Modeling Visual Cortex V4 in Naturalistic Conditions with Invariant and Sparse Image Representations

Bin Yu
Departments of Statistics and EECS
University of California at Berkeley

Rutgers University, May 2, 2014
Co-authors

Yu Group: Julien Mairal and Yuval Benjemani (leads)

Gallant Lab at UC Berkeley: Ben Willmore, Michael Oliver, Jack Gallant

Supported by NSF STC, Center for Science of Information (CSoI)
Brain Science 2013 – new "genomics"
Macaque ventral stream visual areas

- V1
- V2
- V4
- TEO/PIT
- IT

Orientation markers:
- anterior/rostral
- dorsal
- ventral
Receptive field complexity across areas

<table>
<thead>
<tr>
<th>V2</th>
<th>V4</th>
<th>posterior IT</th>
<th>anterior IT</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="V2" /></td>
<td><img src="image2" alt="V4" /></td>
<td><img src="image3" alt="posterior IT" /></td>
<td><img src="image4" alt="anterior IT" /></td>
</tr>
</tbody>
</table>

Kobatake & Tanaka, 1994

inak, 1994
A contour curvature tuned V4 neuron

Pasupathy & Connor, 1999

Modeling Visual Cortex V4 in Naturalistic Conditions with Invariant and Sparse Image Representations
A bi-modal orientation tuned V4 neuron

![Graph showing spatial frequency vs orientation](image1)

**Highest rank**

**Lowest rank**

**C**

**D**

**E**

*cell v0102*

David, Hayden & Gallant, 2006
Sources of Knowledge About the Brain

Ways of Understanding the Visual Cortex

- study of lesions and associated impairments;
- electrodes (single or arrays);
- imaging studies (fMRI, ...); image below from Hansen et al. [2007]
Classical Models for V1 “Simple” Cells

Image from Olshausen and Field [2005]:

V1

- Models based on Gabor filters achieve impressive prediction performance with experimental data (single neuron and fMRI);
- V1 receptive fields are relatively small and well localized.

V1 is the most well understood area, but not all is ... It serves as a performance benchmark for other areas.
V4: an intermediate area on the "what" pathway
[see Roe et al., 2012]

What we know about V4
- affected by attention;
- diverse selectivity;
- larger receptive fields and is more invariant than V1/V2;
- no good predictive model with natural image inputs.

Question about V4
- what are the roles of V4? Roe et al advocated a background-foreground thesis, among other things.
Experimental set-up in the Gallant Lab for single neuron data collection

![Diagram of experimental set-up]

- Image Sequence
- Subject
- Recording
- Response

- Signal Filtering
- Spike Sorting
- Time Binning
- Firing Rate

Modeling Visual Cortex IV in Naturalistic Conditions with Sparse, Invariant Image Representations
Objectives and data

Aim: a statistical/computational model with
- good prediction performance on natural scenes (validation data);
- elucidation of properties of a population of V4 neurons;
- biologically interpretability.

Data
- consists of 4000 – 12000 grayscale images (with no motion or color content) and average firing rates for 71 neurons;
- the image sequence is shown at 30 Hz;
- the stimuli is centered on an estimated receptive field (RF) while the subject performs a fixation task;
- the stimuli size is 2 – 4 times larger than the estimated RF.
Outline of Today’s Talk

- Multi-layer invariant feature extraction
- Prediction model via low-rank regularization
- Model interpretation
Part I: Invariant Image Representation: Feature Extraction
Methodology

Classical computer vision image representation for scene analysis

1. dense low-level feature extraction (local histograms of gradient orientations) [Lowe, 2004];
2. feature encoding into visual words using vector quantization or sparse coding [Olshausen and Field, 1997];
3. feature pooling.

A state-of-the-art pipeline for scene and object recognition [Lazebnik et al., 2006, Yang et al., 2009, Boureau et al., 2010];

Can we exploit some ideas from this line of thinking to mimic invariance properties of V4 neurons?

be as biologically compatible as possible?
Methodology

Classical computer vision image representation for scene analysis

1. dense low-level feature extraction (local histograms of gradient orientations) [Lowe, 2004];
2. feature encoding into visual words using vector quantization or sparse coding [Olshausen and Field, 1997];
3. feature pooling.

