Tracking in space and features with phase synchrony

Sabine Muzellec$^{\star 1,2}$, Drew Linsley$^{\star 1,3}$, Alekh K. Ashok^{1,3}, Rufin VanRullen^{2}, Thomas Serre^{1,3}
Carney Institute for Brain Science, Brown University^{1}
CerCo - CNRS^{2}
Department of Cognitive Linguistic & Psychological Sciences, Brown University^{3}
{sabine_muzellec,drew_linsley}@brown.edu

Abstract

Healthy humans depend on their ability to track objects while they move through the world even as they change in appearance. Here, we introduce the *FeatureTracker* challenge to systematically evaluate and compare the abilities of humans and state-of-the-art deep neural networks (DNNs) to track objects that change in appearance over time. While humans can effortlessly solve this task, DNNs cannot. Drawing inspiration from cognitive science and neuroscience, we describe a novel recurrent neural circuit that can induce this tracking capability in DNNs by leveraging the oscillatory activity of its neurons to follow objects even as their appearances change. The resulting complex-valued recurrent neural network (CV-RNN) outperformed all other DNNs and approached human accuracy on the *FeatureTracker* challenge. The success of this novel neural circuit provides computational evidence for a long-hypothesized role of phase synchronization for visual attention and reasoning.

Keywords: Neural circuits; Object tracking; Synchrony

Introduction

Human observers find it effortless to track objects in their surroundings even as they change in appearance or state over time (Corbetta et al., 1990; Blaser et al., 2000). For example, when preparing a meal, we have no trouble tracking ingredients even as chopping and cooking them changes their shapes, sizes, colors, and textures. Converging lines of research in cognitive science and neuroscience have made it clear that humans rely on at least three distinct strategies to track objects, each of which can be flexibly selected and combined based on task demands. (*i*) Humans can recognize objects by "re-identifying" them over time (DiCarlo et al., 2012; Jia et al., 2021). (*ii*) It has also been known for many decades that biological visual systems — including those of humans — are exquisitely sensitive to the motion of objects, and this resource can be used for tracking (Cavanagh, 1992). (*iii*) Humans can also track objects by following their *feature dynamics*, or the rate of change of specific visual features over time. This strategy was demonstrated through a psychophysics paradigm depicted in Blaser et al. (2000): humans could track one target Gabor overlaid with a distractor Gabor, as both smoothly changed in spatial frequency, color, and orientation over time. While there has been extensive progress made in identifying mechanisms underlying human object tracking by re-recognition and/or motion (Jia et al., 2021; Linsley et al., 2021), and extending those mechanisms into modern deep learning-based systems for object tracking (Chen et al., 2022, 2021; Linsley et al., 2021), there is still little known about how humans track objects by their *feature dynamics*. What neural mechanisms support this object-tracking strategy?

Method and Results

The *FeatureTracker* challenge. We began by developing a large-scale synthetic challenge that could be used to test the abilities of human and machine observers to track objects by their *feature dynamics*. Our *FeatureTracker* challenge does this by asking observers to determine if an object that begins in a red square travels to the blue square by the end of the video (Fig. 1A). Our challenge is built on an earlier one that tested the abilities of humans and machines to track objects in videos as they cross and occlude the view of each other (Linsley et al., 2021). Inspired by seminal psychophysics work (Blaser et al., 2000), we extended this earlier challenge by causing the objects in each video to change in appearance over time. We reasoned that such regularities in *feature dynamics* would make the task of object tracking easier for those observers who could leverage them.

Each video in the *FeatureTracker* challenge consists of a sequence of 32 frames that are 32 × 32 pixels, depicting a red "start" square, a blue "goal" square, and 11 objects, one of which begins each video in the red square and is meant to be tracked. As the objects move over the course of the video, their shapes, colors, or shapes and colors change. By holding out regions of color and shape space for training versus testing, we first trained observers on feature variations, then tested the extent to which they could leverage an object's *feature dynamics* to track it.

