# Invariant recognition drives neural representations of action sequences

**Leyla Isik*** **(lisik@mit.edu)**

**Andrea Tacchetti*** **(atacchet@mit.edu)**

**Tomaso Poggio (tp@ai.mit.edu)**

Center for Brains Minds and Machines, MIT
* denotes equal contribution

## Abstract

**Recognizing the actions of others from visual stimuli is a crucial aspect of visual perception. Humans are able to identify similar behaviors and discriminate between distinct actions despite transformations, like changes in viewpoint or actor, that substantially alter the visual appearance of a scene. This ability to generalize across complex transformations is a hallmark of human visual intelligence. Advances in understanding motion perception at the neural level have not always translated in precise accounts of the computational principles underlying what representation of action sequences our visual cortex evolved or learned to compute. Here we test our hypothesis that invariant action discrimination might fill this gap. We show that spatiotemporal CNNs appropriately categorize video stimuli into actions, and that deliberate model modifications that improve performance on an invariant action recognition task lead to data representations that better match human neural recordings. Our results suggest that performance on invariant discrimination dictates the neural representations of action sequences computed by visual cortex. Moreover, these results broaden the scope of the invariant recognition framework understand human visual intelligence to the study of perception of action sequences.**

**Keywords:** Action recognition; Convolutional Neural Networks; Magnetoencephalography; Represenational Similarity Analysis

## Introduction

Humans' ability to recognize the actions of others is a crucial aspect of visual perception. Remarkably, the accuracy with which we can finely discern what others are doing is largely unaffected by transformations that, while substantially changing the visual appearance of a given scene, do not change the semantics of what we observe (e.g. a change in viewpoint). Fueled by advances in computer vision methods for object and scene categorization, recent studies have made progress towards linking computational outcomes to their biological substrate by highlighting a correlation between performance optimization on discriminative object recognition tasks and the accuracy of neural predictions both at the single recording site and neural representation level (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cichy, Pantazis, & Oliva, 2014). However, these results, have not been extended to action perception and dynamic stimuli. What specific computational goals underly the representations of action sequences in human visual cortex?

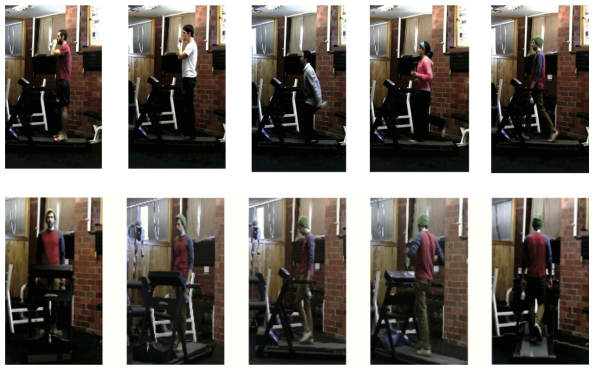


Figure 1: Action recognition dataset consisting of five actors performing five actions from five views.

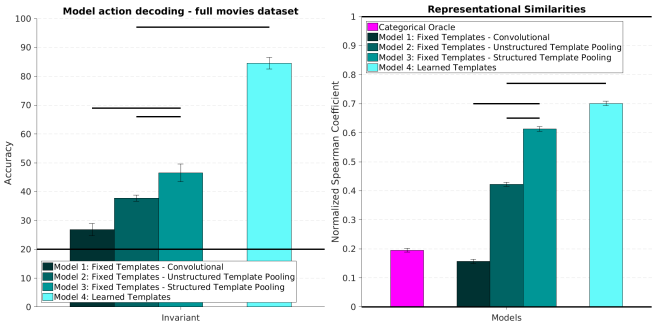
Here we test the hypothesis that invariant recognition might fill this gap. We use artificial systems for action recognition and compare their data representations to human magnetoencephalography (MEG) recordings (Isik, Tacchetti, & Poggio, 2016). We show that, within the Spatiotemporal Convolutional Neural Networks (ST-CNNs) model class, deliberate modifications that result in better performing models on invariant action recognition, also result in a better representational match to human neural recordings. Our results suggest that performance optimization on discriminative tasks, especially those that require generalization across complex transformations, alongside the constraints imposed by the hierarchical organization of motion processing in visual cortex dictates the representation of action sequences in the brain.

