoncelli - unpublished]
Anesthetized macaque
  • V1: 102 neurons
  • V2: 103 neurons

Stimuli presented for 100ms
within a 4° aperture

20 repetitions each

[Example V1 neuron]

[Ziemba, Freeman, Movshon, Simoncelli - unpublished]
Anesthetized macaque
- V1: 102 neurons
- V2: 103 neurons

Stimuli presented for 100ms within a 4° aperture

20 repetitions each

[Ziembta, Freeman, Movshon, Simoncelli - unpublished]
Decoding

Family classification

![Graph showing the relationship between V1 and V2 performance with different values of n (1, 3, 10, 30, 100). The graph includes error bars and a trend line.]
Decoding

Family classification

Exemplar identification

V1 performance vs. V2 performance graph with data points and a linear trend line.

$n = 1, 3, 10, 30, 100$

Chance
The power of synthesis tests...

Intriguing properties of neural networks, ArXiv 2014
Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus
The power of synthesis tests...

Shows that, at the very least, these networks are NOT good models for human vision!

Intriguing properties of neural networks, ArXiv 2014
Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus