Use of local image information in depth edge classification by humans and neural networks

Krista A. Ehinger[1], Wendy J. Adams[2], Erich W. Graf[2], James H. Elder[1]

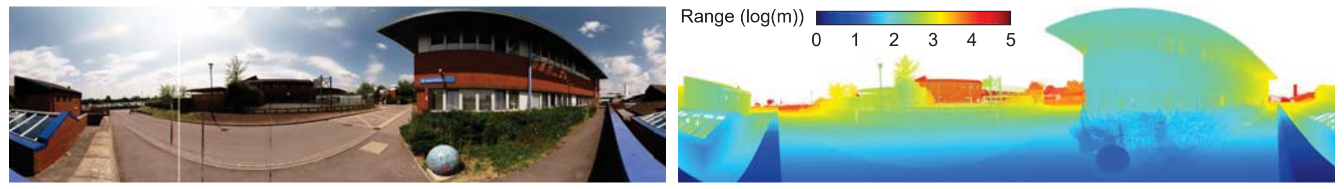
[1]Centre for Vision Research, York University [2]University of Southampton

Introduction

Distinguishing edges caused by a change in depth from other types of edges and establishing figure-ground are important problems in early vision. We compare the performance of humans and a convolutional neural network (CNN) on this task.

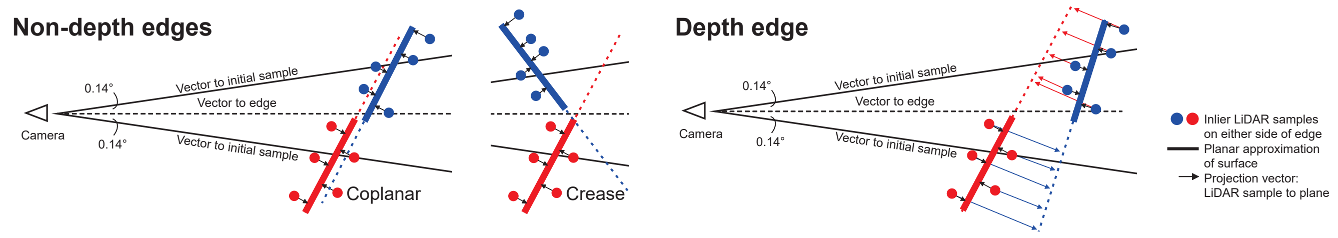
Stimuli

Southampton-York Natural Scenes (SYNS) database [1]: Spherical high dynamic range (HDR) imagery and LiDAR range data from 60 randomly-sampled outdoor locations.



We project images over a uniform sampling of the view sphere and use a multi-scale edge detector [2] to find luminance edges in each view. To identify "depth" and "non-depth" edges, we characterize the 3D surface at the edge:

- Identify two LiDAR samples about 0.14° to either side of the edge.
- Use an adaptive multiscale surface fitting method [1] to estimate local planar approximations to the surfaces at these two points and identify the set of LiDAR samples which are inliers on each plane.
- Mark edges as "non-depth" if surfaces are coplanar (inlier samples from one plane are inliers on the other) or form a crease (planes intersect between the two view vectors). Otherwise, mark edges as "depth."
- Measure the depth change across depth edges, defined as the difference in the distances to the two surfaces.



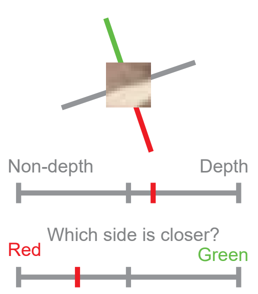
Projected HDR image | Image edges | Depth contrast | Labeled edges

Method

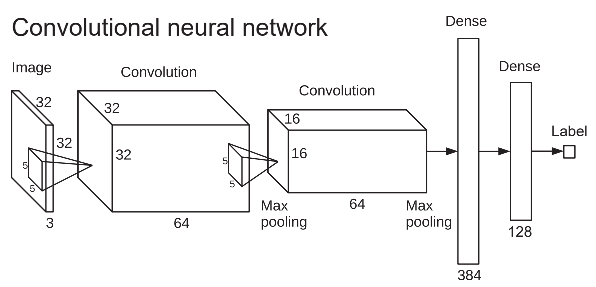
Observers were shown a small square color image patch centred at each edge (width = 8-32 px = 0.6-2.4°) and asked to classify the edge as a "depth" or "non-depth" edge.

