Local depth edge detection in humans and deep neural networks

Krista A. Ehinger
York University
kehinger@yorku.ca

Wendy J. Adams
University of Southampton
w.adams@soton.ac.uk

Erich W. Graf
University of Southampton
e.w.graf@soton.ac.uk

James H. Elder
York University
jelder@yorku.ca

Abstract

Distinguishing edges caused by a change in depth from other types of edges is an important problem in early vision. We investigate the performance of humans and computer vision models on this task. We use spherical imagery with ground-truth LiDAR range data to build an objective ground-truth dataset for edge classification. We compare various computational models for classifying depth from non-depth edges in small images patches and achieve the best performance (86%) with a convolutional neural network. We investigate human performance on this task in a behavioral experiment and find that human performance is lower than the CNN. Although human and CNN depth responses are correlated, observers' responses are better predicted by other observers than by the CNN. The responses of CNNs and human observers also show a slightly different pattern of correlation with low-level edge cues, which suggests that CNNs and human observers may weight these features differently for classifying edges.

1. Introduction

Edge detection is an important early step in both human and computer vision. However, edges in images are produced by multiple causes. Some edges are produced by a change in 3D surface normal or a change in depth, but edges also occur on flat surfaces wherever there is a change in surface reflectance or illumination. Distinguishing these different edge types is important for segmenting objects from background and recovering 3D scene structure. Humans are able to perform edge classification seemingly effortlessly, but the mechanism is poorly understood.

Computer vision researchers are increasingly turning to convolutional neural networks (CNNs) to solve complex visual tasks. The current generation of deep CNNs are able to match or exceed human performance on a range of tasks [8, 13, 12, 5, 18]. The hierarchical structure of these networks mirrors the architecture of the human visual system, and the representations learned at different CNN layers is similar to those in visual regions such as V4 and IT [20, 9, 19]. Despite similar task accuracy for CNNs and humans, there is some evidence the two systems perform tasks in different ways. For example, humans and CNNs rely on different image regions when doing the same classification task [11] and CNNs are much more susceptible to image noise [16, 21, 3].

In this paper, we investigate the performance of CNNs and humans on the task of local edge classification in monocular images. Although edge classification is not necessarily only a local problem – feedback from other image areas probably plays a role – we are primarily interested in how local image information is used in the feed-forward processing of edges. Unlike previous studies of this problem, which have generally relied on human-labeled edges, we use spherical imagery and LiDAR range data to build an objective ground-truth dataset for edge classification.

2. Prior work

Most previous work has focused on classifying human-labeled occlusion edges, without considering the amount of depth change at the edge. (Some occlusion edges, such as the edge between the base of an object and a supporting surface, may involve little or no change in depth.)

Balboa *et al*. [2] compared image statistics at regions with and without occlusion edges and found that luminance contrast tends to be higher at an occlusion edge. Ing *et al*. [10] compared human and model performance in discriminating occlusion edge patches from non-occlusion patches in images of outdoor foliage. They found that both humans and models could perform the task with about 80% accuracy, although humans required some task-specific training to reach the same performance as the models. The primary features used by both humans and models were lu-

minance and color contrast. DiMattina *et al*. [4] investigated the same task using a larger variety of outdoor scenes. They found that human observers relied on both luminance and texture cues to perform the task and that a neural network classifier could match human performance, although simpler linear classifiers could not. Sarkar *et al*. [14] proposed a CNN to distinguish occlusion edge patches from non-occlusion patches in indoor scenes. Unlike the previous studies, they defined occlusion edges as edges with a change in depth and obtained ground-truth depth information using a Microsoft Kinect sensor. However, their approach is limited to fairly small indoor environments due to the lighting and range restrictions of the Kinect.

In this study, we model the detection of depth edges as a two-part problem. First all visible edges in the image are detected. Then, these edges are labelled as either depth or non-depth. This may be a more efficient way to approach the problem, since edges (of any kind) make up only a very small percentage of an image. Previous work on this approach comes from Vilankar *et al*. [17], who investigated depth versus non-depth edge classification by human observers. They found that luminance contrast was a particularly strong cue for this task and predicted human edge classification with 83% accuracy. Like prior studies, they used human annotators to label the ground truth depth and non-depth edges.

3. Dataset

We use the publicly-available Southampton-York Natural Scenes (SYNS) dataset (https://syns.soton.ac.uk) [1]. This dataset includes spherical HDR imagery and LiDAR range data from 60 different outdoor locations in the Southampton area, randomly sampled from a set of 20 land use categories. Figure 1 shows an example scene. 20 scenes (one from each land use category) were reserved as a test set and the remaining 40 scenes were used for training.

