Nonlinear Image Representation with Cascaded Local Gain Control

Eero Simoncelli
Center for Neural Science, and Courant Inst. of Mathematical Sciences, New York University
• How do populations of neurons extract/represent visual information?
• In what ways is this matched to, or optimized for, our visual environment?
• How do these representations enable/limit perception?
• What new principles may be gleaned from these representations, and applied to engineered imaging or vision systems?
Retina
Optic Nerve
LGN
Optic Tract
Visual Cortex

[Hegde & van Essen, 2000
Ito & Komatsu, 2004
Anzai et.al., 2007
etc]

V2?
V1

[stimulus] → [response]

V1

V2?

Simple cell
Complex cell
Canonical functional models for sensory neurons

- "Unreasonably effective" [after Wigner, 1960]
- Evolving...
Temporally adaptive gain control

[Image: Diagram showing the relationship between luminance and time, with luminance gain control and contrast gain control depicted.]

[Supplementary Fig. 1 from Mante et al. 2005]
Local spatial gain control

- Stimuli are gratings varying in diameter and contrast.

- Curves show predictions of size tuning and contrast saturation.

- Saturation monotonically decreases with contrast (Kremers et al., 2001; Solomon et al., 2004).

- Local spatial gain control is illustrated in the diagram.

- Gain control is observed in ON-center cells more prominently than in OFF-center cells.

- The model captures this effect with predicted preferred tuning from 3.8.

- Decreases with contrast (Kremers et al., 2001; Solomon et al., 2004).

- Stimuli centered on the receptive field based on responses to 10% contrast and 37.

- Responses are smaller than the peak responses (mean responses to the largest stimulus tested, as a fraction of the peak responses).

- Local spatial gain control is ineffectual, and size tuning is not close to 1 (Derrington and Lennie, 1982; Shapley and Victor, 1978).

- The suppressive field is ineffectual, and size tuning is not close to 1 (Derrington and Lennie, 1982; Shapley and Victor, 1978).

- This expression is known to capture the contrast–response curve at each stimulus diameter. Power-law fitting is used.

- Saturation is pronounced in ON-center cells than in OFF-center cells, as has been reported to occur even during paralysis. When such drifts occurred, we estimated accuracy from the unsigned magnitude of the first harmonic. Most units (28 of 34) were of the X type.

- Cells had receptive fields with eccentricities ranging from 2° to 45°, with the shift occurring 1% since the previous sequence iteration. The result of fitting one experiment was held fixed in fits to subsequent experiments.

- We repeated this sequence of fits until parametric estimates obtained when fitting one experiment were held fixed in fits to subsequent experiments. We repeated this sequence of fits until parametric estimates obtained when fitting one experiment were held fixed in fits to subsequent experiments.

- Most units (28 of 34) were of the X type.

- The sequence of ocular preference changes along the penetration was accounted for slow changes in neural responsiveness and spontaneous activity.

- The expression is known to capture the fact that gain control is more pronounced in ON-center cells than in OFF-center cells, as has been reported to occur even during paralysis. When such drifts occurred, we estimated accuracy from the unsigned magnitude of the first harmonic. Most units (28 of 34) were of the X type.

- The sequence of ocular preference changes along the penetration was accounted for slow changes in neural responsiveness and spontaneous activity.
V1: Surround suppression

![Graph showing surround suppression with response in impulses per second on the y-axis and grating patch diameter in degrees on the x-axis. The graph includes data points and error bars, highlighting optimal and surround diameters with a suppression range.](image)
Surround contrast

- 0
- 0.03
- 0.06
- 0.13
- 0.25
- 0.5

Response (imp/sec)

Contrast

[Cavanaugh et al. 02]
V1: Cross-orientation suppression

[Carandini et al. 1997]
V1 Normalization Model

The linear model of simple cells

Retinal image

Firing rate

The normalization model of simple cells

Retinal image

Other cortical cells

Firing rate

[Carandini, Heeger, and Movshon, 1996; Carandini & Heeger, 2012]
Example: Area MT

Retinal image

Linear operator  Gain control  Output nonlinearity

V1

Linear operator  Gain control  Output nonlinearity

MT

[Simoncelli & Heeger, 1998]
Sparse marginal statistics

[Burt & Adelson 82; Field 87; Mallat 89; etc]
Independent Components Analysis (ICA)

For linearly-transformed-factorial sources:
guaranteed independence
(with some minor caveats)

[Cardoso 89; Jutten & Herault 91; Comon 94; Bell & Sejnowski 96; etc]
ICA on image blocks

[Bell/Sejnowski ’97; see also Olshausen/Field ’96]
Linearly-transformed factorial model

Coefficient density:

- X
- X
- X

Basis set:

- Image:
Marginal Gaussianization

\[ p(y) \]

\[ p(x) \]

[Chen & Gopinath 01]
Indications that the model is weak...

