What can computational models tell us about face processing?

Garrison W. Cottrell
Gary's Unbelievable Research Unit (GURU)
Computer Science and Engineering Department
Temporal Dynamics of Learning Center
Institute for Neural Computation
UCSD

Collaborators, Past, Present and Future:
Ralph Adolphs, Luke Barrington, Serge Belongie, Kristin Branson, Tom Busey, Andy Calder, Matthew Dailey, Piotr Dollar, Michael Fleming, AfmZakaria Haque, Carrie Joyce, Brenden Lake, Tim Marks, Janet Metcalfe, Jonathan Nelson, Nam Nguyen, Curt Padgett, Honghao Shan, Maki Sugimoto, Matt Tong, Brian Tran, Keiji Yamada, Lingyun Zhang,
What can computational models tell us about face processing?

Garrison W. Cottrell
Gary's Unbelievable Research Unit (GURU)
Computer Science and Engineering Department
Institute for Neural Computation
UCSD

Collaborators, Past, Present and Future:
Ralph Adolphs, Luke Barrington, Serge Belongie, Kristin Branson, Tom Busey, Andy Calder, Matthew Dailey, Piotr Dollar, Michael Fleming, AfmZakaria Haque, Carrie Joyce, Brenden Lake, Tim Marks, Janet Metcalfe, Jonathan Nelson, Nam Nguyen, Curt Padgett, Honghao Shan, Maki Sugimoto, Matt Tong, Brian Tran, Keiji Yamada, Lingyun Zhang,
What can computational models tell us about face processing?

Garrison W. Cottrell
Gary's Unbelievable Research Unit (GURU)
Computer Science and Engineering Department
Institute for Neural Computation
UCSD

Collaborators, Past, Present and Future:
Ralph Adolphs, Luke Barrington, Serge Belongie, Kristin Branson, Tom Busey,
Andy Calder, Matthew Dailey, Piotr Dollar, Michael Fleming, AfmZakaria Haque, Carrie Joyce, Brenden Lake, Tim Marks, Janet Metcalfe, Jonathan Nelson, Nam Nguyen, Curt Padgett, Honghao Shan, Maki Sugimoto, Matt Tong, Brian Tran, Keiji Yamada, Lingyun Zhang,
What can computational models tell us about face processing?

Garrison W. Cottrell
Gary's Unbelievable Research Unit (GURU)
Computer Science and Engineering Department
Institute for Neural Computation
UCSD

Collaborators, Past, Present and Future:
What can computational models tell us about face processing?

Garrison W. Cottrell
Gary’s Unbelievable Research Unit (GURU)
Computer Science and Engineering Department
Institute for Neural Computation
UCSD

Collaborators, Past, Present and Future:
What can computational models tell us about face processing?

Garrison W. Cottrell
Gary's Unbelievable Research Unit (GURU)
Computer Science and Engineering Department
Institute for Neural Computation
UCSD

Collaborators, Past, Present and Future:
Ralph Adolphs, Luke Barrington, Serge Belongie, Kristin Branson, Tom Busey, Andy Calder, Matthew Dailey, Piotr Dollar, Michael Fleming, AfmZakaria Haque, Carrie Joyce, Brenden Lake, Tim Marks, Janet Metcalfe, Jonathan Nelson, Nam Nguyen, Curt Padgett, Honghao Shan, Maki Sugimoto, Matt Tong, Brian Tran, Keiji Yamada, Lingyun Zhang,
Why model?

• Models rush in where theories fear to tread.
• Models can be manipulated in ways people cannot.
• Models can be analyzed in ways people cannot.
Models rush in where theories fear to tread

- Theories are high level descriptions of the processes underlying behavior.
  - They are often not explicit about the processes involved.
  - They are difficult to reason about if no mechanisms are explicit -- they may be too high level to make explicit predictions.
  - Theory formation itself is difficult.
- Using machine learning techniques, one can often build a working model of a task for which we have no theories or algorithms (e.g., expression recognition).
- A working model provides an “intuition pump” for how things might work, especially if they are “neurally plausible” (e.g., development of face processing - Dailey and Cottrell).
- A working model may make unexpected predictions (e.g., the Interactive Activation Model and SLNT).
Models can be manipulated in ways people cannot

