Euclidean coordinates are the wrong prior for models of primate vision

Garrison W. Cottrell Shubham Kulkarni Martha Gahl
Computer Science and Engineering Department
UCSD, 9500 Gilman Drive
La Jolla, CA 92093-0404 USA

Abstract:

The mapping from the visual field to V1 can be approximated by a log-polar transform. In this domain, scale is a left-right shift, and rotation is an up-down shift. When fed into a standard shift-invariant convolutional network (CNN), this provides scale and rotation invariance. However, translation invariance is lost. This is compensated for by multiple fixations on an object. Due to the high concentration of cones in the fovea with the dropoff of resolution in the periphery, fully 10 degrees of visual angle take up about half of V1, with the remaining 170 degrees (or so) taking up the other half. This layout provides the basis for the central and peripheral pathways. Simulations with this model closely match human performance in scene classification, and competition between the pathways leads to the peripheral pathway being used for this task. Remarkably, despite the property of rotation invariance, this model provides a novel explanation for the inverted face effect. We suggest that using Euclidean image coordinates is the wrong prior for models of primate vision.

Keywords: Primate vision; log-polar transform; deep learning; face inversion effect; scene perception.

Introduction

Well-trained deep convolutional neural networks (CCNs) have provided the best model so far of primate vision (Kietzmann, et al., 2019; Yamins, et al. 2014; Storrs, et al. 2021), with early layers predicting activation in early visual areas in monkey cortex, and deep layers predicting activation of IT cortex. An advantageous property of CNNs is their built-in translation invariance. Other invariances, such as size and rotation, must be learned from the training data.

Here we propose that this built-in prior, translation invariance, is not the best one to account for how primate vision works. It is well-known that the retina is foveated, with a high concentration of (cone) photoreceptors in the fovea and foveola, and a nearly linear drop-off of rods from the center to the peripheral region of the retina (Curcio, et al., 1990). Second, the way in which the fibers innervate V1 leads to a coordinate system change from Euclidean (x,y) coordinates to polar (r, θ) coordinates. In particular, due to the falloff of photoreceptors from central to peripheral vision, the mapping is actually well approximated by a (log r, θ) coordinate system (Polimeni, et al., 2006).

This representation leads to scale changes becoming a left-right shift (equivariance), and rotation becoming an up-down shift (equivariance). See Figure 1. When this representation is input to an otherwise standard convnet, we get scale and rotation *invariance* due to the translation invariant property of convolutional networks. However, translation invariance in the original domain is lost. We make up for this by training on multiple fixations (Figure 2). Note that in this representation, we can think of fixating different locations on the face as a form of autonomous data augmentation. Furthermore, with central vision on one side, and peripheral vision on the other, the representation is now formatted such that we can have a central pathway and a peripheral pathway arising from this (Figure 3).

Experiments with The Model 2.0

We call this model "The Model 2.0" (TM2.0), following on earlier work with a shallower version called The Model™ (Cottrell & Hsiao, 2011). The input image is foveated, log-polarized, and split into central and

Figure 1. Log-polar equivariances. (a) Scale. (b) Rotation.

Figure 2. The effect of differing fixations on the log-polar representation; this results in self-augmentation of the data.

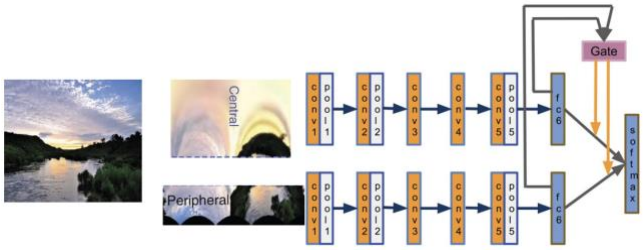

Figure 3. Original image (left), central & peripheral pathways.

peripheral images. The log-polar format separates central from peripheral input and supports two pathways, as shown in Figure 3, corresponding to the central and peripheral pathways in humans and primates. The model's structure is consistent with the various functional distinctions in humans across the mid-fusiform sulcus: central/peripheral, faces/places, small/large (Grill-Spector & Weiner, 2014). The two pathways compete to solve the problem, using a mixture of experts design (Jacobs et al., 1991).

An example of importance of the central/peripheral distinction was highlighted in a series of experiments reported earlier (Wang & Cottrell, 2017). We compared TM2.0 to human data from an experiment by Larson & Loschky (2009). In L&L's experiment, subjects were asked to verify whether a scene was a certain category or not. They were shown stimuli masked in such a way that either the central (scotoma) or peripheral (tunnel vision) portions of the stimulus were masked to different degrees. TM2.0 learned to use the peripheral pathway for this judgment, consistent with the faces/places distinction, and L&L's finding that peripheral vision is more efficient for scene recognition. Furthermore, TM2.0 demonstrated an excellent fit to the subjects' performance. Critically, the fit to the data improved the more anatomical constraints were used (standard CNN < Foveated input < Log-polar input). TM2.0 also closely matched the "critical radius", the point that produces equal scene recognition performance between the two conditions. For human subjects, it was 7.48° degrees; for the model, it was 8.0°, not significantly different from the human result.

