Extreme Translation Tolerance in Humans and Machines

Ryan Blything (ryan.blything@bristol.ac.uk)[1], Ivan I. Vankov (i.i.vankov@gmail.com)[2],
Casimir J. Ludwig (c.ludwig@bristol.ac.uk)[1], Jeffrey S. Bowers (j.bowers@bristol.ac.uk)[1]
1 University of Bristol, School of Psychological Science
2 New Bulgarian University, Department of Cognitive Science and Psychology

Abstract:

What mechanism supports our ability to recognize objects over a wide range of different retinal locations? Most research in psychology and neuroscience suggests that learning to identify a novel object at one retinal location only supports the ability to identify that object at nearby retinal locations, and to date, neural network models of object identification show a similar restriction in generalization. As a consequence, it is widely assumed that objects need to be learned at multiple locations. We challenge this view and show the capacity to generalize across retinal locations (what we call on-line translation tolerance) has been underestimated in humans and artificial neural networks. Two eye tracking studies demonstrate that novel objects can be recognized following translations of 9° and even 18°. Additionally, computational studies showed that convolutional neural networks can achieve similarly robust generalization when a mechanism (Global Average Pooling) was built in to generate larger receptive fields.

Keywords: Translation Tolerance; Translation Invariance; Object Recognition; Vision

Introduction

We can identify familiar objects despite the variable images they project on our retina, including variation in image size, orientation, illumination, and position on retina. How the visual system succeeds under these conditions is still poorly understood. Here we focus on our ability to identify objects despite variations in retinal location and consider the extent to which the visual system relies on "on-line" vs. "trained" translation tolerance. In the case of on-line tolerance, learning to identify an object at one location immediately affords the capacity to identify that object at multiple retinal locations. Trained tolerance, by contrast, refers to the hypothesis that we learn to identify familiar objects across locations by explicitly training the visual system to identify each object across a broad range of retinal locations.

Most of the empirical research in psychology and neuroscience suggests that on-line tolerance is restricted to a few degrees of visual angles (e.g., Afraz & Cavanagh, 2008; Cox & DiCarlo, 2008; Dill & Fahle, 1997; Dill & Fahle, 1998; Nazir & O'Regan, 1990; Newell, Sheppard, Edelman, & Shapiro, 2005). *Figure 1* outlines a selection of studies that used different experimental paradigms and found highly limited (in one case no) translation tolerance. Based on the outcome of such studies, Chen et al. (2017) state that translation tolerance is limited to a few degrees and write: "Given limited translation-invariance from a single glance in human vision, it is reasonable to conclude that saccades (rapid eye movements) are the mechanism for translation-invariant recognition in practice". (p. 544)

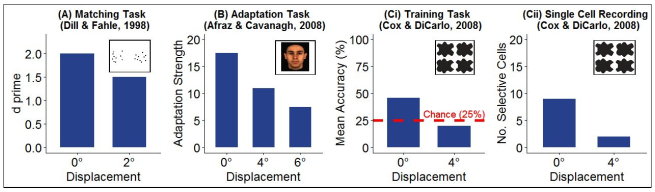

Figure 1. Behavioral studies of translation tolerance.

Work with artificial neural network models has also reported limited generalization when stimuli are presented at untrained locations. For example, Elliffe, Rolls and Stringer (2002) showed that a biologically inspired neural network model called VisNet supported on-line translation tolerance, but each stimulus had to be trained at multiple spatial locations (after training in 7 locations the model generalized to an 8th and 9th location), and the authors only tested small translations (8 pixels in a 128x128 retina).

The above findings have led most theorists in psychology and neuroscience to endorse the 'trained' account of translation tolerance (Chen et al., 2017; Cox & DiCarlo, 2008; Kravitz et al. 2008). Despite empirical and computational results, there are still reasons to question this hypothesis. With regards to the behavioral studies, the stimuli that show limited translation tolerance are typically unlike real objects (e.g., Dill & Fahle, 1998; see *Figure 1a*) and/or are very similar to each other (e.g., Cox & DiCarlo, 2008; see *Figure 1c*). Differentiating between unfamiliar and highly similar items may rely on low-level visual representations that are retinotopically constrained (Kravitz et al., 2008).

With regards to the computational studies, there is reason to believe deep convolutional neural networks (CNNs) might support more robust on-line translation tolerance. Indeed, these models are designed to support translation tolerance by including convolutional layers and pooling layers in order to speed up training and ensure that high-level units have larger receptive fields. That said, we are only aware of four studies that have assessed on-line tolerance in CNNs, and surprisingly, these studies have reported highly restricted tolerance (see below). Here we show that some classes of CNNs can support robust on-line translation invariance.

Behavioral Experiments

Here we report two experiments each with 10 participants. Eye-movements were monitored using the Eyelink 1000 plus system (SR Research). In the learning phase participants were trained to categorize the 24 objects as ‘A’ or ‘B’ (see *Figure 2*). Participants were required to maintain their gaze on a centrally located fixation-cross for 1000ms for an object to appear. If gaze moved 1.5° beyond the fixation-cross, a mask replaced the object.