- We can see the effects of variations in cortical architecture (e.g., split (hemispheric) vs. non-split models (Shillcock and Monaghan word perception model)).
- We can see the effects of variations in processing resources (e.g., variations in number of hidden units in Plaut et al. models).
- We can see the effects of variations in environment (e.g., what if our parents were cans, cups or books instead of humans? I.e., is there something special about face expertise versus visual expertise in general? (Sugimoto and Cottrell, Joyce and Cottrell, Tong & Cottrell)).
- We can see variations in behavior due to different kinds of brain damage within a single “brain” (e.g. Juola and Plunkett, Hinton and Shallice).
Models can be analyzed in ways people cannot

In the following, I specifically refer to neural network models.

- We can do single unit recordings.
- We can selectively ablate and restore parts of the network, even down to the single unit level, to assess the contribution to processing.
- We can measure the individual connections -- e.g., the receptive and projective fields of a unit.
- We can measure responses at different layers of processing (e.g., which level accounts for a particular judgment: perceptual, object, or categorization? (Dailey et al. J Cog Neuro 2002).
How (I like) to build Cognitive Models

- I like to be able to relate them to the brain, so “neurally plausible” models are preferred -- neural nets.
- The model should be a *working model* of the *actual* task, rather than a cartoon version of it.
- Of course, the model should nevertheless be *simplifying* (i.e. it should be constrained to the essential features of the problem at hand):
  - Do we really need to model the (supposed) translation invariance and size invariance of biological perception?
    - As far as I can tell, NO!
  - Then, take the model “as is” and fit the experimental data: 0 fitting parameters is preferred over 1, 2, or 3.
The other way (I like) to build Cognitive Models

• Same as above, except:
• Use them as *exploratory* models -- in domains where there is little direct data (e.g. no single cell recordings in infants or undergraduates) to suggest what we might find if we *could* get the data. These can then serve as “intuition pumps.”
• Examples:
  • Why we might get specialized face processors
  • Why those face processors get recruited for other tasks
Today:
Talk #1: The Face of Fear:
A Neural Network Model of Human Facial Expression Recognition

Joint work with Matt Dailey
A Good Cognitive Model Should:

- Be *psychologically relevant* (i.e. it should be in an area with a lot of real, interesting psychological data).
- Actually be *implemented*.
- If possible, perform the *actual task* of interest rather than a cartoon version of it.
- Be *simplifying* (i.e. it should be constrained to the essential features of the problem at hand).
- *Fit* the experimental data.
- Make new *predictions* to guide psychological research.
The Issue: Are Similarity and Categorization Two Sides of the Same Coin?

• Some researchers believe perception of facial expressions is a new example of *categorical perception*:
  • Like the colors of a rainbow, the brain separates expressions into discrete categories, with:
    • Sharp boundaries between expressions, and...
    • Higher discrimination of faces near those boundaries.
Percent of subjects who pressed this button
The Issue: Are Similarity and Categorization Two Sides of the Same Coin?

- Some researchers believe the underlying representation of facial expressions is **NOT** discrete:
  - There are two (or three) underlying dimensions, e.g., intensity and valence (found by MDS).
  - Our perception of expressive faces induces a *similarity structure* that results in a circle in this space.
The Face Processing System

- Gabor Filtering
- PCA
- Happy
- Sad
- Afraid
- Angry
- Surprised
- Disgusted
- Neural Net

Levels:
- Pixel (Retina) Level
- Perceptual (V1) Level
- Object (IT) Level
- Category Level
The Face Processing System

Pixel (Retina) Level $\rightarrow$ Perceptual (V1) Level $\rightarrow$ Object (IT) Level $\rightarrow$ Category Level

Gabor Filtering $\rightarrow$ PCA

Bob, Carol, Ted, Alice

Neural Net
The Face Processing System

Gabor Filtering

PCA

Pixel (Retina) Level

Perceptual (V1) Level

Object (IT) Level

Feature level

Category Level

Bob
Carol
Ted
Cup
Can
Book
Neural Net
The Face Processing System

Pixel (Retina) Level → Gabor Filtering → Perceptual (V1) Level → LSF PCA HSF PCA → Object (IT) Level → Category Level

- Bob
- Carol
- Ted
- Cup
- Can
- Book
- Neural Net
The Gabor Filter Layer

- Basic feature: the 2-D Gabor wavelet filter (Daugman, 85):