Experiment 1
- 8 participants
- 800 edges from 20 scenes (half depth)
- Edges classified as "depth" or "non-depth" (keypress response)
- Binocular presentation

Experiment 2
- 6 participants
- 300 edges from 20 scenes (half depth)
- Edges on very small/complex surfaces (e.g. foliage) excluded
- Ground truth labels verified by two raters
- Depth and figure-ground classification (slider response)
- Monocular presentation



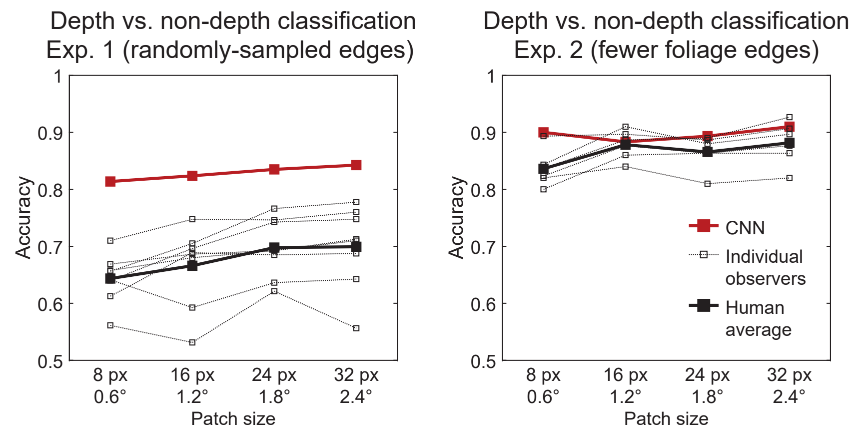
We compared human edge depth classification to the performance of a CNN trained on 200,000 edge patches from 40 scenes not used in the behavioral experiment.



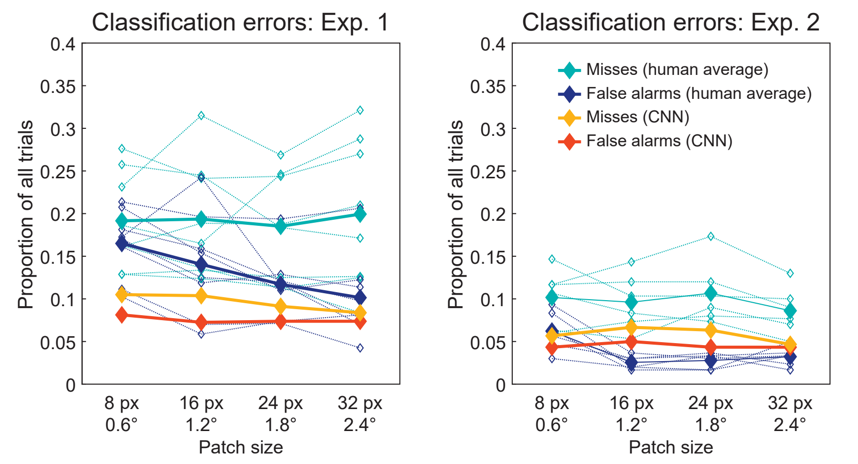
Results: Depth edge classification

Accuracy

Human accuracy increases with patch size (65-70%) but is well below CNN performance (81-85%) in Exp. 1. Human performance was higher in Exp. 2 and comparable to the CNN.