The challenge begins with a training phase of videos, where the observer is trained on videos in which the colors and shapes of objects are drawn from half of the total range of values that these can take (Fig. 1A, first row of examples). DNNs are trained on 100,000 videos, while humans are trained on just 20. Next, the observer is tested on versions of the task where object shapes and colors are drawn from the same or different ranges as those seen during training (Fig. 1A, second-fourth row of examples). The ability to track objects by their *feature dynamics* would support high performance regardless of the manipulation that occurs during test time, and a failure to generalize means that an observer did not learn to implement the appropriate tracking strategy.

While humans achieved high performance on every version of the *FeatureTracker* challenge, state-of-the-art DNNs (trained using a BCE loss) were far less successful (Fig. 1B).

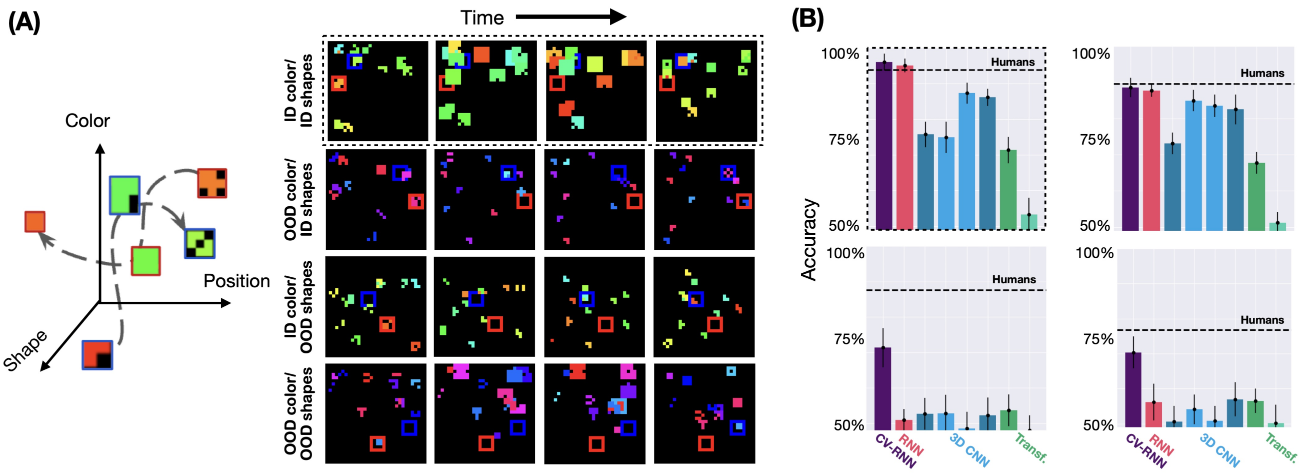

Figure 1: (**A**) Inspired by Blaser et al. (2000), we develop a large-scale synthetic object tracking challenge that we call *FeatureTracker*. The goal of *FeatureTracker* is to watch a video and say if the object that begins in the red square travels to the blue square by the end of the video. The task is difficult because (*i*) objects occlude each other over time, and (*ii*) the appearances of each object also change over time in prescribed ways. Specifically, humans and models are trained on videos where the shapes and colors of objects vary within prescribed ranges and then tested on versions where shapes and colors vary within the same or different ranges (middle column; each row depicts a distinct testing condition). (**B**) Humans and recurrent neural networks (the proposed complex-valued recurrent neural circuit, CV-RNN, and a real-valued version of the same model, RNN) rival human accuracy when tested on shapes/colors that vary in the same range as those seen during training (first row of **A**) or shapes that are distinct from those seen during training (third row of **A**). However, only the CV-RNN comes close to human performance when tested on versions of *FeatureTracker* where colors vary in different ranges than those seen during training (second and fourth row of **A**). Models tested also include 3D convolutional neural networks (3D CNNs) and Transformers. 3D CNNs and Transformers pre-trained on natural videos are denoted with darker bars.