- These model the processing in early visual areas

Subsample in a 29x36 grid
Principal Components Analysis

- The Gabor filters give us 40,600 numbers
- We use PCA to reduce this to 50 numbers
- PCA is like Factor Analysis: It finds the underlying directions of Maximum Variance
- PCA can be computed in a neural network through a competitive Hebbian learning mechanism
- Hence this is also a *biologically plausible* processing step
- We suggest this leads to representations similar to those in Inferior Temporal cortex
How to do PCA with a neural network

(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O'Toole et al. 1991)

- A self-organizing network that learns whole-object representations

(features, Principal Components, Holons, eigenfaces)

![Diagram of a neural network with Holons and Perceptual Layer](image.png)
How to do PCA with a neural network
(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O’Toole et al. 1991)

- A self-organizing network that learns whole-object representations
  *(features, Principal Components, Holons, eigenfaces)*

[Diagram of a neural network, showing layers labeled as Holons (Gestalt layer) and Input from Perceptual Layer]
How to do PCA with a neural network
(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O’Toole et al. 1991)

- A self-organizing network that learns whole-object representations
  *(features, Principal Components, Holons, eigenfaces)*

Holons
*(Gestalt layer)*

Input from Perceptual Layer
How to do PCA with a neural network
(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O'Toole et al. 1991)

- A self-organizing network that learns whole-object representations
  
  *(features, Principal Components, Holons, eigenfaces)*

![Diagram of a neural network with Holons (Gestalt layer) and Input from Perceptual Layer.](image)
How to do PCA with a neural network

(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O’Toole et al. 1991)

- A self-organizing network that learns whole-object representations

   \[\text{features, Principal Components, Holons, eigenfaces}\]
How to do PCA with a neural network
(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O’Toole et al. 1991)

- A self-organizing network that learns whole-object representations
  (features, Principal Components, Holons, eigenfaces)

Holons
(Gestalt layer)

Input from Perceptual Layer
How to do PCA with a neural network

(Cottrell, Munro & Zipser, 1987; Cottrell & Fleming 1990; Cottrell & Metcalfe 1990; O’Toole et al. 1991)

- A self-organizing network that learns whole-object representations

*(features, Principal Components, Holons, eigenfaces)*

Holons

*(Gestalt layer)*

Input from Perceptual Layer
Holons

- They act like face cells (Desimone, 1991):
  - Response of single units is strong despite occluding eyes, e.g.
  - Response drops off with rotation
  - Some fire to my dog’s face
- A novel representation: **Distributed templates** --
  - each unit’s optimal stimulus is a ghostly looking face (template-like),
  - but many units participate in the representation of a single face (distributed).
  - For this audience: Neither exemplars nor prototypes!
- Explain holistic processing:
  - *Why? If stimulated with a partial match, the firing represents votes for this template:*
    - Units “downstream” don’t know what caused this unit to fire.
  - *(more on this later…)*
The Final Layer: Classification


The holistic representation is then used as input to a categorization network trained by supervised learning.

Output: Cup, Can, Book, Greeble, Face, Bob, Carol, Ted, Happy, Sad, Afraid, etc.

- Excellent generalization performance demonstrates the sufficiency of the holistic representation for recognition
The Final Layer: Classification

- Categories can be at different levels: basic, subordinate.
- Simple learning rule (~delta rule). It says (mild lie here):
  - *add* inputs to your weights (synaptic strengths) when you are supposed to be *on*,
  - *subtract* them when you are supposed to be *off*.
- This makes your weights “look like” your favorite patterns – the ones that turn you on.
- When no hidden units => No back propagation of error.
- When hidden units: we get task-specific features (most interesting when we use the basic/subordinate distinction)
Outline

An overview of our facial expression recognition system.

- The internal representation shows the model’s prototypical representations of Fear, Sadness, etc.
- How our model accounts for the “categorical” data
- How our model accounts for the “two-dimensional” data
- Discussion
- Conclusions for part 1
Facial Expression Database

- Ekman and Friesen quantified muscle movements (Facial Actions) involved in prototypical portrayals of happiness, sadness, fear, anger, surprise, and disgust.
  - 70% agreement on emotional content by naive human subjects.
- 110 images, 14 subjects, 7 expressions.

