Atoms of Recognition in
Human and Computer Vision
Efficient use of limited information: recognizing local configurations

© Proceedings of the National Academy of Sciences. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/.

Minimizing variability

Useful for the interpretation of complex scenes

© Proceedings of the National Academy of Sciences. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/. Source: Ullman, Shimon, Liav Assif, Ethan Fetaya, and Daniel Harari. "Atoms of recognition in human and computer vision." Proceedings of the National Academy of Sciences 113, no. 10 (2016): 2744-2749.
• Useful for dealing with complex scenes but challenging: non-redundant images

• Human studies
• Computational models
• Implications: representation for recognition, brain processing, CBMM
• Dan Harari
• Liav Assif
• Guy Ben-Yossef
• Eitan Fetaya
• Leyla Isik
• Yena Han

• ERC Advanced Grant ‘Digital Baby’
Searching for Minimal Images

Over 15,000 subjects, laboratory controls
‘MIRC’ (Minimal Recognizable Configuration): all 5 descendants are unrecognizable

Sharp transition
Pairs

Parent – MIRC,
Child – ‘sub-MIRC’
Example:
Original images

Plane  Ship  House fly  Bald eagle  Horse

Bike  Car door  Human eye  Eyeglasses  Suit&tie

© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/.
# Visual Elements

<table>
<thead>
<tr>
<th>14</th>
<th>13</th>
<th>12</th>
<th>13</th>
<th>7</th>
<th>8</th>
<th>6</th>
<th>6</th>
<th>10</th>
</tr>
</thead>
</table>

| 24 | 26 | 19 | 30 | 29 | 25 | 20 | 18 | 21 |

© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [https://ocw.mit.edu/help/faq-fair-use/](https://ocw.mit.edu/help/faq-fair-use/).
Average 16.9 / class
Highly redundant
Each MIRC is non-redundant
• Sensitive tool to compare representations

• Differences between MIR Cs and sub-MIR Cs to infer visual features

• Recognition features not captured by human feed-forward models and computer vision representations
Testing computational models
• Training of object images,
• Testing on minimal images
• MIRCs and sub-MIRCs
Testing on:

• DPM Deformable Parts Models
• Bag-of-Words / VLAD (vector of locally aggregated descriptors)
• R-CNN (Deep Convolutional Neural Network) Malik
• Hmax -- model of recognition in the cortex
• Consistent winners of standardized recognition competitions
• (PASCAL, ImageNet)
Hmax Model

© Scholarpedia. License BY-NC-SA

Riesenhuber and Poggio, 1999
Serre et al 2007
Deep Network Models

Krizhevsky et al. [NIPS2012]

- Same model as LeCun’98 but:
  - Bigger model (8 layers)
  - More data (10^6 vs 10^3 images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)

- 7 hidden layers, 650,000 neurons, 60,000,000 parameters
- Trained on 2 GPUs for a week
‘Pascal Challenge’ Airplanes
The recognition gap is not reproduced
R-CNN ‘Deep-net’ Recognition Model
All Classifiers

Low accuracy for minimal images

Statistical significance

<table>
<thead>
<tr>
<th>H0: Binary gap equal to multi-class. H1: otherwise</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0: Human gap equal to Binary. H1: Human gap greater than Binary</td>
<td>0.2443</td>
</tr>
<tr>
<td>H0: Human gap equal to multi-class. H1: Human gap greater than multi-class</td>
<td>1.3290e-04</td>
</tr>
<tr>
<td>H0: Binary MIRC recall equal to sub-MIRC. H1: otherwise</td>
<td>4.5180e-06</td>
</tr>
<tr>
<td>H0: Human MIRC recall equal to sub-MIRC. H1: Human MIRC greater than sub-MIRC</td>
<td>0.1711</td>
</tr>
<tr>
<td>H0: AP full object equal to MIRC. H1: AP full object greater than MIRC</td>
<td>1.7502e-12</td>
</tr>
<tr>
<td>H0: AP full object equal to sub-MIRC. H1: AP full object greater than sub-MIRC</td>
<td>5.3537e-05</td>
</tr>
</tbody>
</table>

Data

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>MIRC</th>
<th>sub-MIRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human recall</td>
<td>NaN</td>
<td>0.81</td>
<td>0.10</td>
</tr>
<tr>
<td>Binary recall</td>
<td>NaN</td>
<td>0.84</td>
<td>0.67</td>
</tr>
<tr>
<td>Multiclass recall</td>
<td>NaN</td>
<td>0.74</td>
<td>0.63</td>
</tr>
<tr>
<td>AUC: AP</td>
<td>0.87</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC: AP (normalized)</td>
<td>0.84</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>AUC: ROC</td>
<td>0.99</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>EER: ROC</td>
<td>0.02</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>Gap (MIRC recall 50%-90%) – Binary: max, mean</td>
<td>NaN</td>
<td>0.38</td>
<td>0.27</td>
</tr>
<tr>
<td>Gap (MIRC recall 50%-90%) – Multi: max, mean</td>
<td>NaN</td>
<td>0.27</td>
<td>0.20</td>
</tr>
<tr>
<td>Num of negatives patches</td>
<td>245970.93</td>
<td>414795.17</td>
<td>414795.17</td>
</tr>
<tr>
<td># normalized negative images</td>
<td>2259.50</td>
<td>1428.78</td>
<td>1428.78</td>
</tr>
<tr>
<td>Num of positives</td>
<td>28.20</td>
<td>10.00</td>
<td>15.20</td>
</tr>
<tr>
<td>Gap s.d. (human, Binary, multi-class)</td>
<td>0.05</td>
<td>0.27</td>
<td>0.18</td>
</tr>
</tbody>
</table>
• Recognition of minimal images does not emerge by training any of the existing models tested.