A potential puzzle with a rotation invariant model is: How could it possibly account for the face inversion effect, that inverted faces are much harder to recognize, match, and remember than upright faces (Farah, Tanaka & Drain, 1995; Kohler, 1940; Yin, 1969)? The explanation lies in a critical distinction between the representation of scale and rotation invariance. Scale is just a shift left or right. However, rotation is a circle group, but the cortex is flat, not cylindrical. In V1, the representation goes from 90 degrees to 270 degrees. When an object is rotated, part of it "falls off" the top and reappears at the bottom (Figure 1(b)): Notice that for small rotations, Elon's nose is next to his left eye, but after inversion, it's next to his right eye – a different configuration of his features.

In face recognition, a form of visual expertise, the configuration of features is important, commonly called holistic processing (Young, Hellawell & Hay, 1987). In the log-polar representation, the facial features, the nose, the eyes, etc., stay upright (Fig. 1b), but their configuration differs when the face is inverted. This in consistent with visual scientists' common interpretation of face inversion effect: Since the configuration is disrupted, human subjects resort to feature processing (Farah & Tanaka, 1993).

This qualitative observation is borne out in experiments comparing TM2.0 with a vanilla convnet. To model face processing experience over development, we trained the networks to gradually learn more faces, starting with 4, then 8, etc., until they learned 128 faces. Faces were presented at random tilts from -15 to +15 degrees. At each stage of training, we compared the performance of the network on the upright faces in the holdout set to performance on the same faces inverted; this gap is our measure of the inversion effect. As the network became more of a face expert, the inversion effect increased. The vanilla convnet's performance on inverted faces fell nearly to chance (5%), while TM2.0 was still able to perform at 30%, which is more consistent with human performance. In addition, the inversion effect was similar when TM2.0 was trained to be a dog or car expert, consistent with the expertise hypothesis concerning face recognition (Diamond & Carey, 1986; Gauthier & Tarr, 1997). Moreover, the inversion effect was much smaller for objects trained at the basic level (screwdriver, ladle, dumbbell, etc.). Objects known at the basic level (Rosch et al. (1976) are recognized based on their features; configuration does not matter as much (Karimi-Rouzbahani et al., 2017). Finally, houses, which are a mono-oriented object, fell between objects and faces, cars, and dogs. This is also what Yin (1969) found: In a recognition experiment, subjects were worst at inverted faces, then inverted houses, then inverted objects.

Conclusions

We have argued here that, as models of primate vision, the standard CNN has the wrong prior for representing images. While the standard approach has the advantage of being relatively translation invariant, it is severely disrupted by inversion of faces.

On the other hand, the anatomical data shows that the transformation from the visual field to V1 is well approximated by a log-polar transform. This representation has advantages in that, when given to a standard CNN, results in scale and rotation invariance.

Translation invariance is lost, but can be remedied in the same way we do; by fixating objects of interest at multiple locations. Our experiments with this model demonstrate a better fit to human data than a standard CNN in both scene, face, and object processing, and provides a novel explanation for the face inversion effect. We conclude that the Euclidean representation of the inputs is the wrong prior for models of the primate visual system.

Acknowledgments

This research was supported in part by NSF grant CRCNS-2208362, and NSF cooperative agreement SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center.

References

Cottrell, G.W. & Hsiao, J.H. (2011) Neurocomputational Models of Face Processing. In A.J. Calder, G. Rhodes, M. Johnson, and J. Haxby (Eds.) *The Oxford Handbook of Face Perception*. Oxford, UK: Oxford University Press.

Curcio, C. A., Sloan, K. R., Kalina, R. E., and Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, **292**(4):497–523.

Diamond, R. and Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, **115**(2):107–117.

Farah, M. J., & Tanaka, J. W. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, **46**:225–245.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995) What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):628-634.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring face recognition mechanisms. *Vision Research*, **37**:1673–1682.

Grill-Spector, K. and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience* 15:536–548.

Jacobs, Robert A., Jordan, Michael I., Nowlan, Steven J., and Hinton, Geoffrey E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, **3**(1):79-87.

Karimi-Rouzbahani, H., Bagheri, N. & Ebrahimpour, R. (2017) Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models. *Scientific Reports* **7**, article 14402.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O. and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, **116**(43):21854–21863.

Kohler, W. (1940). *Dynamics in Psychology.* New York: Liveright Publishing Corporation.

Polimeni, J., Balasubramanian, M. & Schwartz, E. (2006). Multi-area visuotopic map complexes in macaque striate and extra-striate cortex. *Vision Research*, **46**(20):3336–3359.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Storrs, K. R;, Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, **33**(10):2044–2064.

Wang, P., and Cottrell, G. W. (2017) Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of Vision* **17(4):**9-22.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of SciencesI,* **111**:8619–8624.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, **81**:141–145.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, **16**:747–759.