HDR images were captured using a Spheron SpheroCam HDR with a 360° by 180° field of view and a resolution of 4 arcmin, which produces a 5,400 x 2,700 pixel image in equirectangular projection. 26 exposure stops were used to produce the HDR image. LiDAR range maps were captured using a Leica ScanStation P20 with a 360° by 135° field of view, a maximum range of 125 m, and a resolution of 2.2 arcmin, which gives 10,054 x 3771 samples in equirectangular projection. In each scene, the LiDAR range data were co-registered to the HDR image by finding a rigid transformation which would align the positions of three calibration targets visible in both images, as described in [1].

4. Ground truth for depth edge classification

In each spherical image, we sampled 42 camera angles roughly uniformly over the view sphere using an icosahedral grid. This sampling included views pointing directly up (elevation = 90°) and directly down (elevation = -90°), but the view pointing directly down was discarded because there were no LiDAR samples in this view. Each point on the view sphere was assigned to the closest camera angle, tesselating the sphere into 42 polygonal regions each with a radius of about 17°. At each camera angle, we projected a 48° x 48° image from both the spherical HDR image and LiDAR range map (Figure 2(a)); this field of view is slightly larger than the polygonal region sampled by the camera to avoid problems with edge detection algorithms at the image boundaries. The HDR images in this dataset have a range of 0.0-3696.0. We converted the projected HDR image to an 8-bit image using an exponential compression algorithm [7], with the exposure chosen so that the median value in the HDR image would map to the central grayscale value (127). This formula reduces to:

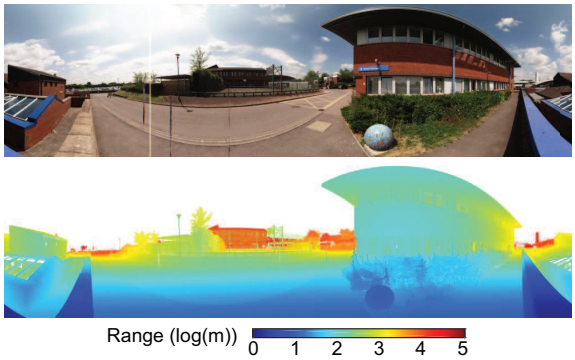


Figure 1: Spherical HDR image (top) and LiDAR range data (bottom) for an example scene from the Southampton-York Natural Scenes dataset [1].

$$I' = 255\left(1 - 0.5^{\frac{I}{m}}\right) \quad (1)$$

where I' is the 8-bit pixel value, I is the HDR pixel value, and m is the median HDR value in the projected image.

In this study, we restricted our attention to luminance edges, converting the 8-bit color images were to grayscale luma (Y') using the Rec. 601 standard. We detected edges in the image and range map using a multiscale edge detector [6] with noise parameter set to 3 gray levels for image edge detection (Figure 2(c)) and 1 mm for range edge detection (Figure 2(d)). We discarded edges in image regions with missing LiDAR data, as well as edges generated by motion artifacts or the calibration targets used to align the HDR and LiDAR images. Motion artifacts, which appear as vertical streaks in the HDR image, were identified automatically by comparing vertical filter responses in two HDR images taken a few minutes apart. Pixels above