A state-of-the-art pipeline for scene and object recognition [Lazebnik et al., 2006, Yang et al., 2009, Boureau et al., 2010];

Can we exploit some ideas from this line of thinking to

- mimic invariance properties of V4 neurons?
- obtain a model tuned to natural images?
- be as biologically compatible as possible?
Our pipeline

**Invariant Image Representation**

**Feature Pooling**
- Multiscale pooling
  - Number of pooling regions: between 1 and 19

**Computation of Feature Maps**
- Sparse coding
  - Number of feature maps: 2048
- Patch extraction and contrast normalization
  - Patch sizes: 4 x 4 x 8

**Computation of Orientation Maps**
- Convolution with Gaussian filter and subsampling
  - Orientation map sizes: 32 x 32
  - Number of orientation maps: 8
- Convolution with local oriented filters and rectification
  - Filter sizes: 7 x 7

**Input Image**
- Image size: 256 x 256
First Layer

**Computation of Orientation Maps**

- Convolution with Gaussian filter and subsampling
  - Orientation map sizes: 32 x 32
  - Number of orientation maps: 8

- Convolution with local oriented filters and rectification
  - Filter sizes: 7 x 7

**Input Image**

- Image size: 256 x 256
Patch Extraction Across Orientation Maps

- Patch extraction and contrast normalization
- Patch sizes: 4 x 4 x 8

Computation of Orientation Maps

- Convolution with Gaussian filter and subsampling
- Orientation map sizes: 32 x 32
- Number of orientation maps: 8

The 3D-patches

- are of size $4 \times 4 \times 8 = 128$ (8 orientations);
- correspond in the original image domain to $32 \times 32$ patches;
- are invariant to local deformations in the original image domain;
What is the effect of the first processing layer?
Comparing patches on orientation maps and in the image domain leads to a different similarity measure.

In blue, correlation in the original image domain.
In red, correlation in the new domain.
Constrast Normalization

Given a 3D-patch \( x \) in \( \mathbb{R}^{128} \), we apply \( x \leftarrow x / \max(\|x\|_2, c) \).

In short, our 3D-patches have similar properties as dense SIFT descriptors [Lowe, 2004] but are based on simple filtering/subsampling and normalization step.
Second Layer

Computation of Feature Maps

- Sparse coding

Number of feature maps: 2048

Patch extraction and contrast normalization

Patch sizes: 4 x 4 x 8

Computation of Orientation Maps
Sparse Coding

1. learn a dictionary on a database of 3D-patches once in for all;
2. encode all 3D-patches of an image to obtain feature maps.

Dictionary Learning Formulation

\[
\min_{A \in \mathbb{R}^{p \times n}, D \in \mathcal{D}} \sum_{i=1}^{n} \frac{1}{2} \| x^i - D \alpha^i \|_2^2 + \lambda \| \alpha^i \|_1,
\]

We also force the codes \( \alpha \) and the dictionary \( D \) to be non-negative.

The dictionary is fairly large \( (p = 2048) \).

The original formulation of Olshausen and Field [1997] was in the image domain. It was successfully used on image descriptors in computer vision [Yang et al., 2009, Boureau et al., 2010].
Third Layer

Invariant Image Representation

Feature Pooling
Multiscale pooling
Number of pooling regions: between 1 and 19

Computation of Feature Maps
Sparse coding
Number of feature maps: 2048

High-dimensional vector
Feature Pooling

\[
A = \begin{pmatrix}
\alpha^1 \\
\alpha^2 \\
\vdots \\
\alpha^{29*29}
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
\beta_1 \\
\vdots \\
\beta_p
\end{pmatrix}
\]

The pooling operation is the \( \ell_2 \)-norm of features \( \beta_k \triangleq \sqrt{\sum_{i=1}^{29*29} (\alpha_i^k)^2} \) for one pooling region and \( k = 1, \ldots, 2048 \).

Another alternative is the max-pooling operation [Riesenhuber et al., 1999, Cadieu et al., 2007], often used in computer vision [Lazebnik et al., 2006, Yang et al., 2009, Boureau et al., 2010].
Summary

Invariant Image Representation

High-dimensional vector

Feature Pooling

Multiscale pooling
Number of pooling regions: between 1 and 19

Computation of Feature Maps

Sparse coding
Number of feature maps: 2048

Patch extraction and contrast normalization
Patch sizes: 4 x 4 x 8

Computation of Orientation Maps

Convolution with Gaussian filter and subsampling
Orientation map sizes: 32 x 32
Number of orientation maps: 8

Convolution with local oriented filters and rectification
Filter sizes: 7 x 7

Input Image
Image size: 256 x 256
Our Feature Extraction Process in a Nutshell

It consists of local simple operations and uses the sparse coding principle. In particular, it

- has some invariance to small image deformation (first layer);
- has some selectivity to features learned from natural image statistics (second layer);
- is shift invariant within the receptive field (third layer);
Part II: Prediction Model based on Extracted Features
Prediction Pipeline

Input Image Sequence

Nonlinear Encoding

Neuron Responses

Linear Model
Temporal Aspect of Data

Typical Time Response to an Excitatory Stimulus

![Graph showing mean excitation over time]

- **Mean Excitation**
  - **vals** range from -0.002 to 0.004
  - **Lag (in frames)**:
    - t-9 to t-1

The graph illustrates the typical time response to an excitatory stimulus, highlighting the peak excitation point.
Prediction with Low-Rank Regularization