Anger, Disgust, Neutral, Surprise, Happiness (twice), Fear, and Sadness

This is actor “JJ”: The easiest for humans (and our model) to classify
The Final Layer: Classification

- The final layer is trained based on the putative emotion being expressed by the actor (what the majority of Ekman’s subjects responded when given a 6-way forced choice)
- Uses a very simple learning rule: The delta rule
- It says:
  - *add* inputs to your weights (synaptic strengths) when you are supposed to be *on*,
  - *subtract* them when you are supposed to be *off*.
- This makes your weights “look like” your favorite patterns – the ones that turn you on.
- Also: no hidden units => No back propagation of error
### Results (Generalization)

<table>
<thead>
<tr>
<th>Expression</th>
<th>Network % Correct</th>
<th>Human % Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>100.0%</td>
<td>98.7%</td>
</tr>
<tr>
<td>Surprise</td>
<td>100.0%</td>
<td>92.4%</td>
</tr>
<tr>
<td>Disgust</td>
<td>100.0%</td>
<td>92.3%</td>
</tr>
<tr>
<td>Anger</td>
<td>89.2%</td>
<td>88.9%</td>
</tr>
<tr>
<td>Sadness</td>
<td>82.9%</td>
<td>89.2%</td>
</tr>
<tr>
<td>Fear</td>
<td>66.7%</td>
<td>87.7%</td>
</tr>
<tr>
<td>Average</td>
<td>89.9%</td>
<td>91.6%</td>
</tr>
</tbody>
</table>

- Kendall’s *tau* (rank order correlation): 0.667, p=0.0441
- Note: This is an *emergent property* of the model!
Correlation of Net/Human Errors

• Like all good Cognitive Scientists, we like our models to make the same mistakes people do!
• Networks and humans have a 6x6 confusion matrix for the stimulus set.
• This suggests looking at the off-diagonal terms: The errors
• Correlation of off-diagonal terms: $r = 0.567$. \[ F(1,28) = 13.3; p = 0.0011 \]
• Again, this correlation is an emergent property of the model: It was not told which expressions were confusing.
Outline

• An overview of our facial expression recognition system.
• The internal representation shows the model’s prototypical representations of Fear, Sadness, etc.
• How our model accounts for the “categorical” data
• How our model accounts for the “two-dimensional” data
• Discussion
• Conclusions for part 1
Examining the Net’s Representations

- We want to visualize “receptive fields” in the network.
- But the Gabor magnitude representation is noninvertible.
- We can *learn* an approximate inverse mapping, however.
- We used linear regression to find the best linear combination of Gabor magnitude principal components for each image pixel.
- Then projecting each unit’s *weight vector* into image space with the same mapping visualizes its “receptive field.”
Examining the Net’s Representations

- The “y-intercept” coefficient for each pixel is simply the average pixel value at that location over all faces, so subtracting the resulting “average face” shows more precisely what the units attend to:

- Apparently local features appear in the global templates.
Outline

• An overview of our facial expression recognition system.
• The internal representation shows the model’s prototypical representations of Fear, Sadness, etc.
• How our model accounts for the “categorical” data
• How our model accounts for the “two-dimensional” data
• Discussion
• Conclusions for part 1
Morph Transition Perception

- Morphs help psychologists study categorization behavior in humans
- Example: JJ Fear to Sadness morph:

Young et al. (1997) Megamix: presented images from morphs of all 6 emotions (15 sequences) to subjects in random order, task is 6-way forced choice button push.
Results: classical Categorical Perception: sharp boundaries...

iness  Surprise  Fear  Sadness  Disgust  Anger  Happ-

and higher discrimination of pairs of images when they cross a perceived category boundary
Results: Non-categorical RT’s

- “Scalloped” Reaction Times

![Graph showing reaction times for different emotions](image-url)
Results: More non-categorical effects

- Young et al. Also had subjects rate 1\textsuperscript{st}, 2\textsuperscript{nd}, and 3\textsuperscript{rd} most apparent emotion.