An RNN with attention (Linsley et al., 2021) and multiple 3D Convolutional Neural Networks (3D CNNs) rivaled human performance when tested on objects that had similar appearances as seen during training (left corner). However, each model fell to chance when they were asked to track objects with features (mainly colors) that were dissimilar to training.

Phase synchrony supports object tracking by *feature dynamics*. We hypothesized that the failure of existing models at solving the *FeatureTracker* challenge stems from interference in their feature representations. Specifically, models are unable to separately represent information about an object's appearance and its position over time. Evidence from neuroscience has implied neural oscillations as a mechanism for multiplexing different sources of information, with minimal interference, within the same neuronal population (Sternshein et al., 2011; Drew et al., 2009). These findings are strongly related to the role of phase synchronization in perceptual grouping (Woelbern et al., 2002; Elliott & Müller, 2001), suggesting a global key mechanism linking object-based attention and phase synchronization. Inspired by this body of research, we developed a novel attentional mechanism that could model neural oscillations, and use them to solve the *FeatureTracker* challenge. Specifically, we modified the attention head of an existing RNN for object tracking, the Index-and-Track RNN (Linsley et al., 2021), giving it the capacity to use the phase information of its complex-valued activity to encode object positions (agnostic to color and shape), and amplitude to encode object appearance (color and shape). These modifications simply consisted of transferring the existing operations from the real to the complex domain, following the framework proposed by Reichert & Serre (2013). Our resulting model, the Complex-Valued RNN (CV-RNN) was significantly better than the other DNNs on each condition of the *FeatureTracker* challenge.

Conclusion

Humans have an extraordinary ability to track objects through the world even as they change in appearance, state, or visibility through occlusion. Our results imply that oscillations act as a key mechanism underlying this ability and that inducing DNNs with this capability can help them behave more like humans.

Acknowledgments

This work was funded by an ONR grant N00014-24-1-2026 to TS. We acknowledge the Center for Computation and Visualization at Brown University for hardware support funded by NIH Office of the Director grant S10OD025181.

References

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. *Nature*, *408*(6809), 196–199.

Cavanagh, P. (1992). Attention-based motion perception. *Science*, *257*(5076), 1563–1565.

Chen, X., Yan, B., Zhu, J., Wang, D., & Lu, H. (2022, March). High-Performance transformer tracking.

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021, March). Transformer tracking.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., & Petersen, S. E. (1990). Attentional modulation of neural processing of shape, color, and velocity in humans. *Science*, *248*(4962), 1556–1559.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012, February). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Drew, T., McCollough, A. W., Horowitz, T. S., & Vogel, E. K. (2009). Attentional enhancement during multiple-object tracking. *Psychonomic Bulletin & Review*, *16*, 411–417.

Elliott, M. A., & Müller, H. J. (2001). Effects of stimulus synchrony on mechanisms of perceptual organization. *Visual Cognition*, *8*(3-5), 655–677.

Jia, X., Hong, H., & DiCarlo, J. J. (2021, June). Unsupervised changes in core object recognition behavior are predicted by neural plasticity in inferior temporal cortex. *Elife*, *10*.

Linsley, D., Malik, G., Kim, J., Govindarajan, L. N., Mingolla, E., & Serre, T. (2021). Tracking without re-recognition in humans and machines. *Advances in Neural Information Processing Systems*, *34*, 19473–19486.

Reichert, D. P., & Serre, T. (2013). Neuronal synchrony in complex-valued deep networks. *arXiv preprint arXiv:1312.6115*.

Sternshein, H., Agam, Y., & Sekuler, R. (2011). Eeg correlates of attentional load during multiple object tracking. *PloS one*, *6*(7), e22660.

Woelbern, T., Eckhorn, R., Frien, A., & Bauer, R. (2002). Perceptual grouping correlates with short synchronization in monkey prestriate cortex. *Neuroreport*, *13*(15), 1881–1886.