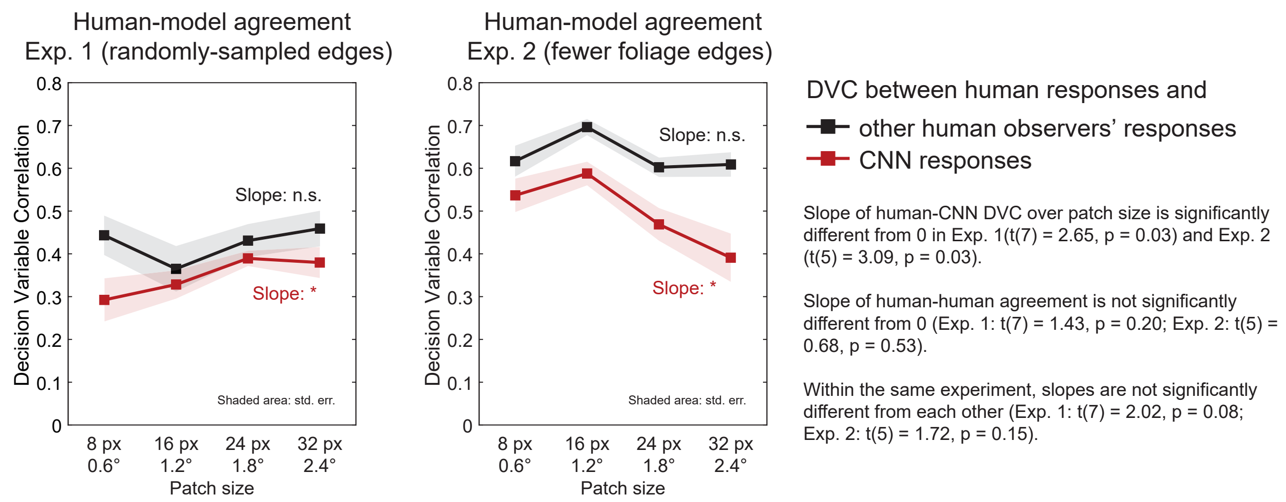


Human observers show a bias towards labeling edges as "non-depth": misses are more common than false alarms.
The CNNs show a smaller bias in the same direction.



Human-model agreement

Decision variable correlation (DVC) [3] was used to measure agreement between human observers and the CNN. DVC uses a signal detection framework to model the similarity between two observers in a 2AFC task. Correlation between human observers and CNN is above chance but lower than human-human agreement.



Slope of human-CNN DVC over patch size is significantly different from 0 in Exp. 1(t(7) = 2.65, p = 0.03) and Exp. 2 (t(5) = 3.09, p = 0.03).

Slope of human-human agreement is not significantly different from 0 (Exp. 1: t(7) = 1.43, p = 0.20; Exp. 2: t(5) = 0.68, p = 0.53).

Within the same experiment, slopes are not significantly different from each other (Exp. 1: t(7) = 2.02, p = 0.08; Exp. 2: t(5) = 1.72, p = 0.15).

Results: Figure-ground classification

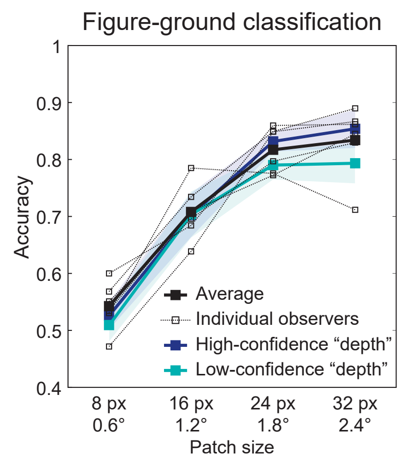
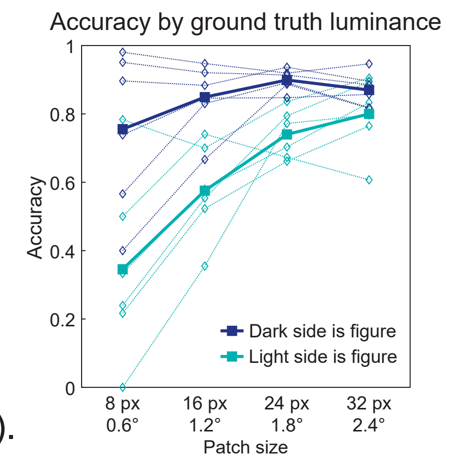


Figure-ground accuracy increases with patch size from 53-84% correct.

Accuracy is only slightly higher for edges labeled "depth" with high confidence (confidence score in the upper half of the distribution of scores from this observer).



Observers show a bias towards labeling the darker side of the edge as "figure," although this is not a reliable cue (the lighter side is figure in 51% of edges).

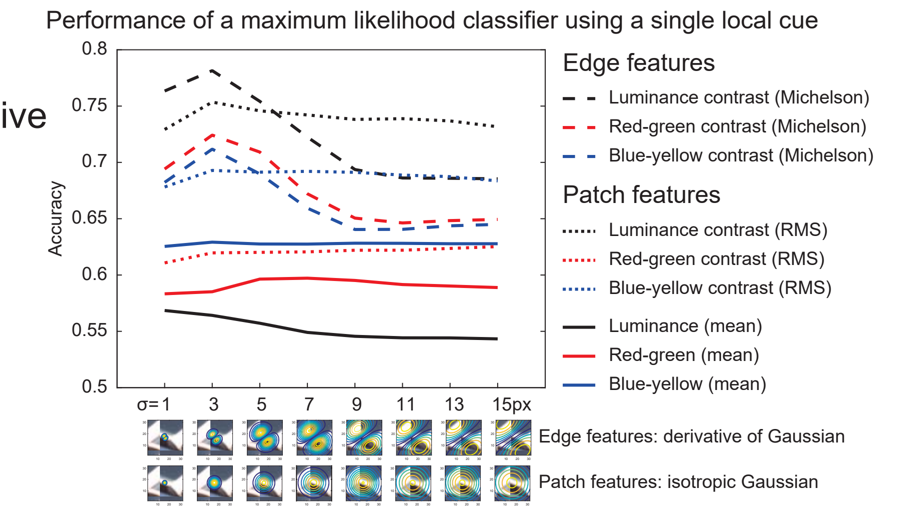
Luminance and color cues for edge depth classification