- At the 70/30 morph level, subjects were above chance at detecting mixed-in emotion. These data seem more consistent with \textit{continuous} theories of emotion.
Modeling Megamix

- 1 trained neural network = 1 human subject.
- 50 networks, 7 random examples of each expression for training, remainder for holdout.
- Identification = average of network outputs
- Response time = uncertainty of maximal output (1.0 - \(y_{\text{max}}\)).
- Mixed-in expression detection: record 1st, 2nd, 3rd largest outputs.
- Discrimination: 1 – correlation of layer representations
- We can then find the layer that best accounts for the data
Modeling Six-Way Forced Choice

- Overall correlation $r = .9416$, with NO FIT PARAMETERS!
The model fits the data best at a precategorical layer: the layer we call the "object" layer; NOT at the category level.
Discrimination

• Classically, one requirement for “categorical perception” is higher discrimination of two stimuli at a fixed distance apart when those two stimuli cross a category boundary.

• Indeed, Young et al. found in two kinds of tests that discrimination was highest at category boundaries.

• The result that we fit the data best at a layer before any categorization occurs is significant: In some sense, the category boundaries are “in the data,” or at least, in our representation of the data.
Outline

• An overview of our facial expression recognition system.
• The internal representation shows the model’s prototypical representations of Fear, Sadness, etc.
• How our model accounts for the “categorical” data
• How our model accounts for the “non-categorical” data
• Discussion
• Conclusions for part 1
Reaction Time: Human/Model

HUMAN SUBJECTS REACTION TIME

MODEL REACTION TIME (1 - MAX_OUTPUT)

Correlation between model & data: .6771, p<.001
Mix Detection in the Model

Can the network account for the continuous data as well as the categorical data? YES.
Human/Model Circumplexes

These are derived from similarities between images using non-metric Multi-dimensional scaling.
For humans: similarity is correlation between 6-way forced-choice button push.
For networks: similarity is correlation between 6-category output vectors.
Outline

• An overview of our facial expression recognition system.
• How our model accounts for the “categorical” data
• How our model accounts for the “two-dimensional” data
• The internal representation shows the model’s prototypical representations of Fear, Sadness, etc.

Discussion
• Conclusions for part 1
Discussion

• Our model of facial expression recognition:
  • Performs the same task people do
  • On the same stimuli
  • At about the same accuracy
• Without actually “feeling” anything, without any access to the surrounding culture, it nevertheless:
  • Organizes the faces in the same order around the circumplex
  • Correlates very highly with human responses.
  • Has about the same rank order difficulty in classifying the emotions
Discussion

• The discrimination correlates with human results most accurately at a *precategorization* layer: The discrimination improvement at category boundaries is in the representation of data, not based on the categories.

• These results suggest that for expression recognition, the notion of “categorical perception” simply is not necessary to explain the data.

• Indeed, most of the data can be explained by the interaction between the similarity of the representations and the categories imposed on the data: Fear faces are similar to surprise faces in our representation – so they are near each other in the circumplex.
Conclusions from this part of the talk

- The best models perform the same task people do.
- Concepts such as “similarity” and “categorization” need to be understood in terms of models that do these tasks.
- Our model simultaneously fits data supporting both categorical and continuous theories of emotion.
- The fits, we believe, are due to the interaction of the way the categories slice up the space of facial expressions, and the way facial expressions inherently resemble one another.
- It also suggests that the continuous theories are correct: “discrete categories” are not required to explain the data.
- We believe our results will easily generalize to other visual tasks, and other modalities.
Outline for the next two parts

• Why would a face area process BMW’s?:
  • Behavioral & brain data
  • Model of expertise
  • results

• Why is this model wrong, and what can we do to fix it?
Are you a perceptual expert?

Take the expertise test!!!

“Identify this object with the first name that comes to mind.”

**Courtesy of Jim Tanaka, University of Victoria**
“Car” - Not an expert

“2002 BMW Series 7” - Expert!
“Bird” or “Blue Bird” - Not an expert

“Indigo Bunting” - Expert!
“Face” or “Man” - Not an expert

“George Dubya”- Expert!

“Jerk” or “Megalomaniac” - Democrat!
How is an object to be named?

- **Animal**
  - **Superordinate Level**
- **Bird**
  - **Basic Level** (Rosch et al., 1971)
- **Indigo Bunting**
  - **Subordinate Species Level**
Entry Point Recognition

Animal
Semantic analysis

Bird
Visual analysis

Indigo Bunting
Fine grain visual analysis

Downward Shift Hypothesis

Entry Point
Dog and Bird Expert Study

- Each expert had at least 10 years experience in their respective domain of expertise.
- None of the participants were experts in both dogs and birds.
- Participants provided their own controls.