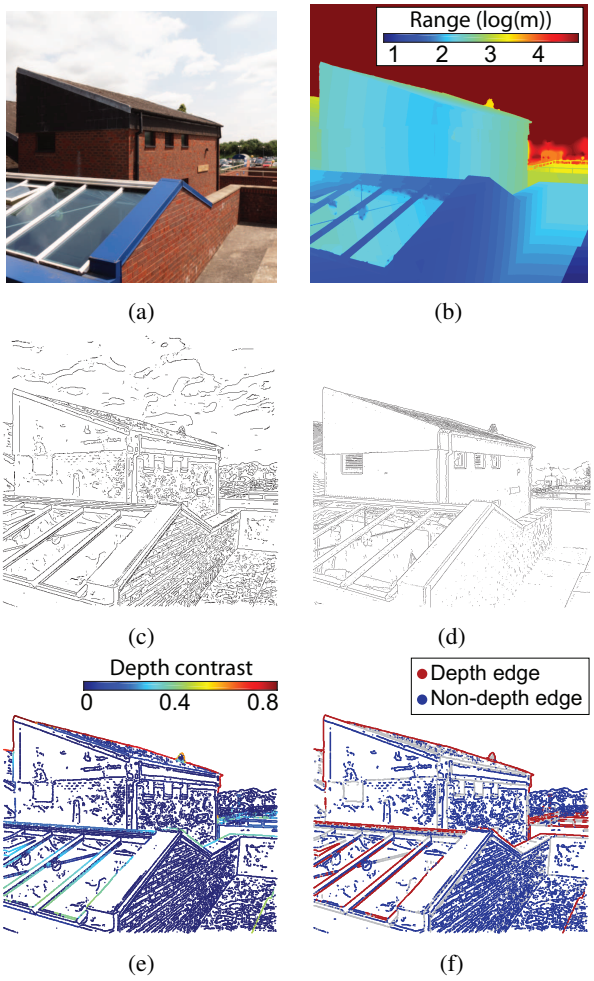


Figure 2: Overview of our method to compute ground-truth depth contrast at image edges. We project views from (a) the spherical HDR image and (b) range map. We use a multiscale edge detector to detect (c) image edges and (d) range edges. For each image edge, we look for a matching depth edge. (e) Matched edges are labeled according to their depth contrast; unmatched edges are labeled with a depth contrast of zero. (f) We assign edges to two classes: "non-depth" edges, which have no matching edge in the range map, and "depth" edges with a depth contrast >0.1.

an empirically-determined threshold difference were identified, and regions above a threshold size were labeled as motion artifacts. The calibration targets were labeled by hand.

The rigid alignment of the spherical HDR and LiDAR images in the database is quite good: 66% of points are aligned to within 0.067° (the width of an HDR pixel at the horizon). However, the remaining alignment errors can be as high as 0.4° (about 6 pixels), which causes significant problems when trying to measure the ground truth depth change at image edges. To correct for these remaining alignment errors, we associated each image edge with the range edge most likely to be a match, by finding the range edge i which maximized the posterior probability:

$$\frac{p(x_i|match)p(\theta_i|match)p(match)}{p(x_i|nonmatch)p(\theta_i|nonmatch)p(nonmatch)} \quad (2)$$

where x_i is the angular distance between the image and range edge, θ_i is their difference in orientation, and $p(match)$ and $p(nonmatch)$ are the prior probabilities on an image edge matching (or not matching) any randomly-selected range edge. The priors and the distributions for matched edges $p(x_i|match)$ and $p(\theta_i|match)$ were learned from training data by matching a subset of edges by hand. The distributions for nonmatching edges assume that both edges are selected at random: the distribution of distances for nonmatching edges $p(x_i|nonmatch)$ is the distance between any two random points in an image and the distribution of orientation difference $p(\theta_i|nonmatch)$ is uniform.

If no range edge with posterior probability above a threshold was found, then the image edge was labeled as a "non-depth" edge (25% of all edges). We chose a threshold of 0.05 based on visual inspection of the results. For Spheron edges with a match, we computed depth contrast: the difference between the range values on either side of the edge divided by their sum. We measured the range value on each side of the edge by averaging the LiDAR range values at 3 sample locations in the direction of the gradient, at a distance of 0.07, 0.11, and 0.14° from the edge. Edges with depth contrast greater than 0.1 were labeled as "depth" edges (13% of all edges). Estimated edge depth contrasts and edges labelled as depth and non-depth for a sample image are shown in Figure 2(e) and 2(f), respectively.

5. Computational methods: A CNN for local edge depth classification

We built a small neural network for depth edge classification. The network consists of two convolutional layers followed by two fully-connected layers. The first convolutional layer uses 64 kernels of size 5 pixels x 5 pixels x 3 color channels with a stride of 1 pixel; the second convolution layer uses 64 kernels of size 5 pixels x 5 pixels x 64 channels with a stride of 1 pixel. Each convolutional layer is followed by a normalization and max pooling in a 2 x 2 pixel window, which reduces the image size by one half. The third and fourth layers of the network are fully connected, consisting of 384 and 128 nodes, respectively. Each fully-connected layer is followed by 50% dropout during training. The last fully-connected layer outputs to a softmax classifier.

The input to the network is an image patch centered on a depth or non-depth edge (patch size = 8, 16, 24, or 32 pixels). We doubled the size of the training set by including left-right mirror-reversed patches; no other data augmentation methods were used. Patches were normalized to have a mean intensity of 0 and standard deviation of 1. However, the original mean value and standard deviation of the patch were concatenated onto the output of the final convolutional layer and provided as input to the first fully-connected layer.

We used 200,000 edges sampled from the 40 training scenes for training, and tested on 100,000 edges sampled from the 20 test scenes. Each set included 2,500 "non-depth" edges per scene and an equal number of "depth" edges drawn randomly from the pool of all edges with depth contrast over 0.1. The network was trained using stochastic gradient descent with a learning rate of 0.1 and rate decay of 0.1. The network was trained for 20,000 epochs with a batch size of 128.