Sample from model

Image, ICA-transformed and marginally Gaussianized
Subbands are *heteroskedastic* (they have variable variance):

We can model this behavior using a Gaussian scale mixture (GSM):

[Wainwright & Simoncelli 2000]
GSM

Model generalized coefficient neighborhood as a Gaussian scale mixture (GSM) [Andrews&Mallows ‘74]:

\[ \vec{x} = \sqrt{z} \vec{u} \]

- \( \vec{u} \) is Gaussian, \( z > 0 \)
- \( z \) and \( \vec{u} \) are independent
- \( \vec{x} \) is elliptically symmetric, with covariance \( zC_u \)
- marginals of \( \vec{x} \) are leptokurtotic

[Wainwright&Simoncelli, ‘99]
joint histogram of natural image band-pass filter responses with separation of 2 samples
Joint statistics - sound

joint histogram of natural audio signal
gammatone filter responses with separation of 0.1 msec
non-Gaussian elliptical models of natural images:

- Simoncelli, 1997;
- Zetzsche & Krieger, 1999;
- Huang & Mumford, 1999;
- Wainwright & Simoncelli, 2000;
- Hyvärinen and Hoyer, 2000;
- Parra et al., 2001;
- Srivastava et al., 2002;
- Sendur & Selesnick, 2002;
- Teh et al., 2003;
- Gehler and Welling, 2006
- Lyu & Simoncelli, 2008
- etc.
• Density is elliptical, but *not* Gaussian

• Whitening makes spherical, but not independent!
radial Gaussianization (RG)

Gaussianize the radial component of the density

Approximate version: estimate local L2 norm, and divide (i.e., local gain control)

[Lyu & Simoncelli, 2008]
... all paths lead to a spherical/factorial Gaussian

[Lyu & Simoncelli, 2008]
Densities and their factorizations

[Lyu & Simoncelli, 2008]
RG vs. ICA on coefficient pairs

RG eliminates most dependency for *nearby* coeffs
ICA offers minimal advantage over PCA
Similar behaviors for coefficient blocks

[Lyu & Simoncelli, 2008]
Joint densities

- Nearby: densities are approximately circular/elliptical
- Distant: densities are approximately factorial

[Simoncelli, ‘97; Wainwright&Simoncelli, ‘99]
How do we build a global model that captures the full range of observed statistical behaviors?

1) Random Field of Gaussian Scale Mixtures
   [Lyu & Simoncelli, 2008]

2) Build an implicit model, using local gain control. I’ll show three recent examples...
Example 1: Density estimation

[Ballé, Laparra, Simoncelli, ICLR-16]
Density estimation (parametric density)

\[ p_x(x) = \frac{1}{Z(\theta)} \exp\left(-f(x; \theta)\right) \]

\[ Z(\theta) = \int \exp\left(-f(x; \theta)\right) \, dx \]

tractable?

[Balle, Laparra, Simoncelli, ICLR-16]
Density estimation (parametric transformation)

$x \sim p_x$

Gaussianization

$g(x; \theta)$

$y \sim \mathcal{N}$

"inferred" density:

$$p_x(x) = \left| \frac{\partial g(x; \theta)}{\partial x} \right| \mathcal{N}(g(x; \theta))$$

Friedman, 1984
Chen & Gopinath, 2001
Lyu & Simoncelli, 2009
Laparra et al., 2010
Dinh et al., 2015

[Balle, Laparra, Simoncelli, ICLR-16]
Parameter estimation

\[ x \rightarrow g(x; \theta) \rightarrow y \]

\[- \log p_x(x) = \frac{\partial g(x; \theta)}{\partial x} \log \left( \frac{1}{\sqrt{2\pi\|\theta\|^2}} \exp \left( -\frac{1}{2\|\theta\|^2} \|g(x; \theta) - x\|^2 \right) \right) C \]

minimize wrt. \( \theta \) using stochastic gradient descent

[Balle, Laparra, Simoncelli, ICLR-16]
Marginal distribution of linear filter responses

Burt & Adelson, 1981
Field, 1987
Mallat, 1989

[Balle, Laparra, Simoncelli, ICLR-16]
\[ y = \frac{x}{(\beta + \gamma |x|)^\varepsilon} \]

\[ p_x(x) = \frac{\partial y}{\partial x} \mathcal{N}(y) \]

[Balle, Laparra, Simoncelli, ICLR-16]
\[
\begin{align*}
  y_0 &= \frac{x_0}{\left(\beta_0 + \gamma_0 |x_0|^{\alpha_0} \right)^{\varepsilon_0} + \gamma_0 |x_0|^{\alpha_0}}^{\varepsilon_0} \\
  y_1 &= \frac{x_1}{\left(\beta_1 + \gamma_1 |x_1|^{\alpha_1} \right)^{\varepsilon_1} + \gamma_1 |x_1|^{\alpha_1}}^{\varepsilon_1}
\end{align*}
\]

[Balle, Laparra, Simoncelli, ICLR-16]
Contour lines, linear filter responses

[Balle, Laparra, Simoncelli, ICLR-16]
Figure 1: Mutual information in pairs of wavelet coefficients after various transformations, plotted as a function of the spatial separation between the coefficients.