Tanaka & Taylor, 1991
Object Verification Task

Superordinate
- Animal
- Plant

Basic
- Bird
- Dog

Subordinate
- Robin
- Sparrow

YES         NO

YES         NO
Dog and bird experts recognize objects in their domain of expertise at subordinate levels.
Is face recognition a general form of perceptual expertise?

George W. Bush

Indigo Bunting

2002 Series 7 BMW
Face experts recognize faces at the individual level of unique identity

Tanaka, 2001
**Event-related Potentials and Expertise**

**Face Experts**

- Bentin, Allison, Puce, Perez & McCarthy, 1996

**Object Experts**


---

**Novice Domain**

**Expert Domain**
Neuroimaging of face, bird and car experts

Car Experts

Bird Experts

Gauthier et al., 2000
How to identify an expert?

Behavioral benchmarks of expertise
• Downward shift in entry point recognition
• Improved discrimination of novel exemplars from learned and related categories

Neurological benchmarks of expertise
• Enhancement of N170 ERP brain component
• Increased activation of fusiform gyrus
End of Tanaka Slides
• Kanwisher showed the FFA is specialized for faces
• But she forgot to control for what???
Greeble Experts (Gauthier et al. 1999)

- Subjects trained over many hours to recognize individual Greebles.
- Activation of the FFA increased for Greebles as the training proceeded.
The “visual expertise mystery”

• If the so-called “Fusiform Face Area” (FFA) is specialized for face processing, then why would it also be used for cars, birds, dogs, or Greebles?

• Our view: the FFA is an area associated with a process: fine level discrimination of homogeneous categories.

• But the question remains: why would an area that presumably starts as a face area get recruited for these other visual tasks? Surely, they don’t share features, do they?

Solving the mystery with models

• Main idea:
  • There are multiple visual areas that could compete to be the Greeble expert - “basic” level areas and the “expert” (FFA) area.
  • The expert area must use features that distinguish similar looking inputs -- that’s what makes it an expert
  • Perhaps these features will be useful for other fine-level discrimination tasks.

• We will create
  • Basic level models - trained to identify an object’s class
  • Expert level models - trained to identify individual objects.
  • Then we will put them in a race to become Greeble experts.
  • Then we can deconstruct the winner to see why they won.

• A network that can differentiate faces, books, cups and cans is a “basic level network.”

• A network that can also differentiate individuals within ONE class (faces, cups, cans OR books) is an “expert.”
Model

- Pretrain two groups of neural networks on different tasks.
- Compare the abilities to learn a new individual Greeble classification task.
Expertise begets expertise

Learning to individuate cups, cans, books, or faces first, leads to faster learning of Greebles (can’t try this with kids!!!).

The more expertise, the faster the learning of the new task!

Hence in a competition with the object area, FFA would win.

If our parents were cans, the FCA (Fusiform Can Area) would win.

Training Time on first task

- Learning to individuate cups, cans, books, or faces first, leads to faster learning of Greebles (can’t try this with kids!!!).
- The more expertise, the faster the learning of the new task!
- Hence in a competition with the object area, FFA would win.
- If our parents were cans, the FCA (Fusiform Can Area) would win.
Entry Level Shift: Subordinate RT decreases with training

\( rt = \text{uncertainty of response} = 1.0 - \max(output) \)
How do experts learn the task?

• Expert level networks must be *sensitive* to within-class variation:
  - Representations must **amplify** small differences
• Basic level networks must *ignore* within-class variation.
  - Representations should **reduce** differences
Observing hidden layer representations

- Principal Components Analysis on hidden unit activation:
  - PCA of hidden unit activations allows us to reduce the dimensionality (to 2) and plot representations.
  - We can then observe how tightly clustered stimuli are in a low-dimensional subspace.
- We expect basic level networks to separate classes, but not individuals.
- We expect expert networks to separate classes and individuals.
Subordinate level training magnifies small differences *within* object representations.
Greeble representations are spread out prior to Greeble Training