In addition to the standard network, we also trained and tested networks on manipulated images. We removed color information from the patches by converting them to grayscale luma (Y') using the Rec. 601 standard. In another condition, we removed edge orientation information by rotating all patches so that the detected edge was vertical. In a third condition, we removed color and edge orientation information by combining these manipulations.

6. Computational results

The performance of the CNN classifier across a range of patch sizes and image conditions is shown in Figure 3. When classifying unmanipulated images, performance increases with patch size, from 83% correct for the smallest patch size (8 pixels) to 86% for the largest patch size (32 pixels). The fact that performance improves only slightly with increasing patch size suggests that the local information at the edge provides much of the information needed to distinguish depth from non-depth edges.

Both orientation and color seem to be important cues for edge classification. Performance drops about 5% when patches are rotated to a standard orientation, which suggests that the orientation of the edge, or of structures near an edge, are important for depth classification. Removing color results in a larger performance drop, particularly for the smallest patches. The information provided by color and orientation is somewhat independent, since removing both cues results in a larger performance drop than removing either cue alone.

6.1. Network parameters

Figure 4 shows the effect of varying training set size and the depth contrast threshold for labeling edges as "depth." All networks were trained and tested on 32-pixel image patches in their original color and orientation. Performance

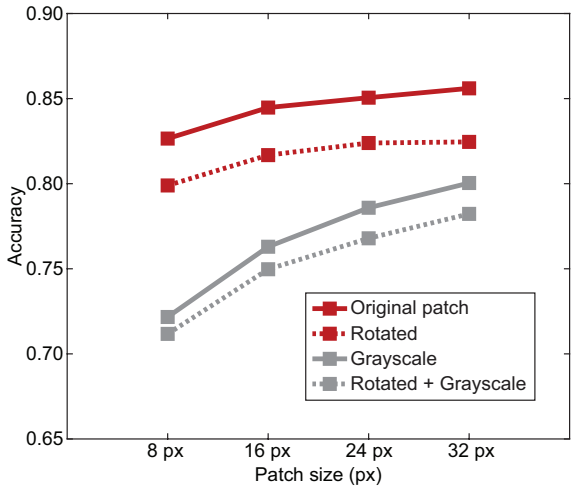


Figure 3: CNN classification performance.

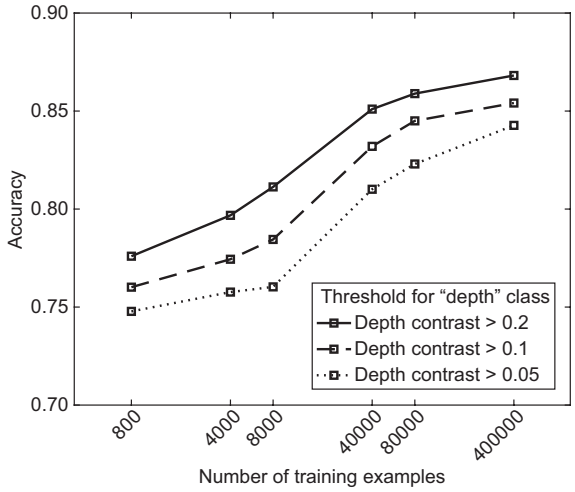


Figure 4: CNN performance over different training set sizes and threshold levels for labeling edges as "depth."

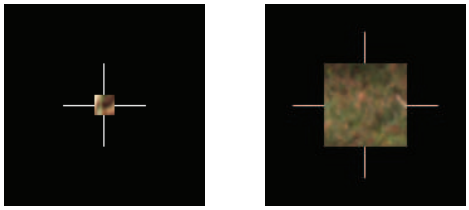
increases with the number of training examples, though the increase slows down after about 40,000 examples. Performance also increases as the threshold used to identify "depth" edges increases. This may be because high-depth edges are less varied or have more reliable image cues.

6.2. Comparison with other methods

Table 1 shows the performance of various methods in classifying depth versus non-depth edges using a 32 pixel patch centered on the edge. Patches were rotated so that the edge was vertical, with the higher-luminance side of the edge on the left, and the original edge orientation was provided to the classifier as a separate input. Classifiers were trained in MATLAB using 40,000 edges for training (50%

Classifier	Accuracy
CNN	0.83
SVM (Gaussian kernel)	0.78
Logistic regression	0.71
K-NN (K = 10)	0.56

Table 1: Comparison of different edge classification methods, classifying 32-pixel image patches.



(a) (b)

Figure 5: Examples of patches presented in the behavioral experiment: (a) 8 pixels (b) 32 pixels.

depth) with 5-fold cross-validation and 20,000 edges for test. The CNN trained with 40,000 examples achieves an accuracy of 83%, while the next best performance, 78%, comes from an SVM using a Gaussian kernel.