- **Input:** image sequence and neuron responses $y^1, y^2, \ldots, y^T$;
- **Preprocessing:** center and normalize the neuron responses;
- **Image feature vector:** $(2048*5)$-dimensional $\beta^t$ vector at time $t = 1, ..., T$;
- **Model:** $y^t \approx \sum_{j=1}^{\tau} \beta^{t-j} w^j$ with a lag $\tau = 9$ starting at time $t + 1$;
- **Constraint on the weights $w^j$:**

$$W = [w^1, w^2, \ldots, w^{\tau}] \in \mathbb{R}^{p \times \tau} \text{ should be low-rank.}$$

- **Formulation:**

$$\min_{W \in \mathbb{R}^{p \times \tau}} \sum_{t=1}^{T} \frac{1}{2} (y^t - \sum_{j=0}^{\tau} \beta^{t-j} w^j)^2 + \gamma \| W \|_*.$$

$\| \cdot \|_*$ is the trace norm ($\ell_1$-norm or sum of singular values) [see Fazel et al., 2001].
The baseline is a well engineered non-linear Gabor model [see, e.g., Nishimoto and Gallant, 2011] (state of the art for V1/V2 prediction).
Prediction Performance on Validation Data

![Histogram showing correlation coefficient distribution](image)
Prediction Performance on Validation Data

Example for a neuron with $\rho = 0.67$ on validation data
Prediction Performance:

- first model to achieve similar prediction performance on natural images as the ones achieved for V1 and V2 cells;
- significantly outperforms the Gabor model;
- 5-fold cross-validation leads to semi time-separable models.
Part III: Model Interpretation
Visualization of Dictionary Elements

Difficulty: 3D-patches are not in the image domain

We build a database of one million correspondence pairs of \((32 \times 32\) image patches, \(4 \times 4 \times 8\) 3D-patches) and find best matches.

(A) Complex Feature Visualization

(B) Feature Classification by Type

Others …

Straight line
Curve inward
Curve outward
Multiple curves
White bar
Black bar
Stripes
White corner
Black corner
Acute white corner
Acute black corner
White blob
Black blob
Double white blobs
Double black blobs
Complex blobs
Crosses and Junctions
Smooth Transition

1
2
3
4
5
6
7
8
9
10
11
12
Categorization of Dictionary Elements 1/2

(B) Feature Classification by Type

- Straight line
- Curve inward
- Curve outward
- Multiple curves
- White bar
- Black bar
- Stripes
- Smooth Transition
- White corner
- Black corner
- Acute white corner
- Acute black corner

Modeling Visual Cortex V4 in Naturalistic Conditions with Invariant and Sparse Image Representations
We also categorize each dictionary element by

- orientation;
- scale;
- texture affinity.
**Visualization of Feature Channels**

<table>
<thead>
<tr>
<th>curves/edges</th>
<th>bars</th>
<th>corners</th>
<th>blobs</th>
<th>image</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Modeling Visual Cortex V4 in Naturalistic Conditions with Invariant and Sparse Image Representations
Model Analysis

Methodology for Population Analysis

1. find the peak excitation lag;
2. look at the center pooling region;
3. compute 19 impact values: 8 feature type, 3 scales, 3 texture affinity, 5 orientation;
4. perform a sparse principal component analysis [Zou et al., 2006].

Methodology for Individual Neuron Analysis

- retrieve most excitatory and inhibitory images;
- compute impact values (possibly with refined categories).
Individual Neuron Analysis

neuron 64, $\rho = 0.64$
Individual Neuron Analysis

neuron 40, $\rho = 0.76$
Individual Neuron Analysis

neuron 71, $\rho = 0.75$
Individual Neuron Analysis

neuron 22, $\rho = 0.64$
Individual Neuron Analysis

neuron 24, $\rho = 0.63$
Individual Neuron Analysis

neuron 31, $\rho = 0.66$
Individual Neuron Analysis

neuron 59, $\rho = 0.67$
Population Analysis via Sparse PCA

Modeling Visual Cortex V4 in Naturalistic Conditions with Invariant and Sparse Image Representations
Population Analysis via Sparse PCA
Summary

Main conclusions

- first quantitative model with natural images that meet the benchmark performance;
- the most striking observation is the role of contours vs texture discrimination;
- V4 neurons are selective to a large diversity of features types such as bars, edges, corners...
- some V4 neurons are selective to orientation, some of them not.
Current Directions

- include time and color to deal with fully naturalistic conditions;
- using the new V4 model for movie reconstruction based on fMRI data;
- theoretical and simulation studies of deep learning methods;
- sparse coding: a key step in analyzing fruitfly TF images.


References II


SIFT Representation

[Lowe, 2004]

1. compute pixel orientations: \( \rho = \| \nabla I \|, \theta = \arctan(\nabla_y I / \nabla_x I) \).

2. binning + histograms.

3. normalization.

standard parameters: 8 orientations, 16 \( \times \) 16 patches, 4 \( \times \) 4 bins.