## Results

We filmed a video dataset showing five actors, performing five actions (drink, eat, jump, run and walk) at five different viewpoints (Fig. 1). We used these videos to develop four variants of feedforward hierarchical models of visual cortex explicitly designed to exhibit varying degrees of invariance to changes in 3D viewpoint. These models were instances of ST-CNNs (Karpathy et al., 2014), which are direct extensions of the CNNs used to recognize objects or faces in static images (LeCun, Bengio, & Hinton, 2015), to input stimuli that extend both in space and time. Subsequently, in order to assess each model's robustness to changes in viewpoint, we extracted feature representations of videos showing two different viewpoints, frontal and side using each model. We then trained a machine learning classifier to discriminate video se-

quences into different action classes based on each model's output and using a single viewpoint, and evaluated the classifier accuracy in predicting the action depicted in new, unseen videos at a mismatching viewpoint (e.g. a classifier trained on the frontal viewpoint would be tested on the side viewpoint).

All the models we considered produced representations that were, at least to a minimal degree, invariant to changes in viewpoint (Fig. 2a). Importantly, it was possible to rank the models we considered based on performance on this invariant action recognition task and, in particular, the end-to-end trainable models (model 4) performed better than models 1, 2 and 3, which used fixed templates. Within this group, models that employed a Structured Channel Pooling mechanism to increase robustness performed best (Leibo, Mutch, & Poggio, 2011).




a b

Figure 2: ST-CNN model performance on a viewpoint invariant recognition task (a) predicts reprentational matches with neural responses (b). Horizontal lines between conditions indicate significant difference $p < 0.05$ (ANOVA and Bonferroni corrected t-test).

We used Representational Similarity Analysis (RSA) to assess how well each model feature representation matched human neural data. RSA produces a measure of agreement between artificial models and brain recordings based on the correlation between empirical dissimilarity matrices. We used video feature representations extracted by each model from a set of new, unseen stimuli to construct model dissimilarity matrices and MEG sensor recordings of the neural activity elicited by those same stimuli to compute a neural dissimilarity matrix (Isik et al., 2016). Finally, we constructed a dissimilarity matrix using an action categorical oracle. In this case, the dissimilarity between videos of the same action was zero and the distance across actions was one. We observed that end-to-end trainable model (model 4) produced dissimilarity structures that better agreed with those constructed from neural data than models with fixed templates. Within models with fixed templates, model 3, constructed using a Structured Pooling mechanism to build invariance to changes in viewpoint, produced representations that agree better with the neural data than models employing Unstructured Pooling (model 2) and purely convolutional models (model 1) (Fig. 2b). This ranking is aligned with what we obtained using models' performance on an invariant action recognition task. Moreover, a dissimilarity matrix based on a categorical oracle matched the neural data worse than data representations based on robust CNNs.

## Conclusions

Recognizing the actions of others from complex visual stimuli is a crucial aspect of human perception. We investigated the relevance of invariant action discrimination to improving model representations' agreement with neural recordings. Our deliberate approach to model design underlined the relevance of both supervised, gradient based methods and memory based, structured pooling methods to the modeling of neural data representations. Importantly, a categorical oracle did not outperform convolutional architectures, highlighting the relevance of both the computational task and the architectural constraints to obtaining quantitatively accurate models of neural representations of action sequences.

Recognizing the semantic category of visual stimuli across photometric, geometric or more complex changes, in very low sample regimes is a hallmark of human visual intelligence. We show that by building data representations that support this kind of robust recognition one obtains empirical dissimilarity structures matching those in human neural data. These results suggest that invariant discrimination is one of the computational principles shaping the representation of action sequences in human visual cortex. In the wider context of the study of perception, our results strengthen the claim that the computational goal of human visual cortex is to support invariant recognition by broadening it to the study of dynamic action perception.

## Acknowledgements

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

## References

Cichy, R. M. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–62.

Isik, L., Tacchetti, A., & Poggio, T. (2016). Fast , invariant representation for human action in the visual system. *arXiv 1601.01358*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. F. (2014). Large-scale video classification with convolutional neural networks. In *CVPR.*

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comp. Bio.*, *10*(11).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Leibo, J. Z., Mutch, J., & Poggio, T. (2011). How can cells in the anterior medial face patch be viewpoint invariant? *MIT-CSAIL-TR-2010-057, CBCL-293*.

Yamins, D. L. K., 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. *PNAS*, *111*(23), 8619–24.