Basic

Face
Variability Decreases Learning Time

Greeble Variance Prior to Learning Greebles

\[ r = -0.834 \]
Examining the Net’s Representations

• We want to visualize “receptive fields” in the network.
• But the Gabor magnitude representation is noninvertible.
• We can *learn* an approximate inverse mapping, however.
• We used linear regression to find the best linear combination of Gabor magnitude principal components for each image pixel.
• Then projecting each hidden unit’s *weight vector* into image space with the same mapping visualizes its “receptive field.”
Two hidden unit receptive fields

**AFTER TRAINING AS A FACE EXPERT**

**AFTER FURTHER TRAINING ON GREEBLES**

**NOTE:** These are not face-specific!
Controlling for the number of classes

- We obtained 13 classes from hemera.com:
  - 10 of these are learned at the basic level.
  - 10 faces, each with 8 expressions, make the expert task.
  - 3 (lamps, ships, swords) are used for the novel expertise task.
Results: Pre-training

- New *initial* tasks of similar difficulty: In previous work, the basic level task was much easier.

- These are the learning curves for the 10 object classes and the 10 faces.
Results

- As before, experts still learned new expert level tasks faster.

Number of training epochs on faces or objects

Number of epochs
To learn swords
After learning faces
Or objects
Outline

• Why would a face area process BMW’s?
• Why this model is wrong...
background

- I have a model of face and object processing that accounts for a lot of data...
The Face and Object Processing System

- **Gabor Filtering**
- **PCA**
- **FFA**
- **LOC**

**Levels**
- **Pixel (Retina) Level**
- **Perceptual (V1) Level**
- **Gestalt Level**
- **Hidden Layer**
- **Category Level**

**Classifiers**
- **Expert Level Classifier**
- **Basic Level Classifier**
Effects accounted for by this model

- Categorical effects in facial expression recognition (Dailey et al., 2002)
- Similarity effects in facial expression recognition (ibid)
- Why fear is “hard” (ibid)
- Rank of difficulty in configural, featural, and inverted face discrimination (Zhang & Cottrell, 2006)
- Holistic processing of identity and expression, and the lack of interaction between the two (Cottrell et al. 2000)
- How a specialized face processor may develop under realistic developmental constraints (Dailey & Cottrell, 1999)
- How the FFA could be recruited for other areas of expertise (Sugimoto & Cottrell, 2001; Joyce & Cottrell, 2004, Tran et al., 2004; Tong et al, 2005)
- The other race advantage in visual search (Haque & Cottrell, 2005)
- Generalization gradients in expertise (Nguyen & Cottrell, 2005)
- Priming effects (face or character) on discrimination of ambiguous Chinese characters (McCleery et al, under review)
- Age of Acquisition Effects in face recognition (Lake & Cottrell, 2005; L&C, under review)
- Memory for faces (Dailey et al., 1998, 1999)
- Cultural effects in facial expression recognition (Dailey et al., in preparation)
I have a model of face and object processing that accounts for a lot of data...

But it is fundamentally wrong in an important way!
I have a model of face and object processing that accounts for a lot of data...

But it is fundamentally wrong in an important way!

It assumes that the whole face is processed to the same level of detail everywhere....
• I have a model of face and object processing that accounts for a lot of data...

• But it is fundamentally wrong in an important way!

• It assumes that the whole face is processed to the same level of detail everywhere....and that the face is always aligned with “the brain”
background

• Instead, people and animals *actively sample* the world around them with eye movements.

• The images our eyes take in “slosh around” on the retina like an MTV video.

• We only get *patches* of information from any fixation

• But somehow (see previous talk), we perceive a stable visual world.
background

• So, a model of visual processing needs to be dynamically integrating information over time.

• We need to decide where to look at any moment:
  • There are both bottom up (inherently interesting things)
  • and top down influences on this (task demands)
Questions for today

- What form should a model take in order to model eye movements without getting into muscle control?

- Do people actually gain information from saccades in face processing?

- How is the information integrated?

- How do features change over time?
answers

• What form should a model take in order to model eye movements without getting into muscle control?
  • Probabilistic

• Do people actually gain information from saccades?
  • Yes

• How is the information integrated?
  • In a Bayesian way

• How do features change over time?
  • Don’t know yet, but we hypothesize that the effective receptive fields grow over time
This (brief!) part of the talk

• What form should a model take in order to model eye movements without getting into muscle control?
  • Probabilistic

• Do people actually gain information from saccades? ✓
  • Yes

• How is the information integrated? ✓
  • In a Bayesian way: NIMBLE

• How do features change over time?
  • Don’t know yet, but we hypothesize that the effective receptive fields grow over time
Where to look?