7. Psychophysical methods: Human depth edge classification

We ran a behavioral experiment in which eight participants classified a subset of the test patches shown to the models.

7.1. Apparatus

Images were displayed on a 26 x 35.5 cm (1024 x 768 pixel) CRT display with a refresh rate of 100 Hz. Participants were seated 53 cm from the screen in a darkened room. A headrest was used to maintain viewing distance.

7.2. Stimuli

The stimuli were 800 edge patches (50% "depth") from the test scenes. Each patch was shown at the angular size it subtended in the original scene (for example, a 32 pixel patch was shown at 2.4° = 66 pixels).

7.3. Observers

Eight participants (three female) took part in the experiment.

7.4. Procedure

Patches were shown individually in the center of a black screen, with cross-hairs and a red dot which flashed once for 300 ms to indicate the location of the central edge pixel (Figure 5). Participants labeled each patch as "depth" or "non-depth" by pressing one of two keys. There was no time limit on responses.

Each patch was shown four times at four different patches sizes: 8, 16, 24, and 32 pixels. Patches were presented in blocks of increasing size, so participants classified all 800 edges in 8 x 8 pixel patches in the first block, then classified the same 800 edges in 16 x 16 pixel patches in the second block, etc. By monotonically increasing patch size, we ensure that participants cannot use remembered information from the larger patches to classify the smaller ones. Presentation order was random within each block.

Prior to the start of the first block, participants did a practice block in which they classified 80 edges from the training scenes. Edges appeared in one of the four patch sizes, selected at random, and participants received feedback after each response in the practice block. There was no feedback during the experimental blocks.

8. Psychophysical results

Human classification performance across patch size is shown in Figure 6, along with the performance of the CNN classifier on the patches shown in the behavioral experiment. On average, human performance increased with patch size. A repeated measures ANOVA shows a significant effect of patch size ($F(3,21) = 8.29$, $p < 0.01$). Posthoc Tukey HSD tests indicate that performance on the smallest patch size is significantly different ($p < 0.01$) from performance on the two largest patch sizes.

The agreement between observers in classifying patches was fairly low. We measured inter-observer agreement using Krippendorf's alpha, which ranges from 0 (no agreement) to 1 (perfect agreement), with values above 0.8 generally considered to mean good agreement. Inter-observer agreement on the depth classification task ranged from 0.32 and 0.28 for the 8- and 16-pixel patches, respectively, to 0.36 for the two largest patch sizes.

At all patch sizes, human observers were less accurate than the CNN classifier. However, this does not necessarily mean that computer models have surpassed humans at real-world depth edge discrimination, since there were some differences between the way patches appeared in the experiment and the way they would have appeared in the real world. The stimuli patches were presented on a flat surface at a fixed distance from the observer, which could make binocular cues misleading. The patches were also lower contrast and lower resolution (about 13 pixels per degree) than they would have been if foveated in natural viewing conditions. These differences may have misled observers who were relying on priors learned from their everyday experience with outdoor scenes. Human observers may be able to match CNN performance if given more extensive

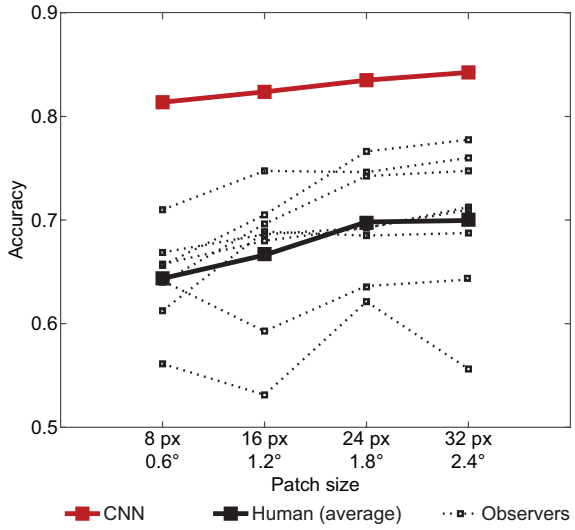


Figure 6: Human depth edge classification performance. Dotted lines show the performance of individual observers. The performance of the CNN on the same patches is shown in red.

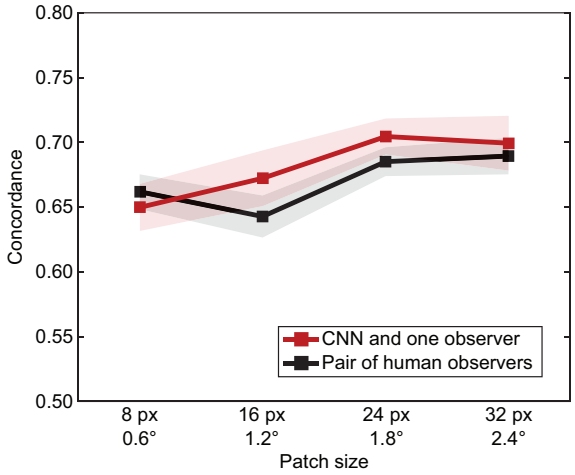


Figure 7: Average concordance between human observers and the CNN versus pairs of human observers. The shaded region indicates standard error.