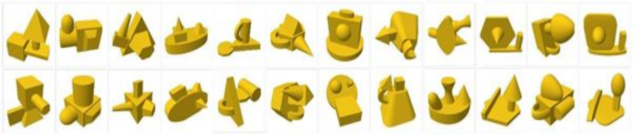

Figure 2. Twenty-four novel objects taken from Leek et al. (2016). Pairs of objects are matched for similar local features but differ in global configuration.

Experiment 1 assessed 9° on-line translation tolerance: All 24 objects were first studied at the center of the screen and then twelve objects were trained 9° to the left of fixation and the other 12 trained 9° to the right of fixation (training continued until 12/12 consecutive correct answers given at each peripheral location). In the test phase, all objects were tested at 9° left, 9° right, and centre of fixation. Experiment 2 assessed 18° on-line translation invariance: half images were presented 9° to the right, and half 9° to the left of central fixation (none at the center) (training continued until 24/24 consecutive correct answers given in each location). All objects were tested 9° to the left and 9° right of fixation.

The results of Experiments 1-2 are summarised in *Table 1*. Near complete on-line tolerance was obtained following 9° shifts (Experiment 1) and robust tolerance extended to 18° (Experiment 2).

***Table 1.* Mean (+/- 95% CI range) Accuracy scores in Experiment 1 and 2.** Columns show degrees by which the test presentation was displaced from the nearest training location and the screen position of that test presentation.

	Mean (+/- 95% CI range) Accuracy			
Displacement	**0°**	**0°**	**9°**	**18°**
Screen Position	**Centre (trained)**	**Peripheral (trained)**	**Peripheral (novel)**	**Peripheral (novel)**
Exp 1 (N=10)	93% (5%)	83% (5%)	81% (6%)	not tested
Exp 2 (N=10)	not tested	97% (3%)	not tested	89% (7%)

Simulations

We examined on-line translation tolerance in convolutional neural networks (CNNs) that rival human capacity at identifying objects in some conditions. These models use convolution operations which learn to detect the same features across all spatial positions, and in most cases, include pooling layers that aggregate information from multiple spatially organized units to a single unit that represents more abstract image features. Although these features might be expected to support on-line translation tolerance the only studies to date have found that on-line tolerance is highly restricted (Chen et al., 2018; Furukawa, 2017; Kauder-Abrams, 2017; Qi 2017) (see *Figure 3*). A key feature of these previous models is that they did not include a Global Average Pooling (GAP) layer designed to provide larger receptive fields that may be relevant to supporting more robust on-line tolerance.

Here, we used a popular CNN with and without a GAP layer (VGG16; Simonyan & Zisserman, 2014) by training the network to classify the 24 ‘Leek’ images (*Figure 2*) as ‘A’ or ‘B’ at restricted locations, and then testing its accuracy at five displaced locations. The Leek images were 50x50 pixels and were presented within 400x400 pixel space to allow for large displacements at test. In all simulations, training continued until the model reached 100% accuracy. The results are displayed in *Table 2*.

***Table 2.* Mean Accuracy of CNN when classifying Leek (2016) stimuli over large translations**

	Displacement				
VGG16 pretrained on imagenet	**30 px**	**110 px**	**190 px**	**270 px**	**310 px**
GAP	0.97	0.99	0.96	0.97	0.92
No GAP	0.84	0.51	0.57	0.49	0.51

Clearly GAP is required in order to support robust on-line translation tolerance.

Discussion

In two behavioral experiments we demonstrated that participants trained to recognise novel objects at one retinal position can recognise the same objects at untrained distal retinal-locations (up to 18° displacement in Experiment 2) with high accuracy, and in simulations we have identified a condition in which CNNs can also support extreme on-line tolerance. *Figure 3* highlights how much greater these translation tolerance effects are compared to past studies.

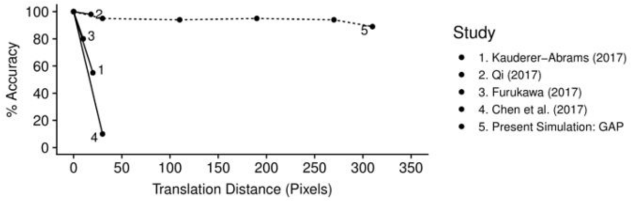

Figure 3: Comparison of accuracy scores of previous CNN simulations with our pre-trained CNN that included GAP, plotted as a function of translation distance (pixels).