Figure 2: Contour plots of pairwise wavelet coefficient densities. Each row corresponds to a model arising from a different transformation (ICA-MG, RG, GDN). Each column corresponds to a pair of coefficients spatially separated by distance $d$ (pixels). Gray: contour lines of histogram density estimate. Black: contour lines of densities induced by best-fitting transformations. As distance increases, the empirical density between the coefficients transitions from elliptical but correlated to separable. The RG density captures the former, and the ICA density captures the latter. Only the GDN density has sufficient flexibility to capture the full range of behaviors.

[Balle, Laparra, Simoncelli, ICLR-16]
Generalized divisive normalization (GDN)

\[ y_i = \frac{z_i}{(\beta_i + \sum_j \gamma_{ij}|z_j|^{\alpha_{ij}})^{\varepsilon_i}} \]

Special cases/related models:
- Independent Component Analysis, Cardoso, 2003
- Independent Subspace Analysis, Hyvärinen & Hoyer, 2000
- Weighted normalization model, Schwartz & Simoncelli, 2001
- Topographic ICA, Hyvärinen et al., 2001
- Radial Gaussianization, Lyu & Simoncelli, 2009
- \( L_p \)-nested symmetric distributions, Sinz & Bethge, 2010
- “Two-layer model”, Köster & Hyvärinen, 2010

[Balle, Laparra, Simoncelli, ICLR-16]
Parameter estimation (multiple layers)

\[ - \log p_x(x) = - \log \left| \frac{\partial g(x; \theta)}{\partial x} \right| - \frac{1}{2} \|g(x; \theta)\|_2^2 + C \]

\[ - \log \left| \frac{\partial g_0(x_0; \theta)}{\partial x_0} \right| - \log \left| \frac{\partial g_1(x_1; \theta)}{\partial x_1} \right| - \ldots \]

minimize wrt. \( \theta \) using stochastic gradient descent

[Balle, Laparra, Simoncelli, ICLR-16]
One layer of joint GDN > many layers of marginal GDN

[marginal GDN]

1 layer 2 layers 3 layers 4 layers

average likelihood [bit/px]

3.8 4.1 4.4 4.7

[joint GDN]

1 layer

[Balle, Laparra, Simoncelli, ICLR-16]
Example 2: Perceptually-optimized rendering

[Laparra, Ballé, Berardino & Simoncelli, in preparation]
Perceptual error is not consistent with Mean Squared Error

Equal MSE:
Why does MSE fail?

Human visual system constructs a nonlinear visual representation, and is sensitive to distortions in this space.
Simple retinal gain-control model:

Multiscale version: Normalized Laplacian Pyramid (NLP)

[Laparra, et. al. 2016]
NLP Distortion metric

\[ I \xrightarrow{\text{NLP}} y^{(0)} \xrightarrow{(\cdot)^2} \sum \xrightarrow{(\cdot)^{\frac{\beta}{2}}} \]

\[ I_r \xrightarrow{\text{NLP}} y^{(1)} \xrightarrow{(\cdot)^2} \sum \xrightarrow{(\cdot)^{\frac{\beta}{2}}} \]

\[ \vdots \]

\[ \vdots \]

\[ I \xrightarrow{(\cdot)^2} \sum \xrightarrow{(\cdot)^{\frac{\beta}{2}}} + \xrightarrow{\left(\cdot\right)^{\frac{1}{\beta}}} D(I, I_r) \]

[Laparra, et. al. 2016]
TID2008 database
[Ponomarenko et al., 2009]

MSE in retinal model response space explains perceptual data

[Laparra et al, 2016]
The image rendering problem

When we take a picture of a scene in the real world ($I$) we have a perception of this scene in our mind ($f(I)$). If the screen was able to reproduce the real scene exactly (i.e. $I_r = I$) our perception would be the same (i.e. $f(I) = f(I_r)$). However due to different restrictions actual screens are far from being able to reproduce real world scenes, then when we render this picture in a screen ($I_r$) the perception that we get is different from what we saw in the original scene (i.e. $f(I) \neq f(I_r)$). For instance while the range of luminances in the real world could go from almost $0$ cd/m$^2$ (dark room) to around $10^9$ cd/m$^2$ (sun light at noon), a regular display can represent only from $1$ to $200$ cd/m$^2$. We are aiming to find an image (that can be reproduced in the display) that produces a perception as similar as possible to the original scene, i.e we want to minimize the distance between $f(I)$ and $f(I_r)$ with the restriction that $I_r$ has to be rendered in a screen.