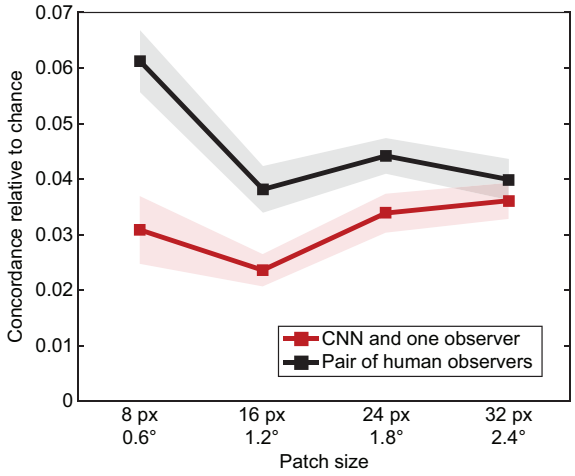


Figure 8: Average corrected concordance between human observers and the CNN versus pairs of human observers. The shaded region indicates standard error.

training on the lab-based classification task. However, we did not do this because we are interested in how people naturally interpret edge depth from images.

9. Human-model comparison

Figure 7 shows the concordance between observers' responses and the CNN compared to the concordance between all pairs of observers. Concordance is computed as the percentage of trials on which both observers give the same response. Concordance between humans and the CNN is similar to concordance between pairs of human observers and increases with patch size. However, this uncorrected measure of concordance is confounded with both accuracy and response bias. In this experiment, accuracy increased with patch size, and most observers (and the CNN) showed a bias to label edges as "non-depth." Both of these factors work to inflate the apparent agreement between the CNN and human observers.

Figure 8 shows the corrected concordance, calculated as the difference between the proportion of trials on which observers agree and the proportion of trials on which they would have agreed by chance given their performance (proportion correct) in each ground truth category. This gives a measure of the agreement not explained by the accuracy or bias of the observers. In general, corrected concordance between pairs of human observers is higher than corrected concordance between observers and the CNN, and the concordance does not increase with patch size.

We also compared the human and model responses to individual edges using the method described in [5]. For each image patch shown in the experiment, we computed a human depth confidence score which was the percentage of observers who labeled the patch as "depth." We used the "depth" logit values from the CNN classifier as a measure of the model's confidence. The correlation between human and model confidence across patch sizes is shown in Figure 9. Human and model confidence is moderately correlated with a correlation coefficient (Spearman's rho) of 0.52 for the smallest patch size and 0.60-0.63 for the larger patch sizes. The increase in correlation with patch size is likely

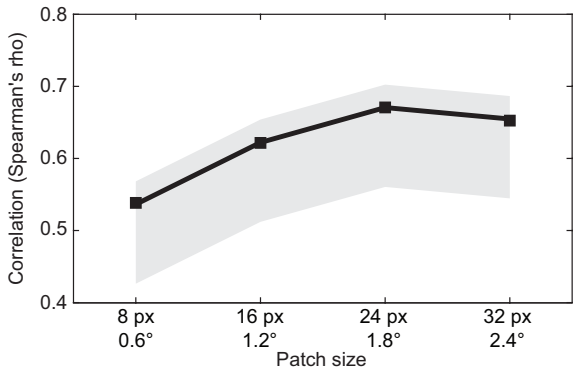


Figure 9: Correlation coefficient (Spearman's rho) between the human observer and CNN classifier confidence in classifying edges as "depth." The shaded region indicates the 95% bootstrap confidence interval.

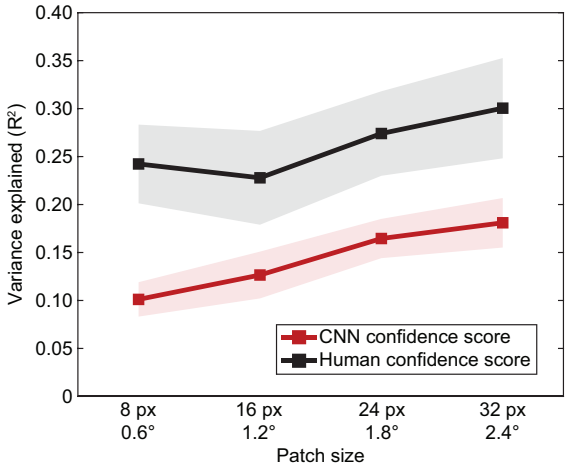


Figure 10: Percent of variance in individual observers' responses which is explained by the CNN or human confidence score. The shaded region indicates standard error.

