Robustness of Object Recognition under Extreme Occlusion in Humans and Computational Models

Hongru Zhu1, Peng Tang1, Jeongho Park2, Soojin Park1, Alan Yuille1
1Johns Hopkins University 2Harvard University 3Yonsei University

Introduction

Motivation. Objects in the visual world are occluded much more than objects in typical visual science experiments. However, both humans and object recognition models are rarely tested on the level of occlusion we encounter in the real world.

Previous Solution. Several studies [1,2] addressed this issue by investigating and developing models that can handle real-world object recognition under constant mask occlusion as shown on the right.

Our Solution. We propose a challenging task of object recognition under extreme occlusion using image dataset from [3] (right), and test humans and computational models including:
- Convolutional Neural Networks (CNNs)
- CNNs w/ recurrent computations that can handle mask occlusion from [1]
- Our two-stage compositional model.

Computational Models

- CNNs
  - AlexNet [4], VGG16 [5] and ResNet-18 [6].
  - CNNs w/ recurrent computations [1]
  - AlexNet fc7 → Hopfield Content-addressable Memory [7] → SVM.
- Two-stage compositional model.

Intuition: With a few parts missing, an occluded object can be recognized as long as the positions of visible parts conform to reasonable spatial constraints.

Stage 1: Object Part Detection (Adopted from [8]) by exploiting spatial constraints over visible subparts.

Stage 2: Object Recognition by aggregating evidence from detected parts and their spatial relationships at different scales via spatial pyramid pooling (SPP).

Behavioral Experiments

Participants. 25 subjects recruited from MTurk to recognize objects in the same 500 occlusion images used to test computational models.

Procedure. Stimuli had bounding boxes around targets. Subjects had unlimited time to observe and type the names of objects in the boxes.

Data Processing and Exclusions. 19,341 (out of 25,000) typed strings that contained keywords belonging to one and only one of the five vehicle categories were used for analysis.

Results

In short. Our two-stage model shows greater robustness and produces the most similar results to humans under extreme occlusion.

Testing Accuracy under No/Extreme Occlusion

<table>
<thead>
<tr>
<th>Models</th>
<th>Humans</th>
<th>No occlusion</th>
<th>Extreme occlusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>98.0%</td>
<td>95.3%</td>
<td>83.3%</td>
</tr>
<tr>
<td>ResNet</td>
<td>90.1%</td>
<td>94.0%</td>
<td>89.8%</td>
</tr>
<tr>
<td>VGG16</td>
<td>94.7%</td>
<td>62.6%</td>
<td>77.7%</td>
</tr>
<tr>
<td>AlexNet+Hopfield</td>
<td>77.7%</td>
<td>46.9%</td>
<td></td>
</tr>
</tbody>
</table>

Category-level Confusion Matrices under Extreme Occlusion

Spatial Heatmaps Learned by our Two-stage Model

In these representational dissimilarity matrices (RDMs), the dissimilarity between two images is measured as the Euclidean distance between two 5-dim vectors representing either human response frequencies or the Softmax model outputs.

References

[1] Tang et al. Recurrent computations for visual pattern completion. PNAS, 2018
[8] Zhang et al. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. CVPR, 2018