where $I$ is a source image, $I_r$ is the image that will be rendered, $D(\cdot,\cdot)$ is a metric for measuring perceptual distance, and $R$ is the set of images that can be rendered. Note that this formulation is general for any rendering problem where the only thing to be changed is the set of possible images to be rendered, and this set is restricted by the display restrictions. The perceptual distance metric we use in the experiments is presented in the next section. This metric distance is differentiable in closed form with respect to the image to be rendered $I_r$, and it allows us to reframe the problem stated above as an optimization problem. It is important to stress that we do not need an extra function to convert $I$ to $I_r$. This is the main characteristic of our proposal.

In order to minimize the functional we perform an alternating projection procedure where we alternate a gradient descent step to minimize the perceptual distance and a projection on the constraints. During gradient descent, we employ the ADAM Regularization method [11]. The derivative of the perceptual distance with respect to the image to be rendered $I_r$ is described in the appendix B.

2.1 Perceptual distance

To quantify perceptual distance, we transform images into a perceptual space using a crude model for the early visual system (retina and lateral geniculate nucleus) and then compute the distance between the transformed images. In other words we define a transformation $f$ such that $f(I)$ and $f(I_r)$ are representations of the human visual system (HVS) perception of images $I$ and $I_r$ respectively. We then compute the distance between $f(I)$ and $f(I_r)$. If we were able to recreate the original scene exactly, our perceptual representation would be the same $f(I) = f(I_r)$. However due to display constraints, it is usually not possible to recreate the original image exactly, i.e. $I_r \neq I$.

We have previously proposed a function, which we call the Normalized Laplacian Pyramid (NLP), that mimics the functions and nonlinearities of the early Human Visual System (Retina and LGN). We showed that distances measured between two images in this domain correlate very highly with human opinions [12]. Here we extend the model to deal directly with luminances (cd/m$^2$). This allows us to work with the same units when setting constraints on acquisition and display. All parameters of this model were optimized to best explain human perceptual ratings of distorted images in a public database [13], and are fixed for all results presented. Figure 2 summarizes the steps performed in the transformation.
Automatic high dynamic range image rendering

original  Paris et. al., 2015  Laparra et. al. (in preparation)
Perceptually-optimized image rendering

L\(_{\text{max}}\) = 10\(^3\)

original

Paris et. al., 2015

L\(_{\text{max}}\) = 10\(^4\)

L\(_{\text{max}}\) = 10\(^5\)
Dithering (binary rendering)

original
(grayscale)

standard
(Floyd-Steinberg 1976)

Laparra et. al.
(in preparation)
Example 3: Compression

[Ballé, Laparra & Simoncelli, PCS-16+]
End-to-end rate-distortion optimization

\[ L[g_a, g_s] = H[P_q] + \lambda \mathbb{E} \| z - \hat{z} \| \]

relaxation to differential entropy:

\[ p_{\hat{y}_i}(t) = \sum_{n=-\infty}^{\infty} P_{q_i}(n) \delta(t - n), \]

\[ P_{q_i}(n) = (p_{y_i} \ast U(0,1))(n), \text{ for all } n \in \mathbb{Z}, \]

[Balle et al., PCS-16]
3-stage GDN cascade

**GDN:**

\[
y_i = \frac{x_i}{(\beta_i + \sum_j \gamma_{ij} x_j^2)^{1/2}}
\]

\(\gamma_{ij}\) symmetric

**IGDN (approximate):**

\[
x_i = y_i \cdot (\beta_i + \sum_j \gamma_{ij} y_j^2)^{1/2}
\]

(one step of fixed-point inverse)

[Balle et al., in preparation]
[Balle et al., in preparation]
original (cropped)

jpeg: 9851 bytes, RMSE: 18.84

jpeg2000: 8127 bytes, RMSE: 18.37

3-stage GDN: 8115 bytes, RMSE: 15.95

[Balle et al., in preparation]
[Balle et al., in preparation]
Local gain control...

- is found throughout biological sensory systems
- can be implemented as an invertible nonlinear transform
- can Gaussianize natural signals, eliminating dependencies
- can mimic human perception of visual distortions
- can be used, cascaded, for image compression
- but we need a more complete characterization/design toolbox!
Thanks

Johannes Ballé
Valero Laparra
Alex Berardino
Siwei Lyu