• We have lots of ideas about this (but, obviously, none as good as Baldi’s surprise!), but for the purposes of this talk, we will just use bottom up salience:

  • Figure out where the image is “interesting,” and look there

  • “interesting” = where there are areas of high contour
Interest points created using Gabor filter variance
Given those fragments, how do we recognize people?

- The basic idea is that we simply store the fragments with labels:
  - Bob, Carol, Ted or Alice
  - In the study set, not in the study set

(Note that we have cast face recognition as a categorization problem.)

- When we look at a new face, we look at the labels of the most similar fragments, and vote, using kernel density estimation, comparing a Gaussian kernel and 1-nearest neighbor.

- This can be done easily in a (Naïve) Bayesian framework
A basic test: Face Recognition

- Following (Duchaine & Nakayama, 2005):
  - Images are displayed for 3 seconds -- which means about 10 fixations - which means about 10 fragments are stored, chosen at random using bottom-up salience
  - 30 lures
  - Normal human ROC area (A’) is from 0.9 to 1.0

- NIMBLE results:

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Face Recall NIMBLE</th>
<th>ROC area Sum of Likelihoods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian ($\sigma = 1$)</td>
<td>0.94±0.03</td>
<td>0.97±0.02</td>
</tr>
<tr>
<td>kNN ($k = 1$)</td>
<td>0.93±0.05</td>
<td>0.96±0.02</td>
</tr>
</tbody>
</table>
Multi-Class Memory Tasks

- An identification task tests NIMBLE on novel images from **multiple classes**
- Face identification:
  - 29 identities from the FERET dataset
  - Changes in expression and illumination

- Object identification:
  - 20 object classes from the COIL-100 dataset
  - Changes in orientation, scale and illumination
## Identification Results

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Face ID Accuracy (%)</th>
<th></th>
<th>Object ID Accuracy (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NIMBLE</td>
<td>Sum of Likelihoods</td>
<td></td>
</tr>
<tr>
<td>Gaussian ($\sigma = 1$)</td>
<td>85.6±2</td>
<td>72.2±2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kNN ($k = 1$)</td>
<td>89.2±.6</td>
<td>85.8±2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Kernel             |                          | NIMBLE   | Sum of Likelihoods     |          |
|--------------------|                          |          |                        |          |
| Gaussian ($\sigma = 10$) | 87±1                   | 73.7±2   |                        |          |
| kNN ($k = 1$)      | 92.7±.7                | 87±.4    |                        |          |

- [Belongie, Malik & Puzicha, 2002] report 97.6% accuracy on a 20-class, 4 training image set from COIL-100.
Do people really need more than one fixation to recognize a face?

- To answer this question, we ran a simple face recognition experiment:
- Subjects saw 32 faces, then were given an old/new recognition task with those 32 faces and 32 more.
- However, subjects were restricted to 1, 2, 3, or unlimited fixations on the face.
• Using a gaze-contingent display, we can limit the fixations on the face itself to 1, 2, 3, or unlimited fixations.

• The average face mask prevents subjects from acquiring information about the face from peripheral vision.
Fixations during training

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16 overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image of fixations 1" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image2.png" alt="Image of fixations 2" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image3.png" alt="Image of fixations 3" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image4.png" alt="Image of fixations 4+" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image of overall fixations" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fixations during testing

Fixation 1

Fixation 2

Fixation 3

Fixation 4

Overall
Overall Data

- Overall fixation maps during training
Overall Data

- Overall fixation maps during testing

1st Fixation  2nd Fixation  3rd Fixation  4+ Fixations  Overall
Results

- This says you gain the most recognition by making two fixations, and a third doesn’t help.
- Hence, if a picture is worth a thousand words, each fixation is worth about 500 words!
Long term questions

• Do people have visual routines for sampling from stimuli?
• Do people have different routines for different tasks on identical stimuli?
• Do the routines change depending on the information gained at any particular location?
• Do different people have different routines?
• How do those routines change with expertise?
• How do they change over the acquisition of expertise?