• Representations used by existing models do not capture differences that human recognition is sensitive to.
Test 2: Train on Patches

© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/.
Example siblings

0.85  0.05
1.00  0.70  0.00  1.00  0.10
0.75  0.15
0.75  0.25  0.95  0.10  1.00  0.35
0.85  0.05
0.90  0.35  0.95  0.15  0.95
0.95  0.45
0.90  0.15  0.80  0.10  0.55  0.05
0.80  0.00
1.00  0.10  0.95  0.00  0.55  0.30

30 ‘siblings’ for each MIRC and sub-MIRC

0.79  0.06

© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/.
### Statistical Significance

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0: Binary gap equal to multi-class. H1: otherwise</td>
<td>0.0284</td>
</tr>
<tr>
<td>H0: Human gap equal to Binary. H1: Human gap greater than Binary</td>
<td>5.2497e-05</td>
</tr>
<tr>
<td>H0: Human gap equals multi-class. H1: Human gap greater than multi-class</td>
<td>1.9853e-06</td>
</tr>
<tr>
<td>H0: Binary MIRC recall equals sub-MIRC. H1: otherwise</td>
<td>0.0016</td>
</tr>
<tr>
<td>H0: Human MIRC recall equals sub-MIRC. H1: Human MIRC greater than sub-MIRC</td>
<td>6.9954e-10</td>
</tr>
</tbody>
</table>

### Data

<table>
<thead>
<tr>
<th></th>
<th>MIRC</th>
<th>sub-MIRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human recall</td>
<td>0.78</td>
<td>0.13</td>
</tr>
<tr>
<td>Binary recall</td>
<td>0.79</td>
<td>0.56</td>
</tr>
<tr>
<td>Multiclass recall</td>
<td>0.75</td>
<td>0.60</td>
</tr>
<tr>
<td>AUC. AP (normalized)</td>
<td>0.70</td>
<td>0.61</td>
</tr>
<tr>
<td>AUC: train-full AP (normalized)</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>AUC: ROC</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>EER: ROC</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Gap (MIRC recall 50%-90%) - Binary, max, mean</td>
<td>0.32</td>
<td>0.23</td>
</tr>
<tr>
<td>Gap (MIRC recall 50%-90%) - Multi, max, mean</td>
<td>0.22</td>
<td>0.17</td>
</tr>
<tr>
<td># normalized (positive/negatives) patches</td>
<td>10.00</td>
<td>5.000000</td>
</tr>
<tr>
<td>Num of positives</td>
<td>21.90</td>
<td>21.90</td>
</tr>
<tr>
<td>Gap s.d. (human, human)</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>Gap s.d. (Binary, multi-class)</td>
<td>0.16</td>
<td>0.10</td>
</tr>
</tbody>
</table>
• No sharp gap between MIRCs and sub-MITCs, humans’ is much larger

• Limited recognition
  – (60% accuracy at 75% recall)

• Humans are much better

• Less false detections, different false detections
Example false detections (DPM)
Internal Interpretation

also for Validation

internal interpretations, produced automatically by a model
Cannot be produced by existing feed-forward models
Interpretation Features

- Shape of an extended contour (grouping)
- ‘Visual words’, texture inside a region (local segmentation)
- Small features at a specific location

Likely Top-Down: complex and class-specific
The Visual Hierarchy

© Oxford University Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/. Source: Felleman, Daniel J., and David C. Van Essen. "Distributed hierarchical processing in the primate cerebral cortex." Cerebral cortex 1, no. 1 (1991): 1-47.
Two stages in recognition

- **Generic classification**
- **Class-specific validation and interpretation**
  - Internal and external
BU – TD Interpretation
Examples from a Model
MEG studies of minimal images

- On the role of top-down processing
- Leyla Isik
- Yena Han
MEG: Decoding Image Identity and Category (MIRCs vs. hard-negatives)

Minimal images as a sensitive tool

- **Eagle:**

![Image Identity: 60 (30 class + 30 non-class) ways classification](image)

![Class/non-class classification](image)
Decoding Image Identity and Category (class/non-class)

- Bike

Image Identity:
60 (30 class + 30 non-class) ways classification

Class/non-class classification
General summary comments on bottom-up and top-down

- Initial feed-forward sweep, DNN type can be very useful

- Triggering computations that depend on complex and class-specific properties and relations, top-down routines

- More related to recurrent, recursive computation, working memory, RNN, LSTM, sequential processes,

- Innate structures are more complex, and contain more information about the world
Summary

• Studies of recognition have focused on the first stage only

• Model: minimal images followed by interpretation and validation

• Features and representations used

• Neural circuits involved, top-down processing

• Full image interpretation

• Actions, agents interactions