The findings have important implication for theories of vision. Most computational models of word and object identification only support highly restricted on-line translation tolerance and instead depend on trained tolerance to explain how humans identify images across a wide range of retinal locations. For example, Chen et al. (2017) developed a CNN model of number identification designed to support translation invariance in a biologically and psychologically plausible manner. The model showed zero on-line tolerance at displacements beyond 12 pixels (which they equate to 2° of visual angle; see *Figure 3*). Our behavioral findings falsify this and other theories that rely on trained tolerance in order to account for our ability to identify objects over a wide range of retinal locations (Dandarund et al., 2013; DiBono & Zorzi, 2013; Elliffe, Rolls & Stringer, 2002).

Researchers have identified neurons in inferior-temporal cortex (IT) with a range of receptive field sizes (ranging from 2.8° to 26°; for review see Kravitz et al., 2008), but the larger receptive fields have been attributed to trained translation tolerance (e.g., Chen et al., 2017; Cox & DiCarlo, 2008). Our findings suggest that these large receptive fields may be the product of mechanisms that support extreme on-line translation tolerance.

Note, we are not claiming that our results support the conclusion that the visual system supports complete on-line translation invariance. There was a small decrement in performance when novel objects were presented to novel locations in Experiment 2. For this reason we can only conclude that the visual system supports extreme on-line translation tolerance. That said, the results do not rule out complete on-line invariance. It is possible that performance was better in the trained locations because performance in this condition was not only supported by newly learned representations within high-level visual systems that support invariance, but also low-level visual systems that support a more limited degree of on-line translation tolerance. There is no reason to assume that the low-level visual system did not contribute to the performance with our stimuli as well as high-level visual systems, leading to slightly better performance in the same location. In any case, our results clearly support the conclusion that high-level visual systems supports extreme on-line translation tolerance, and future research is needed to determine whether on-line invariance is supported.

With regards to the modelling, it is clear why we found much greater on-line translation tolerance compared to past work, namely, we used CNNs with a Global Average Pooling (GAP) layer that is standard in state-of-the art convolutional networks. Although previous models included some degree of pooling, it was clearly not sufficient to produce the large receptive fields that mediated the extreme tolerance we observed. When using GAP, the receptive fields of the neurons at the final layer of the model cover 100% of the pixel space. It is important to emphasize that we are not committed to the hypothesis that the large receptive fields in human visual system are the product of a GAP-like mechanism. This is one possible solution, but the receptive fields that support extreme translation tolerance may have an entirely different source. For example, in GEON theory (e.g., Biederman, 1987), objects are identified on the basis of identifying parts and the relations between the parts rather than some sort of template matching approach that characterizes CNNs. Translation tolerance on this approach is due to the claim that the parts and relations are coded independently of retinal position, but the GAP mechanisms in CNNs are poorly suited for this form of computation (Sabour, Frosst, & Hinton, 2017). We hope our findings generate more research into understanding the mechanisms that support extreme translation tolerance, including assessing the extent to which GAP or alternative mechanisms play a role in high-level human vision.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the

European Union's Horizon 2020 research and innovation programme (grant agreement No 741134).

References

Afraz, S.R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research, **48,*** 42–54.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, **94,*** 115-147.

Bowers, J. Vankov, I. & Ludwig, C. (2016). The visual system supports online translation invariance for object identification. *Psychonomic Bulletin Review, **23**,* 432-438.

Chen, F., Roig, G., Isik, X., Boix, L. & Poggio, T. (2017). Eccentricity dependent deep neural networks: Modeling invariance in human vision. *AAAI Spring Symposium Series*, Science of Intelligence.

Cox, D. D. & DiCarlo, J. J. (2008). Does Learned Shape Selectivity in Inferior Temporal Cortex Automatically Generalize Across Retinal Position? *Journal of Neuroscience*, ***28,*** 10045-10055.

Dandurand, F., Hannagan, T., & Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connection Science, **25,*** 1-26.

Di Bono, M. G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology, **4,*** 635.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, **89,*** 3264-3278.

Dill, M., & Fahle, M. (1997). The role of visual field position in pattern discrimination learning. *Proceedings of the Royal Society B, **264,*** 1031-1036.

Dill, M. & Fahle, M. (1998) Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics, **60,*** 65–81

Elliffe, M. C. M., Rolls E. T., & Stringer S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics, **86,*** 59– 71.

Furukawa, H. (2017). Deep learning for target classification from SAR imagery: Data augmentation and translation invariance. *IEICE Technical Report, **117,*** 11-17.

Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450.*

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, **12,*** 114-122.

Leek, E. C., Roberts, M. V., Oliver, Z. J., Cristino, F., & Pegna, A. (2016). Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. *Neuropsychologia, **89,*** 495–509.

Nazir T, & O'Regan J. K. (1990). Some results on translation invariance in the human visual system. *Spatial Vision, **5,*** 81-100.

Newell, F. N., Sheppard, D. M., Edelman, S., & Shapiro, K. L. (2005). The interaction of shape- and location-based priming in object categorisation: Evidence for a hybrid "what + where" representation stage. *Vision Research, **45,*** 2065-2080.

Qi, W. (2018). A quantifiable testing of global translation invariance in convolutional and capsule networks.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.