Local cues

We examine the discriminative power of two kinds of local luminance and color cues:

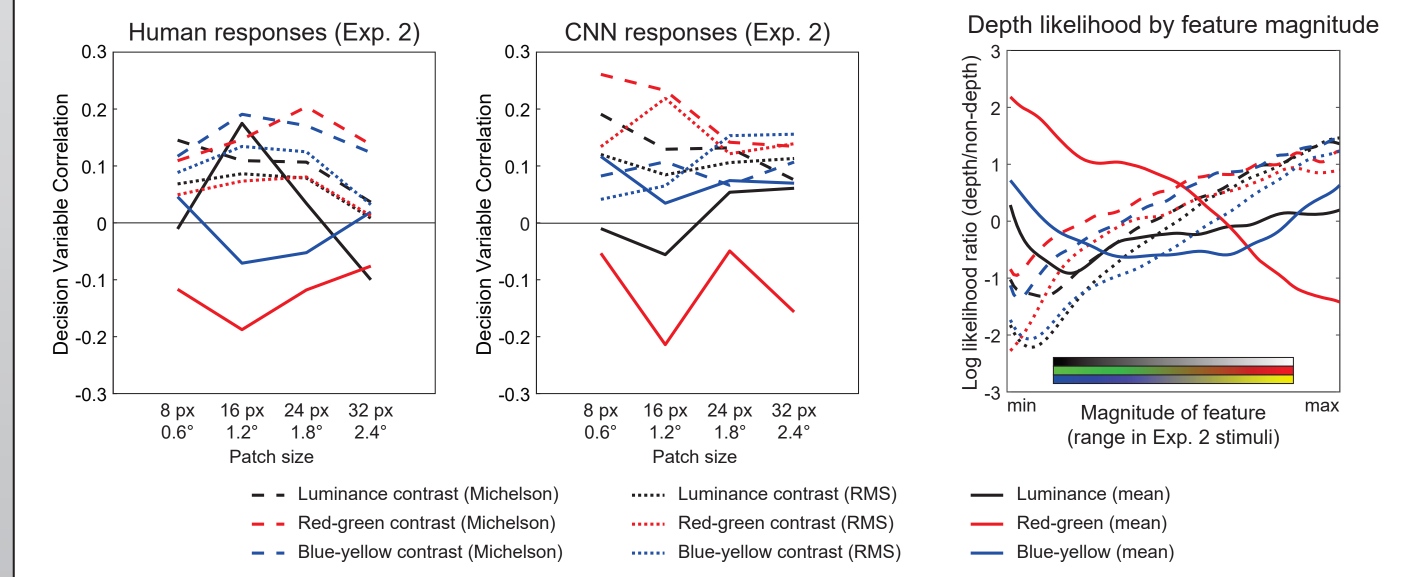
Edge features
Response of a Gaussian derivative filter centered at and aligned with the edge.

Patch features
Response of an isotropic Gaussian filter centered at the edge.



We varied the Gaussian scale constant σ to identify the optimal scale for depth edge discrimination. Contrast cues are the best individual local cues to depth. Performance is highest when contrast is measured in a small region (σ = 0.2°).

Decision variable correlations between the log likelihood ratio of local edge cues and "depth" responses in Experiment 2 show that both human and CNN responses are most associated with contrast cues.



Conclusions

- Observers can accurately discriminate depth from non-depth edges using only a 0.6° window around the edge, but figure-ground discrimination requires a wider view around the edge.
- Human and CNN judgements are highly correlated and rely in part on luminance and color contrast cues.
- But human-human correlation is much higher than human-CNN correlation: there are important determinants of human judgements that the CNN model does not capture.

References

[1] Adams, W.J., Elder, J.H., Graf, E.W., Leyland, J., Lugtigheid, A.J., & Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 6, 35805.
[2] Elder, J. H., & Zucker, S. W. (1998). Local scale control for edge detection and blur estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 699–716.
[3] Sebastian, S. & Geisler, W. S. (2018). Decision-variable correlation. Journal of Vision, 18(4): 3

This research was supported by NSERC Discovery and ORF-RE grants to J.H.E. K.A.E. is funded by a Vision: Science to Applications (VISTA) Award.