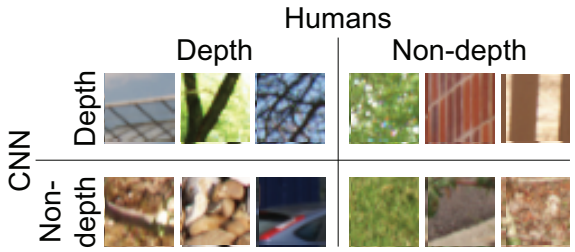


Figure 11: Examples of patches labeled as "depth" or "non-depth" by human observers (columns) or the CNN (rows).

due to the increase in accuracy with patch size for both human observers and the CNN.

We can use the confidence scores to predict the responses of individual observers in a logistic regression. The average variance in human observer responses explained by the CNN and human "depth" confidence scores is shown in Figure 10. For this analysis, the human confidence score is computed using N-1 observers; the observer that will be predicted by the regression is excluded. The variance explained by each type of confidence score increases with patch size, but this is likely due to the increase in accuracy with patch size. Across all patch sizes, the regression based on human confidence scores can better explain individual observers than the regression based on the CNN confidence scores. In other words, we can better predict the "depth" responses of a single human observer by averaging responses from other observers than by using the "depth" responses of the CNN.

Examples of patches on which human observers agree or disagree with the CNN are shown in Figure 11. Human observers tend to agree with the CNN on patches that show foliage or objects against sky, which both label as "depth," and patches which show ground textures, which both label as "non-depth." There is more disagreement about dense foliage, which tends to be labeled as "depth" by the CNN but not human observers, and some patches with high-contrast textures or man-made objects, which human observers are more likely to label as "depth."

9.1. Role of individual edge cues

There are a number of local cues which may be useful for edge classification. Table 2 shows the performance of a Bayesian classifier using kernel density estimation (Gaussian kernel with bandwidth from [15]) and a single local feature. Features are computed in a 3 x 3 pixel patch on either side of the edge. To measure color at the edge, we convert patches to CIE-LAB and use the A and B channels as a measure of red-green and blue-yellow intensity, respectively. We use two measures of contrast: Michelson and root mean square (RMS). Michelson contrast is the difference in the mean intensity on either side of the edge over the sum of the means. RMS contrast is the standard deviation of intensity over the mean intensity of the entire patch.

Although luminance contrast is a useful feature for edge classification, it is only able to predict edge depth in this dataset with 66% accuracy; color contrast measures perform somewhat better. The best single predictor of edge type in our dataset was angular elevation (i.e., the "vertical" location of the edge in the spherical image), which is predictive because depth edges in outdoor scenes mostly occur around or above the horizon.

Figure 12 shows the performance of a Bayesian classifier trained to predict CNN or human "depth" responses from individual features. The classifier uses a Gaussian kernel

Edge cue	Accuracy
Luminance (mean)	0.61
Luminance contrast (Michelson)	0.66
Luminance contrast (RMS)	0.50
Red-green (mean)	0.69
Blue-yellow (mean)	0.59
Red-green contrast (Michelson)	0.68
Red-green contrast (RMS)	0.71
Blue-yellow contrast (Michelson)	0.69
Blue-yellow contrast (RMS)	0.72
Edge orientation	0.50
Edge elevation	0.78
All cues except elevation	0.70
All cues	0.71

Table 2: Classification of depth versus non-depth edges using individual edge cues in a Naïve Bayes classifier.

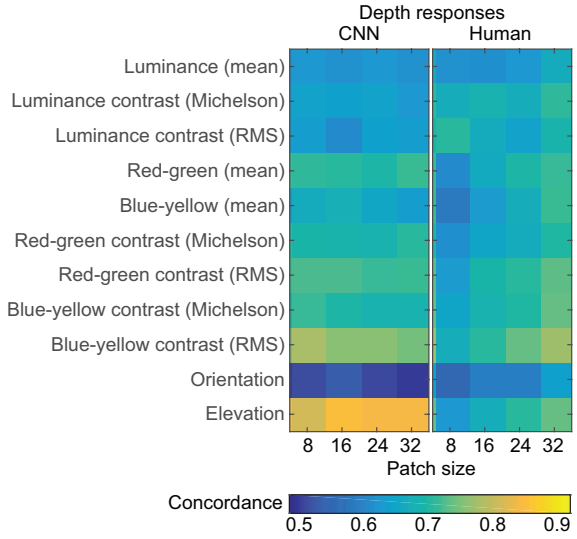


Figure 12: Predictions of the of CNN (right) or human (left) depth responses across patches sizes by a Bayes classifier using various edge cues.

with bandwidth from [15]). The classifier was trained on 50% of the images and tested on the remaining 50%; the reported accuracy is the average of 100 random train-test partitions. For the human responses, the reported accuracy is the average of 100 random train-test partitions for each of the 8 observers.

CNN depth responses are best predicted by elevation and blue-yellow contrast at the edge, which our previous analysis indicates are the cues most strongly related to depth in this dataset. Note that while the CNN cannot directly use elevation as a feature to classify patches, since this information was not available, it may rely on other features that vary with elevation such as foliage color or textures. Blue-yellow contrast is the best predictor of human responses at larger patch sizes, but luminance contrast is a better predictor for the smallest patches. Compared to the CNN, human observers' responses are less well predicted by most features, which may simply reflect the fact that humans were less accurate at this task. Compared to the CNN, human responses are better predicted by the luminance contrast across the edge and the orientation of the edge. These differences may explain some of the disagreement between the CNN and humans (Figure 11). The CNN is more likely to label dense tree foliage as "depth" because it relies on features correlated with elevation, while human observers are more likely to label high-contrast ground textures as "depth" because they rely more on luminance contrast.

10. Conclusion

In this study, we used spherical imagery with LiDAR range data to build an objective ground truth database for local edge classification. We found that CNNs can distinguish depth versus non-depth edges quite accurately, and in fact the best networks exceed the performance of untrained humans at this task. Although the CNN captures some aspects of human performance, it may also be exploiting some cues which humans do not use, which allows it to achieve higher performance. As a result, a CNN is not a perfect model for human depth edge discrimination.

References

[1] W. J. Adams, J. H. Elder, E. W. Graf, J. Leyland, A. J. Lugtigheid, and A. Muryy. The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 6:35805, 2016. 2

[2] R. M. Balboa and N. M. Grzywacz. Occlusions and their relationship with the distribution of contrasts in natural images. *Vision Research*, 40(19):26612669, 2000. 1

[3] T. Borkar and L. Karam. Deepcorrect: Correcting dnn models against image distortions. *arXiv*, 2017. 1

[4] C. DiMattina, S. A. Fox, and M. S. Lewicki. Detecting natural occlusion boundaries using local cues. *Journal of Vision*, 12(13):1–21, Aug. 2012. 2

[5] S. Eberhardt, J. G. Cader, and T. Serre. How deep is the feature analysis underlying rapid visual categorization? In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1100–1108. Curran Associates, Inc., 2016. 1, 6

[6] J. H. Elder and S. W. Zucker. Local scale control for edge detection and blur estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):699–716, 1998. 2

[7] P. Ferschin, I. Tastl, and W. Purgathofer. A comparison of techniques for the transformation of radiosity values to monitor colors. In *Proceedings of 1st International Conference*

on Image Processing, volume 3, pages 992–996 vol.3, Nov 1994. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *arXiv*, 2015. 1

[9] H. Hong, D. L. K. Yamins, N. J. Majaj, and J. J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 2016 Feb 22 2016. 1

[10] A. D. Ing, J. A. Wilson, and W. S. Geisler. Region grouping in natural foliage scenes: Image statistics and human performance. *Journal of Vision*, 10(4):1–19, 2010. 1

[11] D. Linsley, S. Eberhardt, T. Sharma, P. Gupta, and T. Serre. Clicktionary: A web-based game for exploring the atoms of object recognition. *arXiv*, 2017. 1

[12] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with Gaussian Face. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 3811–3819. AAAI Press, 2015. 1

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015. 1

[14] S. Sarkar, V. Venugopalan, K. Reddy, J. Ryde, N. Jaitly, and M. Giering. Deep learning for automated occlusion edge detection in RGB-D frames. *Journal of Signal Processing Systems*, 88(2):205–217, Aug. 2017. 2

[15] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC, 1986. 7, 8

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv*, 2013. 1

[17] K. P. Vilankar, J. R. Golden, D. M. Chandler, and D. J. Field. Local edge statistics provide information regarding occlusion and nonocclusion edges in natural scenes. *Journal of Vision*, 14(9):1–21, Aug. 2014. 2

[18] T. Weyand, I. Kostrikov, and J. Philbin. Planet - photo geolocation with convolutional neural networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 37–55, Cham, 2016. Springer International Publishing. 1

[19] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356365, 2016. 1

[20] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 2014 May 8 2014. 1

[21] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. *arXiv*, 2016. 1